Linguística de *Corpus*

Perspectivas

Organizadoras:
Maria José Bocorny Finatto
Rozane Rodrigues Rebechi
Simone Sarmento
Ana Eliza Pereira Bocorny

INSTITUTO DE LETRAS UFRGS
UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Linguística de *Corpus*

Perspectivas

Organizadoras:
Maria José Bocorny Finatto
Rozane Rodrigues Rebechi
Simone Sarmento
Ana Eliza Pereira Bocorny

FAPERGS
Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul
INSTITUTO DE LETRAS U F R G S
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS
UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

1ª edição: 2018

Capa: Ethel Kawa
Preparação de originais: Carlos Batanoli Hallberg
Revisão: Lia Cremonese
Editoração eletrônica: Fernando Piccinini Schmitt

Esta coletânea foi publicada graças ao apoio recebido da FAPERGS, processo 17/0399-3, Edital 06/2016 EDITAL FAPERGS 06/2016 – AOE, que apoiou o XVI Encontro de Linguística de *Corpus* (ELC 2017) e IX Escola Brasileira de Linguística Computacional (EBRALC 2017). Esta coletânea é um livro derivado do evento, reúne uma seleção de trabalhos gerados a partir de diferentes atividades de ambos os eventos. Todos os trabalhos aqui publicados foram avaliados por Comissão Científica especialmente convidada. Os Anais do evento correspondem a uma outra publicação denominada "Caderno de Resumos do ELC-EBRALC 2017", ISBN: 9788561424183.

Versão DIGITAL gratuita disponível em:
PPG-LETRAS-UFRGS:
https://www.ufrgs.br/ppgletras/ebooks.html

Site do evento:
http://www.ufrgs.br/elc-ebralc2017

L755 Linguística de *corpus* : perspectivas [recurso eletrônico] / Organizadoras: Maria José Bocorny Finatto, Rozane Rodrigues Rebechi, Simone Sarmento, Ana Eliza Pereira Bocorny. — Porto Alegre: Instituto de Letras - UFRGS, 2018.
575 p.

Requisitos do sistema: Adobe Reader.
Modo de acesso: World Wide Web

1. Linguística. 2. Linguística de *corpus*. I. Finatto, Maria José Bocorny. II. Rebechi, Rozane Rodrigues. III. Sarmento, Simone. IV. Bocorny, Ana Eliza Pereira.

CDD 410

Catalogação na publicação: Vladimir Luciano Pinto – CRB 10/1112

ISBN 978-85-64522-36-7

Agradecimentos

Ao trazer a público esta coletânea, *Linguística de Corpus: perspectivas*, importa agradecer ao apoio indispensável da **FAPERGS** – Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, cujo subsídio, em especial, permitiu-nos a produção dessa obra. Agradecemos também ao Instituto de Letras da UFRGS por sua editoria e ao Programa de Pós-Graduação em Letras por sua publicação *on-line*, que nos permite o compartilhamento de produções de acesso livre em formato *e-book*.

O conteúdo aqui apresentado está ligado às atividades do XIV Encontro de Linguística de *Corpus* (ELC) e da IX Escola Brasileira de Linguística Computacional (EBRALC), dois eventos associados, realizados entre 15 e 18 de agosto de 2017, nas dependências da Unisinos, em São Leopoldo (RS). ELC e EBRALC foram fruto de uma parceria institucional entre a UFRGS e a Unisinos, representados pelo PPG-LETRAS-UFRGS e pelo PPGLA da Unisinos. Para esse evento, contamos com os apoios da **CAPES** (Edital PAEP-03-2017, Proc. 88887.138115/2017-00) e do **CNPq** (Chamada ARC-2017, Proc. 403858/2017-8). O evento também teve o suporte da Embaixada Norte-Americana no Brasil, dos certificadores ETS, TOEFL e do nosso Idiomas Sem Fronteiras da UFRGS.

Os artigos aqui reunidos, entretanto, extrapolam os trabalhos apresentados no evento, visto que partem dos temas dos minicursos, palestras e *workshops*. Todos os textos candidatados a esta coletânea foram avaliados por pares, membros de um Comitê Científico especialmente convidado, composto pelos seguintes colegas, aos quais registramos todo o nosso reconhecimento.

Ana Eliza Pereira Bocorny (UFRGS)
Ana Luiza Pires de Freitas (UFCSPA)

Cristiane Krause Kilian (Instituto Superior de Educação Ivoti – RS)
Cristina Becker Lopes Perna (PUCRS)
Guilherme Fromm (UFU)
Heliana Ribeiro de Mello (UFMG)
Larissa Brangel (Unisinos)
Leonardo Zilio (Université Catholique de Louvain – Bélgica)
Lia Cremonese (UFRGS)
Lucelene Lopes (PUCRS)
Luciana Latarini Ginezi (Uninove)
Maria José Bocorny Finatto (UFRGS)
Patricia Tosqui Lucks (ICEA)
Rodrigo Souza Wilkens (Université Catholique de Louvain – Bélgica)
Rove Luiza de Oliveira Chishman (Unisinos)
Rozane Rodrigues Rebechi (UFRGS)
Simone Sarmento (UFRGS)

Conforme acreditamos, a Linguística de *Corpus* (LC) tem se associado a diferentes perspectivas de investigação, seja nos Estudos da Linguagem, com destaque para Linguística Aplicada e Estudos do Texto, seja na área de Ciência da Computação, pela via do Processamento da Linguagem Natural (PLN), sem praticamente nada rejeitar em termos de parcerias de trabalho e de trocas de conhecimentos. A disposição ao diálogo disciplinar e interdisciplinar e o princípio de compartilhamento de materiais de pesquisa têm sido marcas constantes da comunidade brasileira de LC. Assim, esta coletânea pretende ratificar a importância da nossa área e apontar novos modos de seguirmos adiante. Por isso, vale o agradecimento a cada um dos autores dos textos aqui apresentados por sua disposição a esses e a futuros diálogos e por sua confiança no nosso trabalho de organização deste livro.

Maria José Bocorny Finatto
Rozane Rebechi
Simone Sarmento
Ana Eliza Pereira Bocorny

Porto Alegre (RS), abril de 2018.

Sumário

Tradução, Lexicografia e Terminologia

Linguística Aplicada / Ensino (LAP)

E a Linguística de *Corpus* vai desbravando novos horizontes...

Stella Tagnin

Desde a publicação do primeiro *corpus*, o *Brown University Standard Corpus of Present-Day American English*[1], mais conhecido simplesmente como *Brown Corpus* (1964), compilado por Henry Kučera e W. Nelson Francis na década de 1960, a Linguística de *Corpus* vem alçando voo por áreas dificilmente cogitadas naquela época.

Além de ter introduzido os estudos estatísticos da linguagem no que viria a se denominar Linguística de *Corpus* (LC), o *Brown Corpus* serviu de modelo para a construção de outros *corpora*, como o *London-Oslo-Bergen Corpus*[2] (*LOB Corpus,* 1978), uma réplica para o inglês britânico de sua estrutura: 500 textos de 2.000 palavras em 15 gêneros, totalizando 1 milhão de palavras.

Esse montante parece irrisório quando comparado aos *corpora* atuais, que podem ultrapassar 1 bilhão de palavras, como o *News on the Web* (*NOW)*, com mais de 5 bilhões de palavras, ou o *Global Web-Based English* (*GloWbE*), o *Wikipedia Corpus*, o *Hansard Corpus*, e o *Corpus del Español*[3], todos com aproximadamen-

[1] Esse *corpus* pode ser baixado pelo *site* <http://www.nltk.org/nltk_data/>, ou consultado *on-line* pelo Sketch Engine (<https://the.sketchengine.co.uk/open/>).
[2] Esse *corpus* pode ser baixado pelo *site* <http://ota.ox.ac.uk/desc/0167>.
[3] Todos esses *corpora* – e vários outros – fazem parte da plataforma criada por Mark Davies, da Brigham Young University e podem ser acessados em <https://corpus.byu.edu/corpora.asp>.

te 2 bilhões de palavras. O *Corpus do Português*[4] faz parte do mesmo portal e é constituído de um *corpus* histórico com 45 milhões de palavras do século XIII ao século XX e um *corpus* extraído da *Web* com textos do Brasil, de Portugal, Moçambique e Angola. Não podemos deixar de mencionar o *Corpus Brasileiro*[5]. Embora a língua inglesa ainda seja privilegiada em termos de variedade de *corpora* disponíveis, há um considerável número de *corpora* para outras línguas, tanto acessíveis *on-line* quanto *off-line* (VIANA, 2015)[6], esses últimos em geral compilados por pesquisadores para um objetivo específico.

Outro marco instituído pelo *Brown Corpus* foi a etiquetagem morfossintática do *corpus*, ou seja, a cada palavra foi atribuída uma categoria gramatical. Esse procedimento viabilizou análises estatísticas mais sofisticadas do que permitia o *corpus* cru, ou seja, sem anotação. Hoje, essa prática tornou-se corriqueira e pode ser realizada com um dos vários etiquetadores disponíveis[7].

Mas voltemos às áreas de atuação da Linguística de *Corpus*, boa parte delas representada pelos artigos deste volume. Talvez a primeira área que se beneficiou da LC tenha sido a Lexicografia, com a publicação do *American Heritage Dictionary* (MORRIS, 1969), o primeiro dicionário baseado em *corpus* – no caso, o *Brown Corpus*. A partir daí, a maioria das editoras americanas e britânicas passaram a publicar seus dicionários com base em *corpora*, muitas vezes próprios, como é o caso da Collins, Cambridge, Macmillan e Oxford (ATKINS; RUNDELL, 2008), para citar apenas algumas.

A Fraseologia é outra área que se desenvolveu com o ferramental da Linguística de *Corpus*, que permite identificar recorrências lexicais com muita facilidade. Da Fraseologia à Terminologia e, posteriormente, à Terminologia Fraseológica (TAGNIN; BEVILACQUA, 2015), foi um pulo. São incontáveis os trabalhos nessas áreas. Uma rápida busca no Google Books por "fraseologia corpus" e "terminologia corpus" resulta, respectivamente, em 4.090 e 28.000 *hits*. E isso considerando apenas livros, sem levar em conta os inúmeros artigos acadêmicos que abordam esses tópicos e o sem-número de glossários compilados nos mais diversos domínios do saber.

---

[4] Disponível em: <http://www.corpusdoportugues.org/>.

[5] Esse *corpus* pode ser acessado gratuitamente via Linguateca: <http://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS>, ou por meio do Sketch Engine (acesso gratuito por um mês): <https://www.sketchengine.co.uk/corpus-brasileiro/>.

[6] Esse artigo traz uma lista, obviamente não exaustiva, de 100 *corpora* com descrição de suas características. Outras informações sobre *corpora* disponíveis *on-line* podem ser obtidas em <http://martinweisser.org/corpora_site/online_corpora.html>.

[7] Estes são apenas alguns dos etiquetadores disponíveis:
- USAS *on-line* English tagger: <ucrel.lancs.ac.uk/usas/tagger.html>
- Free CLAWS WWW tagger: <http://ucrel.lancs.ac.uk/claws/trial.html>
- Tree-Tagger: <http://corpuslg.org/tools/etiquetagem/>
- Aelius Brazilian Portuguese POS-Tagger: <https://sourceforge.net/projects/aelius/>.

Outra disciplina que se vale da Linguística de *Corpus* é a Tradução (VIANA; TAGNIN, 2015; ZANETTIN, 2012; LAVIOSA, 2008), principalmente a técnica, área em que a interdisciplinaridade se faz evidente, devido ao papel preponderante que a Terminologia aí desempenha. Mas a tradução literária também recorre à LC, tanto para uma pesquisa lexical visando a uma diversificação de vocabulário quanto para uma investigação de como certos vocábulos foram traduzidos, por exemplo, nomes próprios ou termos culturalmente marcados. Para esses estudos, são imprescindíveis *corpora* comparáveis (textos originais em duas ou mais línguas) e *corpora* paralelos (originais e respectivas traduções em duas ou mais línguas).

Os *corpora* muito têm contribuído para o aprimoramento da tradução automática, em especial, a estatística, que cria modelos estatísticos a partir da análise de grandes *corpora* multilíngues (CASELI, 2015), chegando hoje a um modelo baseado em regras até para a tradução para Libras (PAIVA et al., neste volume).

A linguagem oral também mereceu o olhar da Linguística de *Corpus*, e são vários os *corpora* orais disponíveis, dentre os quais destacamos o *MICASE* (*Michigan Corpus of Academic Spoken English*)[8] e o *Santa Barbara Corpus of Spoken American English*[9], nos Estados Unidos (SINCLAIR, 2004), e, em nosso país, o *C-Oral*[10], de fala espontânea do português brasileiro.

Um *corpus* oral ainda pode servir de subsídio para o ensino de língua estrangeira. Um exemplo bastante inovador envolve o recurso a um *corpus* de interações orais entre controladores de tráfego aéreo e pilotos em situações de emergência para o ensino de inglês aeronáutico (Prado, neste volume). Se nos dermos conta da contribuição que esse estudo pode ter para a diminuição de acidentes aéreos, teremos uma ideia de quão alto a Linguística de *Corpus* ainda pode voar.

Na realidade, a área de Linguística Aplicada é uma das que mais se desenvolveu com a LC (SINCLAIR, 2004; HUNSTON, 2002), não só no ensino propriamente dito (VIANA; TAGNIN, 2011; MEUNIER; GRANGER, 2008; O'KEEFFE; MCCARTHY; CARTER 2007), como também na elaboração de gramáticas (CONRAD; BIBER, 2009).

Os estudos de gêneros textuais (DEIGNAN; SEMINO; PAUL, 2017; FLOWERDEW, 2005; HYLAND, 2004) e discursivos (SEMINO et al., 2018; BAKER et al.; 2008; BAKER, 2006) também reconheceram a relevância da Linguística de *Corpus* para obter análises mais objetivas e acuradas. É um campo em franco desenvolvimento, inclusive com congressos exclusivamente dedicados a esses temas, como é o caso da quarta edição do *CAD 2018*[11] (*Corpora and Discourse International Conference*), em que o conceito de "discurso" é entendido

[8] Disponível em: <https://www.lib.umich.edu/database/michigan-corpus-academic-spoken-english-micase>.
[9] Disponível em: <http://www.linguistics.ucsb.edu/research/santa-barbara-corpus>.
[10] Disponível em: <http://www.c-oral-brasil.org/>.
[11] Disponível em: <http://ucrel.lancs.ac.uk/cad2018/>.

de forma abrangente como "linguagem em uso", o que inclui "gêneros e registros específicos".

Todas essas áreas envolvem linguistas, estudiosos da língua e das linguagens. Recentemente, no entanto, Bowker[12] preconizou que a Linguística de *Corpus* tem aplicações que podem ser relevantes não apenas para os linguistas. No caso, a autora referia-se à ciência da informação e biblioteconomia (*Library and Information Science*, *LIS* na sigla em inglês). No *blog*, Bowker, referindo-se a um artigo seu intitulado *Corpus Linguistics: it's not just for linguists!*, afirma que foi instigada por um artigo de Risso (2016, p. 74), em que a autora alega que a área de *LIS* "necessita de novos desenvolvimentos metodológicos que combinem abordagens qualitativas e quantitativas". Bowker inicialmente descreve as técnicas básicas da Linguística de *Corpus* e, em seguida, passa a detalhar como essas podem ser aplicadas em várias áreas da *LIS*. Conclui afirmando que, após 25 anos, desde que a LC se consagrou como metodologia essencial para os estudos linguísticos, "parece ser seguro dizer que a Linguística de *Corpus* não é mais apenas para linguistas".

As principais ferramentas a que Bowker se refere são as listas de frequência, as palavras-chave, as linhas de concordância, as colocações, que são encontradas em *softwares* especialmente criados para a análise de *corpora*. Um dos mais antigos e mais usados é o WST, desenvolvido por Mike Scott (2006), mas Laurence Anthony criou o AntConc em 2014, um *freeware* que vem ganhando cada vez mais adeptos, principalmente porque seu autor vem adicionando tantas novas e variadas ferramentas, só ou em colaboração com outros pesquisadores, que enseja análises complexas e multifacetadas. Toda essa produção fez com que fosse chamado de "o Thomas Edison da Linguística de *Corpus*" no *blog All About Corpora*[13].

Mesmo vislumbrando áreas não linguísticas em que a Linguística de *Corpus* possa fazer uma contribuição significativa, qualquer área em que a língua tenha um papel relevante – e qual é a área que pode dispensar a língua? – só tem a ganhar em termos de objetividade e confiabilidade com o uso da Linguística de *Corpus*, principalmente porque a tecnologia avança a passos largos nesse setor, permitindo a criação de ferramentas que permitem análises cada vez mais específicas e direcionadas.

Referências

ATKINS, B. T. S.; RUNDELL, M. *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press, 2008.

BAKER, P. *Using Corpora in Discourse Analysis.* London: Continuum, 2006.

[12] Disponível em: <https://allaboutcorpora.com/corpus-linguistics-not-just-linguists>. Acesso em: 22 jan. 2018.
[13] Disponível em: <https://allaboutcorpora.com/laurence-anthony>. Acesso em: 22 jan. 2018.

BAKER, P.; GABRIELATOS, C.; KHOSRAVINIK, M.; KRZYZANOWSKI, M.; McENERY, T.; WODAK, R. A useful methodological synergy? Combining critical discourse analysis and *corpus* linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19 (3), p. 273-306, may 2008.

CASELI, H. de M. O uso de *corpora* paralelos para a criação de um tradutor automático estatístico. In: VIANA, V.; TAGNIN, S. E. O. Corpora *na Tradução*. São Paulo: HUB Editorial, 2015, p. 243-267.

CONRAD, S.; BIBER, D. *Real Grammar – A Corpus-Based Approach to English.* White Plains: Pearson Education, 2009.

DEIGNAN, A.; SEMINO, E.; PAUL, S.-A. Metapahors of Climate Science in Three Genres: Research Articles, Educational Texts, and Secondary School Student Talk. *Applied Linguistics*, amx035, out. 2017. Disponível em: <https://academic.oup.com/applij/advance-article/doi/10.1093/applin/amx035/4396285>. Acesso em: 22 jan. 2018.

FLOWERDEW, L. An integration of *corpus*-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against *corpus*-based methodologies. *English for Specific Purposes*, v. 24, n. 3, p. 321-332, 2005.

HUNSTON, S. Corpora *in Applied Lingusitics.* Cambridge: Cambridge University Press, 2002.

HYLAND, K. *Genre and second language writing.* Michigan: University of Michigan Press, 2004.

LAVIOSA, S. *Corpus-based Translation Studies – Theory, Findings, Applications.* Amsterdam: Rodopi, 2008.

MEUNIER, F.; GRANGER, S. *Phraseology in Foreign Language Learning and Teaching.* Amsterdam: John Benjamins, 2008.

O'KEEFFE, A.; McCARTHY, M.; CARTER, R. *From Corpus to Classroom – language use and language teaching.* Cambridge: Cambridge University Press, 2007.

RISSO, V. G. Research methods used in library and information science during the 1970-2010. *New Library World*, 117 (1/2), p. 74-93, 2016.

SCOTT, M. *WordSmith Tools.* 2006. Disponível em: <http://www.lexically.net/wordsmith/downloads/>. Acesso em: 9 jun. 2016.

SEMINO, E.; DEMJEN, Z.; HARDIE, A.; PAYNE, S. A.; RAYSON, P. E. *Metaphor, cancer and the end of life*: a *corpus*-based study. London: Routledge, 2018.

SINCLAIR, J. (Ed.). *How to use* corpora *for language teaching.* Amsterdam: John Benjamins, 2004.

TAGNIN, S. E. O.; BEVILACQUA, C. Corpora *na Terminologia.* São Paulo: HUB Editorial, 2015.

VIANA, V. *Corpora* para consulta *on-line* e *off-line*. In: VIANA, V.; TAGNIN, S. E. O. Corpora *na Tradução*. São Paulo: HUB Editorial, 2015, p. 269-320.

VIANA, V.; TAGNIN, S. E. O. Corpora *no Ensino de Línguas Estrangeiras.* São Paulo: HUB Editorial, 2011.

_____. Corpora *na Tradução.* São Paulo: HUB Editorial, 2015.

MORRIS, W. (Ed.). *American Heritage Dictionary.* Boston: Houghton-Mifflin, 1969.

ZANETTIN, F. *Translation-driven* corpora. Manchester: St. Jerome Publishing, 2012.

Estudos de gêneros textuais e discursivos

ComentCorpus: o uso de mecanismos linguísticos na detecção de ironia e sarcasmo para o português do Brasil em um *corpus* opinativo

ComentCorpus: the use of linguistic mechanisms in the detection of irony and sarcasm for Brazilian Portuguese in a *corpus* of opinion

Gabriela Wick Pedro
Oto Araújo Vale

Resumo: O uso da ironia e do sarcasmo em textos na *Web* é capaz de alterar o sentido de uma sentença ou transformar sua polaridade. A presença de itens lexicais contrastantes em um dado contexto pode ser um indicador de ironia. O presente artigo tem por objetivo detectar padrões linguísticos capazes de indicar e caracterizar ironia e sarcasmo presente em um *corpus* opinativo do português do Brasil. Para examinar as ocorrências de ironia e sarcasmo parte-se da busca e anotação em *corpus* previamente construído por comentários de notícias do portal da *Folha de S. Paulo*, identificando expressões irônicas e sarcásticas. Assim, espera-se colaborar com o desenvolvimento da área e aperfeiçoando ferramentas de identificação automática de opinião.

Palavras-chave: Anotação de *corpus*. Ironia. Análise de Sentimentos. Opinião.

Gabriela Wick Pedro – Mestranda/Bolsista CAPES no Programa de Pós-Graduação em Linguística pela Universidade Federal de São Carlos (UFSCar), bacharel em Linguística pela mesma universidade – gabiwick@gmail.com.
Oto Araújo Vale – Professor Associado na Universidade Federal de São Carlos (UFSCar), doutor pela Universidade Estatual Paulista (UNESP) – otovale@ufscar.br.

Abstract: The use of irony and sarcasm in texts on the *Web* is capable of altering the meaning of a sentence or transforming its polarity. The presence of contrasting lexical items in a given context may be an indicator of irony. The goal of this project is to detect linguistic patterns capable of indicating and characterizing irony and sarcasm present in an opp. *Corpus* of Brazilian Portuguese. To examine the occurrences of irony and sarcasm, one starts with the search and annotation in *corpus* previously constructed by news commentaries of the *Folha de S. Paulo* portal, identifying ironic and sarcastic expressions. Thus, it is expected to collaborate with the development of the area and improving tools for automatic identification of opinion.

Keywords: *Corpus* Annotation. Irony. Sentiment Analysis. Opinion.

1 Introdução

Nas últimas duas décadas, com o crescimento da *Web 2.0*, grande parte da comunicação diária passou a ser *on-line*. Como consequência, as chamadas mídias sociais tornaram-se uma fonte valiosa de informações sobre a opinião pública a respeito de produtos, empresas, políticos, tendências, entre outros (PANG; LEE, 2008, p. 11). Estes dados passam a ser interessantes não só pelo seu amplo volume, mas por servir de subsídio para aplicações em PLN (Processamento de Língua Natural), fornecendo informações relevantes tanto para as Ciências da Computação, que busca identificar automaticamente noções e aspectos linguísticos, quanto para a Linguística Computacional, que procura compreender e descrever fenômenos linguísticos para o processamento computacional de línguas naturais, possibilitando o aperfeiçoamento de ferramentas linguístico-computacionais. Diante disso, a Linguística Computacional começa a se importar não somente com informações objetivas encontradas em textos, mas passa a se dedicar também aos aspectos subjetivos da linguagem, buscando textos não jornalísticos em uma linguagem não padronizada e avaliativa.

Do interesse em processar automaticamente a subjetividade e as emoções subjacentes à língua, surge a Análise de Sentimentos, ou Mineração de Opinião[1], uma área interdisciplinar que busca interpretar e analisar, computacionalmente, opiniões, avaliações, sentimentos e emoções expressas em texto sobre uma determinada entidade e todos seus atributos relacionados (LIU, 2010, p. 1). Estudos sobre a subjetividade detectam relevâncias a respeito de construções e métodos para a delimitação de opinião em textos argumentativos, podendo assumir diferentes particularidades, ou seja, é possível caracterizar um determinado texto de acordo com seus pontos favoráveis ou desfavoráveis sobre um dado objeto ou identificar diferentes perspectivas em um debate político.

[1] Segundo Pang e Lee (2008, p. 9), a área que trata computacionalmente opinião, emoção, avaliação e subjetividade pode ser encontrada por diversas nomenclaturas: Análise de Sentimento, Mineração de Opinião ou Análise de Subjetividade.

Através de estatísticas, descrições com base em teorias linguísticas, a Análise de Sentimentos tenta extrair e caracterizar o conteúdo de sentimento e opinião de um texto. Os pontos principais de interesse para a Análise do Sentimento podem ser vários, e é importante levar em consideração os domínios que a pesquisa está inserida quando cria-se um *corpus*. O mesmo problema de interesse pode estar relacionado a uma determinada língua, uma vez que vários recursos estão disponíveis para algumas línguas quando quase faltam para muitos outros, como o português do Brasil.

De um modo geral, Liu (2010, p. 1) classifica o que é encontrado em formato textual entre fato e opinião. Fato são as sentenças objetivas desprovidas de sentimentos e transmitem apenas informações sobre algo. Já opinião são as sentenças subjetivas que carregam alguma avaliação, sentimento, opinião ou emoção sobre algo ou alguém. No entanto, a grande dificuldade está exatamente na separação do que é apresentado como conteúdo informativo e factual do que é subjetivo e opinião, uma vez que a linguagem não é tão exata como na consideração acima. Consequentemente, os conceitos de subjetividade e emoções estão intimamente ligados. Enquanto o propósito de uma sentença objetiva é apresentar uma informação factual, a sentença subjetiva pode aparecer em várias formas como: opiniões, desejos, crenças, suspeitas, emoção, pensamentos ou especulações (WIEBE et al., 2005, p. 172; RILOFF et al., 2013, p. 704). As emoções, por sua vez, são sentimentos ou pensamentos de alguém.

Estudos sobre a subjetividade detectam relevâncias a respeito de construções e métodos para a delimitação de opinião em textos opinativos, podendo assumir diferentes particularidades, ou seja, é possível caracterizar um determinado texto de acordo com seus pontos favoráveis ou desfavoráveis sobre um dado objeto ou até identificar diferentes perspectivas em um debate político.

Sabe-se que a manipulação da linguagem figurada é uma das tarefas mais desafiadoras do PLN, dado que esse tipo de linguagem é muitas vezes caracterizada por dispositivos linguísticos, como por exemplo, ironia, sarcasmo, metáforas e humor. O seu significado ultrapassa o significado literal e é, portanto, difícil a sua apreensão até mesmo para os seres humanos. Para Councill et al. (2010, p. 56), a ironia é um recurso linguístico comum em textos que expressam opiniões subjetivas, e sua presença representa um obstáculo significativo para uma análise precisa do sentimento em materiais textuais, principalmente quando o que representa são informações relevantes na tomada de decisão da polaridade de uma sentença.

A ironia exige, muitas vezes, um conhecimento de mundo e a familiaridade com o contexto conversacional e o contexto cultural dos participantes da conversa, informações estas que a máquina não pode acessar facilmente. Estudar expressões irônicas em dados coletados em um *corpus* constituído por textos avaliativos pode aprimorar diversas ferramentas e pesquisas baseadas em Processamento de Língua Natural (PLN). Além disso, pelo fato da linguagem figurada modificar-se

frequentemente devido a mudanças no vocabulário e na própria linguagem, o que dificulta o treinamento de algoritmos de aprendizado de máquina, por exemplo, torna-se uma tarefa trabalhosa e complexa.

O problema na detecção automática da ironia é a quebra de todos os aspectos da língua, desde a pronúncia até a escolha lexical, estrutura sintática, semântica e conceitualização. Assim, não é realista buscar uma solução computacional milagrosa para linguagem figurada, e uma resposta geral não será encontrada em nenhuma técnica ou algoritmo. Em vez disso, deve-se identificar aspectos específicos e formas de linguagens figuradas suscetíveis de análise computacional e, a partir desses tratamentos individuais, tentar sintetizar uma solução gradualmente mais ampla.

A ironia verbal[2] é um conceito tradicionalmente definido como o oposto do que realmente se quer dizer. Dessa forma, a precisão da polaridade de uma sentença pode ser significativamente minada pela presença de ironia, como pode-se observar nos exemplos a seguir[3]:

(1) Ainda bem que este homem mais "honesto" do mundo não vai se candidatar.
(2) Como é bom ver os políticos brasileiro gastarem o dinheiro público com propina.

Essas sentenças seriam classificadas, provavelmente, como positivas, enquanto a intenção é, inegavelmente, negativa. Em (1), as aspas no adjetivo "honesto" indicam a presença de ironia. Em contraste, em (2), não há nenhuma indicação explícita de ironia presente. No entanto, é perceptível a presença de ironia, porque, dado o nosso conhecimento do mundo, sabe-se que o ato dos políticos brasileiros gastarem o dinheiro público com propina não é algo realmente bom. Considerando esse contraste, supõe-se que o autor não é sincero sobre o sentimento expresso, mas quer ser irônico.

Para examinar as ocorrências de ironia e sarcasmo, partiremos da busca em um *corpus* previamente construído por comentários de notícias e anotado de acordo com as expressões irônicas e sarcásticas identificadas. Para o português europeu, há o Senti*Corpus*-PT (CARVALHO et al., 2011, p. 565), um *corpus* de compilações de comentários de notícias que antecederam os debates das eleições de 2009 em Portugal e manualmente anotado como alvo entidades humanas, especificamente políticos.

[2] Para Muecke (1995, p.58) a ironia verbal é quando ocorre uma inversão semântica e, assim, é utilizada para dizer uma coisa e significar outra. Em outras palavras, entende-se a ironia como uma expressão verbal cujos constituintes formais, ou seja, palavras, tentam comunicar um significado subjacente oposto ao expresso.
[3] Exemplos extraídos do *ComentCorpus*.

A anotação de *corpus* pode ser definida como o processo de enriquecer um *corpus*, adicionando informações linguísticas inseridas por humanos ou máquinas com um objetivo teórico ou prático. Embora um *corpus* represente um recurso muito útil para estudos linguísticos, um *corpus* anotado constitui um recurso linguístico ainda mais importante e valioso. Isso acontece porque as anotações acrescentam valor ao *corpus*, permitindo que sejam realizados buscas e processamentos mais refinados.

Uma anotação, quando manual, pode ser realizada em uma pequena porção relativa de *corpus* e de forma econômica com fenômenos bastante complexos. Segundo Hovy e Lavid (2010, p.13), o trabalho manual é lento e limitado a pequenos resultados. Por outro lado, os autores afirmam que a anotação automática requer um investimento considerável na construção do *corpus* e na programação do sistema de anotação automatizado. Embora o resultado de uma anotação de *corpus* automática possa ser rápida, em grandes quantidades de *corpora*, o resultado pode ser de má qualidade[4].

Buscamos aqui compreender como a ironia verbal funciona e como pode ser reconhecida em conteúdo gerado pelo usuário (*user-generated content* ou UGC[5]) e propor um modelo de anotação manual para um *corpus* de textos opinativos. Com base na definição clássica de ironia verbal, a hipótese é que a detecção de opiniões irônicas pode ser realizada através de itens lexicais, podendo ser, portanto, um bom indicador para reconhecê-las em um *corpus*. Espera-se que este trabalho seja relevante para o desenvolvimento de recursos linguísticos e aplicações em ferramentas linguístico-computacionais.

O artigo está estruturado da seguinte forma: a seção 2 apresenta uma visão geral dos conceitos de ironia e trabalhos relacionados sobre tratamento automático de ironia. A seção 3 apresenta a área de Análise de Sentimentos e suas aplicações. A seção 4 apresenta o *corpus* da pesquisa e as diferentes etapas de processo de anotação. Na seção 5 discute-se, em uma análise do *corpus* anotado, uma série de estatísticas apresentada para fornecer informações sobre os dados obtidos no *corpus*, além de mostrar os resultados entre anotadores que avalia as diretrizes de anotação. Finalmente, a seção 6 conclui o artigo com algumas perspectivas para futuras pesquisas.

[4] Uma anotação de *corpus* pode acontecer em todos os níveis linguísticos – morfossintático, sintático, semântico, pragmático ou discursivo. Além disso, pode ocorrer nas seguintes formas: i) manual: através de linguistas; ii) automática: por ferramentas de PLN; ou iii) semiautomática, quando a correção da saída de outras ferramentas é manual.

[5] Qualquer tipo de mídia, como comentários ou postagens em redes sociais, no qual o usuário produz de modo espontâneo sobre um produto ou assunto.

2 Conceitos gerais para ironia

A tarefa de compreender significados irônicos é realizada com relativa facilidade pelos humanos, e alguns dos atos de fala envolvidos nesta operação ainda podem causar mal-entendidos comunicacionais. Para Grice (1975, p. 58), a ironia é conhecida como uma forma de comunicar o oposto do significado literal, ou quando existe uma aparente violação de princípios pragmáticos. No entanto, a ironia parece ser um mecanismo mais complexo amplamente estudado na filosofia e na linguística (GRICE, 1975; SPERBER; WILSON, 1981, p. 295; KREUZ; GLUCKSBERG, 1989, p. 374; UTSUMI, 1996, p. 962). Reyes et al. (2013, p. 2) afirmam que a ironia é uma propriedade inesperada ou incongruente em uma situação. Kreuz e Glucksberg (1989, p. 374) acreditam que a presença desse recurso figurado transmite um significado pragmático ao remeter às expectativas falhadas do falante.

Grice (1975, p. 58) tenta explicar o funcionamento da ironia da perspectiva pragmática, que está focada em conceitos que transpassam os níveis tradicionais literários e linguísticos, considerando também elementos extralinguísticos. Segundo o autor, as intenções de comunicação de um indivíduo se dão pela análise do sentido figurado a partir de implicaturas conversacionais (GRICE, 1975, p. 43). Nas implicaturas conversacionais, há a necessidade de inferências pragmáticas para manter o significado do que foi dito. Para ele, a ironia acontece quando há uma violação da máxima de qualidade.

De acordo com Searle (1971), a busca por um significado não literal começa quando o ouvinte percebe que o enunciado do falante é inadequado ao contexto, isto é, um enunciado não consegue fazer sentido em contraste com o contexto. Assim, a sentença "*Que dia lindo*" só será irônica se o dia estiver chuvoso. Consequentemente, em vez de dizer diretamente algo a uma pessoa, o falante menciona alguma certeza para lembrar o ouvinte dessa certeza quando elas não aparecem na situação em que ocorre.

Sperber e Wilson (1981, p. 301) propõem uma explicação para a ironia verbal como um uso ecoico da linguagem, no qual o falante ecoa implicitamente, isto é, remete a pensamento que atribui a outros pensamentos, sejam eles reais ou não, para expressar sua atitude crítica ou ridicularizada, dando-a como falsa, irrelevante ou pouco informativa. O conteúdo desse enunciado ecoico se assemelha ao da proposição superficial.

O eco pode ocorrer após o que foi dito, mas também pode estar retomando pensamentos reais ou imaginários. Em uma noção mais ampla, o eco tem suas restrições definidoras. Assim, uma representação acessível não pode ser definida como eco, tendo em vista que ele será lembrado a partir da inacessibilidade à representação na qual haverá uma checagem com a relevância do enunciado. Para exemplificar, os autores dão o seguinte exemplo:

(3) a. Estou cansado.
b. Está cansado? E o que você acha que estou?

Em suma, Sperber e Wilson compreendem a ironia como uma variedade de uso interpretativo na qual a proposição expressada por um enunciado representa uma crença atribuída implicitamente pelo falante a alguém ou a si mesmo em outra ocasião. Ademais, é um caso particular de uso ecoico; enquanto o falante expressa implicitamente sua própria atitude em relação aos pensamentos que sua afirmação interpreta e finalmente, trata a ironia como uma atitude expressada de maneira implícita pelo falante sobre os pensamentos que ele faz; logo, eco é a dissociação desses pensamentos.

Para Kreuz e Glucksberg (1989, p. 381), na teoria do lembrete ecoico, os interlocutores reconhecem o sarcasmo quando percebem que um falante está se referindo a um conhecimento anterior. Essa origina-se da teoria da menção ecoica de Sperber e Wilson (1981, p. 303). Os casos de menção indicam que o enunciado foi ouvido e entendido e expressam a reação imediata do ouvinte. Os autores acreditam que o tom de voz, o contexto, as palavras escolhidas pelos falantes, são comuns na fala cotidiana e podem indicar a atitude do falante à proposição referida.

Muitas vezes a ironia pode ser percebida como uma mistura de sarcasmo e sátira, cujo efeito não está baseado apenas em expressar um significado oposto, mas também humor. Estudos têm tentado esclarecer o problema da distinção entre ironia verbal e sarcasmo. Alguns autores, como Grice e Sperber e Wilson diferenciam apenas um termo, enquanto outros consideram a ironia e o sarcasmos como o mesmo fenômeno (ATTARDO, 2000, p. 795; REYES et al., 2013, p. 5). No entanto, além do fato do dispositivo representar melhor cada exemplo, destaca-se o fato de que, para muitas pessoas, não há uma distinção clara sobre os limites de diferenciação entre ironia e o sarcasmo, por exemplo. Entretanto, teoricamente falando, podemos argumentar que a ironia tende a ser um modo de comunicação mais sofisticado do que o sarcasmo, pois enquanto a ironia, geralmente, enfatiza uma pretensão divertida, o sarcasmo está preocupado com o sentido mordaz e as provocações desordenadas. Assim, enquanto a ironia atinge ambiguidade e muitas vezes exibe uma grande sutileza, o sarcasmo é entregue ao ouvinte com um tom frio ou seco que raramente é ambíguo. Todavia, essas diferenças dependem de questões de uso e tom, em vez de apenas pressupostos teóricos.

Nossa pesquisa não faz distinção entre os dois fenômenos, mas sim se concentra em descrever expressões que podem abranger tanto aquelas descritas como ironia verbal quanto as que são consideradas outros tipos de linguagem figurada. Ou seja, aqui se diferenciam o objetivo e o efeito da ironia. O objetivo da ironia, de acordo com a definição, é comunicar o oposto do que é literalmente dito. O efeito, entretanto, pode ter uma interpretação sarcástica, satírica ou mesmo divertida que,

sem dúvida, introduz conotações negativas. Nesse contexto, é conveniente tratar ironia e os dispositivos relacionados como diferentes facetas do mesmo fenômeno. Portanto, dispositivos como sarcasmo, sátira ou humor serão considerados como extensões específicas de um conceito geral e amplo de ironia.

Como já dito anteriormente, o reconhecimento de sentença de ironia representa um grande desafio para a Análise de Sentimentos. González-Ibáñez et al. (2011, p. 582) sugerem que características lexicais por si só não são suficientes para identificar sarcasmo e que características pragmáticas e contextuais merecem mais atenção. Hogenboom et al. (2013, p. 13) fazem uma análise de como os *emoticons* podem transmitir sentimentos. Para isso, os autores criaram, manualmente, um léxico de sentimento composto por *emoticons* para melhorar o método de classificação de sentimento baseado em léxico.

Reyes (2012, p. 30) analisa a ironia em termos de um modelo multidimensional de elementos textuais, com a identificação de um conjunto de características discriminativas para difundir automaticamente textos irônicos a partir de textos não irônicos. Os autores construíram um modelo de ironia para o Twitter para o qual confiaram em um conjunto de recursos textuais para capturar *tweets* irônicos.

Para o português, destaca-se o trabalho de Carvalho et al. (2009, p. 2), que elaboraram um conjunto de padrões linguísticos para o português europeu, como *emoticons*, expressões onomatopeicas, pontuação e aspas. Algumas dessas pistas são específicas para o português (padrões morfológicos), enquanto outras parecem ser independentes da linguagem e estão presentes em todos os lugares nas mídias sociais (*emoticons*). Vanin et al. (2013, p. 636) apresentam um trabalho inicial para o português do Brasil sobre alguns padrões de detecção de ironia em *tweets*. No artigo, desenvolveram-se padrões com formas diminutivas, que podem tanto expressar sentimentos positivos, como afeto, ternura e intimidade, como também podem apresentar conotações sarcásticas ou irônicas, quando a intenção é desvalorizar ou insultar uma determinada entidade.

Do ponto de vista da Linguística Computacional, descrever a ironia talvez seja uma tarefa cognitivamente ousada, tendo em vista que o grau de computabilidade é difícil e exaustivo. Entretanto, é imprescindível, pois permite o desenvolvimento de uma série de pesquisas na área, como decodificar informações e associá-las ao uso do interlocutor em UGC.

3 Análise de Sentimentos e Opinião

As opiniões são conteúdos subjetivos encontrados na linguagem, seja ela falada ou escrita. Com a popularização da internet, essas opiniões passaram a ser mais acessíveis, principalmente pelos curiosos e estudiosos da linguagem, que procuram verificar quais são os desejos, sentimentos e avaliações de cada indivíduo.

Opiniões podem influenciar os pensamentos, os comportamentos e o modo de agir de uma pessoa, pois o que cada indivíduo acredita está ligado à maneira como os outros veem o mundo. Antigamente, as pessoas consultavam a opinião de amigos, familiares ou pessoas próximas antes de tomar qualquer decisão, e as empresas realizavam pesquisas através de enquetes, consultores ou grupos de discussões. Contudo, com o avanço das tecnologias e a facilidade de comunicação, as pessoas deixaram o "boca a boca" e passaram a expressar suas opiniões e sentimentos em comentários de *blogs*, fóruns e postagens em redes sociais.

Dentro desse contexto, surge a Análise de Sentimentos, ou Mineração de Opinião: é uma área recente e agrega estudos de mineração de dados, linguística computacional e inteligência artificial (PANG; LEE, 2008; LIU, 2010). Segundo Liu (2010, p. 1), é o estudo computacional de opiniões, sentimentos, avaliações, atitudes, estimativas, emoções, subjetividade, entre outros, em formato textual, e que são encontrados na *web* em *blogs*, *sites* de notícias, fóruns, resenhas de produtos etc.

Descrever e formalizar as emoções humanas não é uma tarefa nem um pouco simples. Liu (2010, p.3) lista três problemáticas encontradas: i) identificar as expressões de opinião sobre um determinado assunto ou alvo em um *corpus*; ii) classificar a polaridade da opinião, isto é, marcá-la como positiva, neutra ou negativa; e iii) apresentar os resultados de forma sumarizada.

Um ponto importante que pode ser destacado no trabalho de Liu (2010, p.3) é o fato dos termos "sentimentos" e "opinião" geralmente serem vistos como sinônimos, sendo representados através de uma atitude, opinião, avaliação ou a emoção sobre um alvo, podendo ser classificados de acordo com a sua polaridade (positiva, negativa ou neutra). Porém, o termo "emoção" é usado para apontar as percepções e pensamentos subjetivos, tais como alegria, tristeza, raiva, medo e surpresa, e não representam necessariamente uma atitude ou pensamento em relação ao alvo.

Formalmente, segundo Liu (2010, p. 5), uma opinião corresponde a uma quíntupla (e_i, a_{ij}, s_i j_{kl} , h_k, t_l), onde: e_i é o nome de uma entidade; a_{ij} é um aspecto da entidade e_i; s_{ijkl} é a polaridade do sentimento sobre aspecto a_{ij} que tem como alvo a entidade e_i; h_k é o detentor do sentimento, ou seja, quem expressou o sentimento, também pode ser chamado de fonte de opinião e t_l é o instante no qual a opinião foi expressa por h_k.

Algumas pesquisas vêm sendo realizadas na área da Análise de Sentimentos, mas poucas delas são voltadas para descrição e análise linguística. Em sua maioria, são trabalhos da Ciência da Computação que buscam desenvolver *softwares* para automatizar processos de análise de *corpus*. Para o português do Brasil, destaca-se o ReLi (FREITAS, 2013, p. 3). O ReLi é um *corpus* de resenha de livros anotado manualmente quanto à expressão de opinião, objeto de opinião, e à sua polaridade. O *corpus* possui 1.600 resenhas de 7 autores, totalizando aproximadamente

260 mil palavras. Para cada livro (que no total são 13 livros) foi coletado por volta de 200 resenhas.

O SAPair (SILVA et al., 2012, p. 4) é um método original para a Análise de Sentimentos no nível da característica[6] usando pares (característica, palavra opinativa). É feita a extração de opiniões de usuários de um domínio particular (empresas, *sites* de compras) e, a partir desse processo, gera-se uma lista de pares. A cada par, a palavra opinativa expressa o sentimento do falante sobre a uma determinada característica.

Em Oliveira et al. (2012, p.4), foram coletados dados do Twitter duas semanas antes e uma semana após as Olimpíadas de 2012. Logo após, foi realizado um pré-processamento desses dados, que, depois, foram submetidos a diferentes algoritmos de agrupamentos. As filtragens e formas de visualizações foram realizadas após o agrupamento.

Nem toda opinião é, necessariamente, expressa com um item lexical de sentimento. Por exemplo, se alguém disser "comprei um celular ontem e a câmera já deixou de funcionar", não há nenhum elemento lexical que faça o ouvinte interpretar a frase negativamente, mas sabe-se que essa avaliação é negativa. O uso de palavras de sentimentos não é uma condição necessária e nem suficiente para detectar conteúdos subjetivos como opinião e classificar sua polaridade. Isso ocorre porque, em primeiro lugar, a polaridade das palavras de opinião dependerá do contexto em que ela se insere. No exemplo (4), o adjetivo "rápido" expressa um sentimento positivo sobre desempenho do congelador da geladeira. Já no exemplo (5), "rápido" questiona a bateria do celular.

(4) O congelamento dessa geladeira é rápido.
(5) A bateria do celular descarrega rápido.

Becker e Tumitan (2013, p. 6) afirmam que a ironia e o sarcasmo são muito comuns em alguns domínios, como o político e o esportivo. Portanto, a ironia e o sarcasmo são um dos maiores desafios na manipulação de opinião.

4 Coleta de *corpus* e anotação

A maioria das abordagens atuais que tratam automaticamente a ironia se utilizam de dados públicos e disponíveis postos em mídias sociais – incluindo *tweets* e revisões de produtos (CARVALHO et al., 2009, p. 2; DAVIDOV et al., 2010, p. 108). Para esta pesquisa, procurou-se trabalhar com um *corpus* composto por comentários retirados da *Web* por se tratar de textos nos quais os usuários

[6] No nível de aspecto (ou entidade), são extraídos e agrupados os aspectos de uma determinada entidade, identificando a polaridade para cada um deles.

expressam-se através de impressões, opiniões, sentimentos, emoções e avaliações. Outro aspecto importante sobre a delimitação do *corpus* se dá pelo fato de serem textos não jornalísticos em uma linguagem informal com a opinião direta do locutor. Por isso, a escolha de comentários de leitores de um jornal de grande circulação e que se tenha acesso *on-line* dessas informações.

O *ComentCorpus* é um *corpus* de comentários de notícias no português do Brasil com anotações semântico-discursivas conforme a intenção de cada sentença identificada. O *corpus* é composto por 6185 comentários extraídos manualmente de 90 notícias relacionadas ao período de *impeachment* da presidente Dilma Rousseff, totalizando aproximadamente 14 mil sentenças e 207 mil *tokens*. A escolha das notícias se deu através da busca pela palavra *impeachment* no caderno Poder[7], do jornal *Folha de S. Paulo*, no período de janeiro a junho de 2016. Na Tabela 1, são apresentadas as características do *corpus*.

Tabela 1 – Característica do *ComentCorpus*

Notícias	Comentários	Sentenças	Tokens
90	6185	14.547	207.866

Fonte: Elaborado pelos autores

4.1 Anotação de ironia em UGC

A anotação do *corpus* é baseada na definição de opinião proposta por Liu (2012, p. 17), segundo a qual toda opinião é composta por, pelo menos, dois elementos fundamentais: alvo (podendo ser uma entidade, aspecto de uma entidade, um produto, pessoa, organização, marca, evento etc.) e sentimento (que representa uma atitude, opinião ou emoção em que o autor da opinião tem a respeito de um determinado alvo).

Após a compilação do *corpus,* foi iniciado o seu processo de anotação, que segue as seguintes etapas: i) identificação do comentário, autor e data; ii) anotação das opiniões de uma sentença; e iii) anotação da intenção de uma opinião (se é irônica, não irônica ou outro tipo de ironia).

Preliminarmente, criou-se um cabeçalho contendo o número do comentário, autor e data de cada comentário analisado. Dessa maneira, cada comentário está delimitado com as etiquetas <coment id:"xxx"> e </coment>. Seguindo, todo comentário tem um autor que expressa a opinião, o qual é indicado pelas etiquetas <author> e </author>, e tal comentário apresenta uma data no qual foi expresso por um autor e é indicada pelas etiquetas <date> e </date>.

[7] Disponível em: <http://www1.folha.uol.com.br/poder>.

Quadro 1 – Exemplo de cabeçalho do comentário

```
<coment id:"00018">
    <author>Viva</author>
        <date>02/01/2016</date>
            <sentence>O que o Aécio fez eu não sei, mas a Dilma comandou o assalto à
                Petrobrás.</sentence>
</coment>
```

Fonte: Elaborado pelos autores

Considera-se a sentença como a unidade mínima de análise segmentada por ponto-final, exclamação, interrogação ou reticências. Cada sentença foi delimitada com as etiquetas <sentence> e </sentence>.

No entanto, há sentenças em que a pontuação pode indicar uma hesitação ou pausa ou então a reprodução da oralidade. Sentenças com dois pontos, ponto e vírgula, parênteses e travessões ou hífen são consideradas como uma única sentença.

(6) Afinados?… duas pessoas que envergonham a nação!
(7) E dá arrepios pensar na linha de sucessão presidencial: vice-presidente ("afinado"…), presidente do Senado (céus!…), presidente da Câmara (sem comentários).

As opiniões são conteúdos subjetivos encontrados na linguagem, seja ela falada ou escrita, e, com a popularização da internet, essas opiniões passaram a ser mais acessíveis, principalmente pelos estudiosos da linguagem que procuram verificar quais são os desejos, sentimentos e avaliações de cada indivíduo.

Algumas expressões e determinados verbos facilitam a identificação de opiniões durante o processo de leitura e anotação do *corpus*. A Tabela 2 apresenta alguns exemplos:

Tabela 2 – Marcas de opinião observadas no *corpus*

Marcas linguísticas	Exemplos
A meu ver	Se no passado a jurisprudência adotou este entendimento, a meu ver o fez erroneamente.
Em minha opinião	Em minha opinião, nunca deveria ser composto por indicados dos presidentes
Para/Pra mim	Para mim ela só demonstrou ressentimento e inveja
	Pra mim isso é uma confissão de que o partido do governo simplesmente se locupletou, encheu os bolsos.
Sou a favor	Eu sou a favor da saída da atual Presidente e sou forte crítico da emenda da reeleição.
Sou contra	Eu sou contra o *impeachment*, sou a favor da cassação da chapa toda e da convocação de novas leis.
Eu acho que	Eu acho que estou sendo sincero.
Concordo	Concordo que "não devemos pagar o pato".
Quero ver	Quero ver qual o político que pode jogar a primeira pedra sem medo.
Achar	Achei que nunca iria concordar com uma palavra dela.
Pensar	Penso que seria melhor se tivesse morrido.
Creio	Creio que devemos ter o bom senso.
Acreditar	Acredito que não sou o único brasileiro verdadeiramente revoltado
Considerar	Evidentemente considero que estava, e estou, certo.
Entendo	Entendo que o impedimento da atual presidente é necessário.
Imagino	Imagino que você teria argumentos altamente sofisticados.

Fonte: Elaborado pelos autores

As opiniões podem ser expressas explícita ou implicitamente no texto:

a) Opiniões explícitas: são opiniões que expressam diretamente uma avaliação ou sentimento.

(8) Acho que ninguém mais aguenta você no poder.

b) Opiniões implícitas: são opiniões que expressam indiretamente uma avaliação ou um sentimento.

(9) Dilma e sua exagerada arrogância e incompetência que destruiu a economia do país.

Neste artigo, são consideradas opiniões as sentenças subjetivas que carregam alguma avaliação, impressão ou sentimento sobre um determinado alvo. Em (10), Moro é o alvo e "corretamente" é a opinião sobre o alvo. No exemplo (11), "ser mais ridículo" é a opinião sobre o alvo presidente, no qual o mais age como um intensificador de "ridículo".

(10) *Moro* agiu *corretamente*.
(11) Esse *presidente* não tem mais como *ser mais ridículo*!

Após a identificação de uma sentença opinativa, anotou-se a intenção da opinião, podendo ser classificada em não irônica, irônica ou conter outro tipo de ironia. Assim, cada sentença identificada como opinião deve ser individualmente indicada por:

i) Opinião não irônica: uma opinião que não contém mecanismos linguísticos que alternam o seu significado. É indicada pelas etiquetas <opiniao_naoironica>.

Quadro 2 – Anotação de opinião não irônicas

<sentence><opiniao_naoironica>Se o povo votar em candidatos processados ou que tenham tido os nomes envolvidos com as empreiteiras, então vai ser um problema daqueles que votaram.</opiniao></sentence>

Fonte: Elaborado pelos autores

ii) Opinião irônica: baseada nas teorias linguísticas que, apesar de diferirem sobre a sua definição, concordam que a ironia se dá pelo choque entre o significado literal de enunciado e o que se espera do falante. Assim, uma opinião irônica é aquela que há inversão no sentido literal da sentença. Tais opiniões são indicadas pela etiqueta <opiniao_ironica>.

Quadro 3 – Anotação de opinião irônica

<sentence><opiniao_ironica>A Sra. Presidente e tão honesta que vai pro ceu..tadinha.. </opiniao></sentence>

Fonte: Elaborado pelos autores

iii) Outro tipo de ironia: opinião em que não há inversão no sentido literal, no entanto, o texto ainda tem a compreensão literal corrompida. É indicada pela etiqueta <opiniao_outraironica>.

Quadro 4 – Anotação de opinião outro tipo de ironia

<sentence><opiniao_outraironica>Já já começará aquela conversinha de ele é "pelsseguido puliticu" e blá blá blá. </opiniao></sentence>

Fonte: Elaborado pelos autores

Para a anotação do *corpus*, foi utilizada a ferramenta Notepad++ que, além de ser utilizado como bloco de notas, permite editar códigos em diversas linguagens, tais como C, C++, HTML, ASP, Pyhton, JAVA, Pascal, XML, dentre mais de outras quarenta linguagens disponíveis. Além disso, o Notepad++ permite a criação de macros. Basicamente, a macro é uma sequência de ações que pode ser gravada, salva e reproduzida. Para a extração e frequência das ocorrências de opinião irônica, não irônica e outro tipo de ironia, foi utilizado o Unitex (PAUMIER, 2002). O Unitex é um ambiente que permite a manipulação e análise de *corpus* de milhões de palavras. A ferramenta possibilita descrições linguísticas formalizadas através de dicionários eletrônicos e gramáticas representadas por autômatos finitos.

A anotação de pistas linguísticas consiste na identificação de indicadores de ironia no texto. A compreensão da ironia é de certo modo "intuitiva", e outros anotadores podem conversar e discutir sobre qualquer dúvida sobre a tarefa de anotação. Um problema observado é a anotação com um único anotador, pois por ser uma tarefa tão repetitiva é possível existirem falhas que prejudiquem a significância dos dados.

As pistas identificadas envolvem a classificação de possíveis palavras que indicam ironia ou sarcasmo e envolvem anotações tanto na categoria "ironia" quanto em "outro tipo de ironia". Os padrões "não irônico" também passaram por anotação. A decisão dessa tarefa se deu pensando na possível comparação de uma sentença irônica *versus* sentença não irônica, dado que funções assim podem ser úteis para criação de futuras *features* utilizadas em Aprendizado de Máquina. Na Tabela 3, é possível observar as pistas encontradas e suas especificações durante essa seção.

Tabela 3 – Pistas linguísticas para ironia

Padrões	Pistas
P1	aspas
P2	diminutivo
P3	neologismo
P4	Expressão de riso

Fonte: Elaborado pelos autores

Trabalhos para o português como Carvalho et al. (2009, p.2) e Freitas et al. (2012, p.629) entendem as aspas como um possível indicador de ironia. Os exemplos (12) e (13) são casos de sentenças anotadas como irônica e marcadas com P1. Mas é necessário atentar-se para casos de P1 que não são irônicos, como nos exemplos (14) e (15).

(12) Parabéns "honesto" Anastasia.
(13) Tudo culpa do fhc e sua "amiga" de paris!!!

(14) Ele citou o ditado “Quem nunca comeu melado, quando come se lambuza”.
(15) O ministro Barroso, tão logo entrou na Suprema Corte, pontuou: “não existe corrupção do PT, existe corrupção”.

É possível observar que as aspas em contexto irônico geralmente vêm marcando adjetivos, como nos exemplos (12) e (13), marcando ditados e citações diretas, como em (14), e apelidos e contextos jocosos, como em (15).

Em Casati (2017), foram anotadas e analisadas ocorrências de diminutivos no *ComentCorpus*. Os exemplos (16) e (17) mostram um caso de ironia através do diminutivo.

(16) Esse partidinho traidor do Brasil já deveria ter sido varrido.
(17) Por favor deixem o santinho, o comprador da reeleição, em paz!

No processo de anotação, foi observado o uso de “coxinha”, que é usado, frequentemente, de modo pejorativo para referir-se a uma pessoa conservadora e geralmente da classe média. Entretanto, apesar de exprimir um caráter irônico à sentença, considerou-se o termo como um falso diminutivo, uma vez que não é o sufixo *-inha* que carrega essa ironia, mas sim todo um contexto atual que o envolve.

(18) Golpistas e coxinhas querem que o povo se exploda.
(19) Se for pro pau, os coxinhas fogem.

Em relação aos neologismos, a busca por possíveis ocorrências foi feita pelo Unitex (PAUMIER, 2002). Para isso, o *corpus* foi processado pelo *software*, e a Wordlist de palavras que não constavam no dicionário eletrônico foi analisada e anotadas caso fossem de fato um neologismo. Os exemplos (20) e (21) são exemplos de sentenças irônicas marcadas por neologismos como “pixulequeiros” e “Ptrevas”.

(20) Aproveita e tenta levar junto o máximo de bolivarianos pixulequeiros que conseguir.
(21) Essa tentativa da mídia em desvincular o PTrevas do Governo é risível.

Observa-se que é bem comum em UGC e em outras pesquisas de detecção automática de ironia que expressões de riso são fortes indicadores de opiniões irônicas.

(22) Dois caras de pau sem nenhum pingo de vergonha, se acham acima das Leis e ainda ameaçam os Juízes,,,kkkk
(23) falou a super fera! Kkkkkkkk

5 Análise do *corpus*

O *ComentCorpus* foi anotado manualmente por dois anotadores, ambos da área de Linguística, que seguiram diretrizes predeterminadas descritas na seção anterior. Para garanti-las, nas anotações realizadas até o presente momento, foi realizada a verificação da confiabilidade da anotação. Após a anotação do *corpus* e extração dos dados anotados através do Unitex, foi possível comparar a anotação entre os anotadores. A Tabela 4 apresenta o resultado dos da anotação de para anotador referente às opiniões não irônicas (NI), irônicas (I) ou, opiniões que contêm outro tipo de ironia (OI).

Tabela 4 – Dados dos anotadores

	Anotador 1			**Anotador 2**			
Mês	NI	I	OI	NI	I	OI	Opiniões
Janeiro	1296	361	246	1321	397	185	1903
Fevereiro	1269	290	307	1310	268	288	1866
Março	1921	508	311	1897	516	327	2740
Abril	1960	702	299	1966	714	281	2961
Maio	1973	362	321	1969	349	338	2656

Fonte: Elaborado pelos autores

A Tabela 5 mostra o cálculo da porcentagem do resultado da anotação de cada anotador em relação às categorias anotadas.

Tabela 5 – Porcentagem de anotação para cada anotador

	Anotador 1			**Anotador 2**			
Mês	NI	I	OI	NI	I	OI	Opiniões
Janeiro	68.10%	18.97%	12,92%	69.41%	20.86%	9.72%	1903
Fevereiro	68%	15.54%	16.45%	70.20%	14.36%	15.43%	1866
Março	70.10%	18.54%	11.35%	69.23%	18.83%	11.93%	2740
Abril	66.19%	23.70%	10.09%	66.39%	24.11%	9.49%	2961
Maio	74.28%	13.62%	12.08%	74.13%	13.14%	12.72%	2656

Fonte: Elaborado pelos autores

O Gráfico 1 mostra a distribuição das categorias estabelecidas na anotação para cada anotador das 12.126 opiniões anotadas. Percebe-se que há um equilíbrio entre os anotadores, o que mostra uma boa eficiência do processo de anotação. Como pode ser inferido no gráfico abaixo, a categoria opinião não irônica é a mais frequente para ambos anotadores, contabilizando 69,43% e 69,79%, respectivamente. Na sequência, as opiniões irônicas aparecem, respectivamente, com 18,33% e 18,51% das anotações. As opiniões que pertencem ao outro tipo de

ironia são as que mais se distinguem entre os anotadores, mesmo não sendo uma diferença significativa – 12,24% para o Anotador 1 e 11,70% para o Anotador 2.

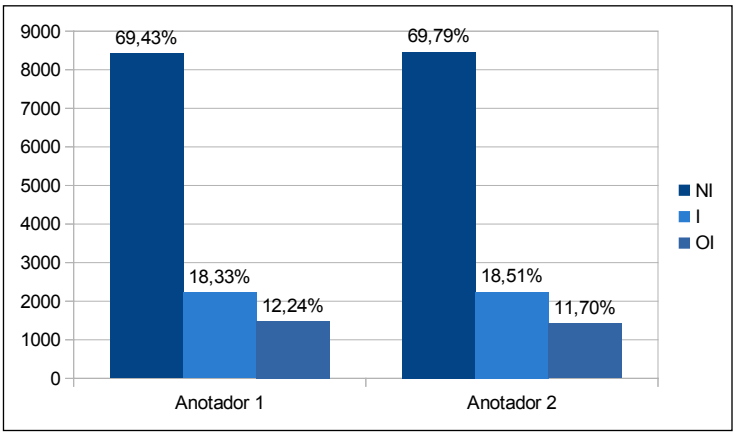


Gráfico 1 – Distribuição das categorias de anotação entre os anotadores
Fonte: Elaborado pelos autores

A Tabela 6 mostra a frequência de ocorrência de cada pista identificada. Percebe-se que a ocorrência de P4 é majoritariamente irônico, seguido pelos diminutivos.

Tabela 6 – Ocorrências de pistas de ironia

Padrões	NI	I	OI	TOTAL
P1	168	234	136	538
P2	42	66	44	152
P3	73	29	48	150
P4	2	26	10	38

Fonte: Elaborado pelos autores

No Gráfico 2, são apresentadas as ocorrências de cada pista linguística nas três categorias de opinião: não irônica, irônica e outro tipo de ironia. As aspas são os dispositivos linguísticos mais frequentes que podem indicar ironia, com 43,49% para opiniões irônicas e 25,28% para outros tipos de ironia. Posteriormente, vêm os casos de diminutivo, com 43,42% de opiniões irônicas e 28,95% de outros tipos de ironia. Seguindo pelas expressões de riso, contam com 68,42% das opiniões irônicas e 26,32% dos outros tipos de ironia. Quanto aos neologismos, apesar de haver casos em que estão em contexto irônico, não é correto afirmar que sejam uma pista de ironia em texto escrito.

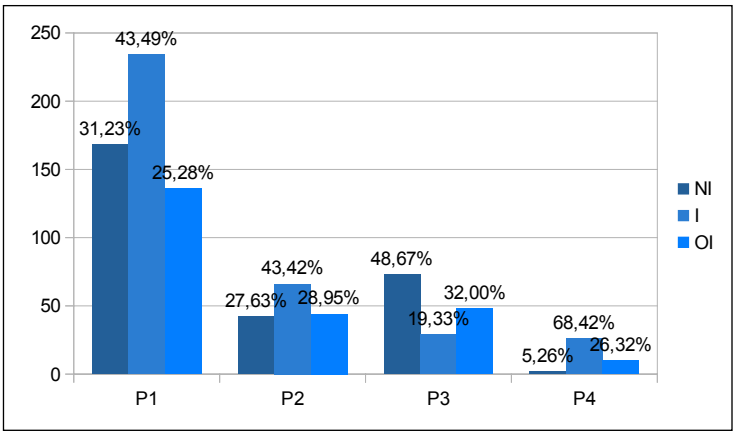


Gráfico 2 – Ocorrências de pistas linguísticas no *corpus*
Fonte: Elaborado pelos autores

Embora as aspas sejam a pista mais frequente, com 538 das ocorrências, a expressão de riso é a pista que mais indica, majoritariamente, uma opinião irônica, pois das 38 ocorrências, 68% estão marcadas em contexto de ironia.

6 Conclusão e trabalhos futuros

As principais dificuldades encontradas estão relacionadas ao processo de compreensão dos textos. Foi possível perceber uma grande dificuldade na identificação de expressões de linguagem figurada no texto, em específico, a ironia e o sarcasmo. Dessa forma, é necessária uma leitura atenta do conteúdo comunicado de cada texto, o que torna a tarefa de anotação complexa e demorada.

A subjetividade, muitas vezes, atrapalha ou não permite uma interpretação por falta de informações ou por tratar-se de um contexto estritamente específico. Além disso, os textos são mal estruturados, o que também dificulta a apreensão do significado. Justamente, a falta de acesso ao autor e às suas características podem dificultar a compreensão de um enunciado irônico; certamente isso é essencial para entender o contexto em que aquela sentença ocorre.

Futuramente, busca-se seguir o processo de anotação do *corpus* das opiniões e também a anotação das expressões irônicas e sarcásticas, focando em mecanismos linguísticos que indiquem ironia. Através dessa identificação, serão extraídas as listas dessas expressões a fim de identificar os padrões mais característicos da ironia.

Acredita-se que esta pesquisa possa contribuir para os estudos voltados para a Análise de Sentimentos e, em especial, para a anotação de *corpus* e o desenvolvimento de novos métodos de identificação das opiniões, em especial as opiniões irônicas de um falante.

Referências

ATTARDO, S. Irony as relevant inappropriateness. *Journal of pragmatics*, v. 32, n. 6, p. 793-826, 2000.

BECKER, K.; TUMITAN, D. Introdução à Mineração de Opiniões: conceitos, aplicações e desafios. In: FERREIRA, J. E. (Org.). *Lectures of the 28th Brazilian Symposium on Databases*. 1. ed. Pernambuco: CIN – UFPE, 2013, p. 27-52

BISOGNIN, T. R. *Sem medo do internetês*. Porto Alegre: AGE, 2009.

CARVALHO, P.; SARMENTO, L.; SILVA, M. J.; OLIVEIRA, E. et al. Clues for detecting irony in user-generated contents: oh...!! "it's so easy". In: INTERNATIONAL CIKM WORKSHOP ON TOPIC-SENTIMENT ANALYSIS FOR MASS OPINION, 1., Hong Kong, 2009. *Proceedings...* Hong-Kong: ACM, 2009, p. 53-56.

CARVALHO, P.; SARMENTO, L.; TEIXEIRA, J.; SILVA, M. J. Liars and saviors in a sentiment annotated *corpus* of comments to political debates. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES: SHORT PAPERS, 49. *Proceedings...* Vol. 2. Portland: Association for Computational Linguistics, 2011, p. 564-568.

CASATI, S. Diminutivos como indicadores de ironia em um corpus do português brasileiro. Monografia (Bacharelado em Linguística). Centro de Educação e Ciências Humanas, Universidade Federal de São Carlos. 2017. p. 1-50.

COUNCILL, I. G.; MCDONALD, R.; VELIKOVICH, L. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. Proceedings of the workshop on negation and speculation in natural language processing. Association for Computational Linguistics. p. 51-59. 2010.

CURCÓ, C. Irony: Negation, echo and metarepresentation. *Lingua*, v. 110, n. 4, p. 257-280, 2000.

_____. Lenguaje figurado y teoría de la mente. *Estudios de Lingüística Aplicada*, v. 30, 1999.

DAVIDOV, D.; TSUR, O.; RAPPOPORT, A. Enhanced sentiment learning using twitter hashtags and smileys. In: Proceedings of the 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, p. 241-249. 2010.

ESULI, A.; SEBASTIANI, F. *SENTIWORDNET: A high-coverage lexical resource for opinion mining*. Pisa: Institute of Information Science and Technologies (ISTI), 2006, p. 1-26.

FREITAS, C. Sobre a construção de um léxico da afetividade para o processamento computacional do português. *Revista Brasileira de Linguística Aplicada*, Belo Horizonte, v. 13, n. 4, 2013.

FREITAS, C.; MOTTA, E.; MILIDÚ, R.; CESAR, J. *Vampiro que brilha… rá!* Desafios na anotação de opinião em um *corpus* de resenhas de livros. São Carlos: Encontro de Linguística de *Corpus*, 2012, p. 1-13.

DE OLIVEIRA, L. P. Linguística de *Corpus*: teoria, interfaces e aplicações. *Revista do Programa de Pós-Graduação em Letras da UERJ*, Rio de Janeiro, v. 16, n. 24, 2009.

FELLBAUM, C. *WordNet*: an eletronic lexical database. Cambridge: [s/n], 1998.

GIBBS, R. W. *The poetics of mind*: figurative thought, language, and understanding. Cambridge: Cambridge University Press, 1994.

GIBBS, R. W.; COLSTON, H. L. Irony in language and thought: a cognitive science reader. Mahwah: Lawrence Erlbaum Associates Publishers, 2007.

GIORA, R. On irony and negation. *Discourse processes*, v. 19, n. 2, p. 239-264, 1995.

GRICE, H. P. *Logic and conversation*. In: COLE, P.; MORGAN, Jerry L. *Syntax and Semantics*. Vol. 3: Speech Acts. New York: Academic Press, 1975, p. 41-58.

GONZÁLEZ-IBÁNEZ, R.; MURESAN, S.; WACHOLDER, N. Identifying sarcasm in Twitter: a closer look. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES: SHORT PAPERS, 49. *Proceedings...* Vol. 2. Portland: Association for Computational Linguistics, 2011, p. 581-586.

HOGENBOOM, A.; BALL, D.; FRASINCAR, F.; BAL, M.; JONG, F. de; KAYMAK, U. Exploiting emoticons in sentiment analysis. In: ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING, 28., Coimbra, 2013. *Proceedings...* [s/l]: ACM, 2013, p. 703-710.

HOVY, E.; LAVID, J. Towards a 'science' of *corpus* annotation: a new methodological challenge for *Corpus* Linguistics. *International Journal of Translation*, v. 22, n. 1, p. 13-36, 2010.

MARCUS, M. P.; MARCINKIEWICZ, M. A.; SANTORINI, B. Building a large annotated *corpus* of English: The Penn Treebank. *Computational Linguistics*, v. 19, n. 2, p. 313-330, 1993.

MUECKE, D. C. *Ironia o irônico*. 1. ed. São Paulo: Editora Perspectiva, 1995.

IDE, N.; PUSTEJOVSKY, J. *Handbook of Linguistic Annotation*. Dordrecht: Springer, 2017.

KENNEDY, G. *An introduction to* Corpus *Linguistics*. New York: Routledge, 2014.

KREUZ, R. J.; GLUCKSBERG, S. How to be sarcastic: the echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, v. 118, n. 4, p. 374, 1989.

KREUZ, R. J.; CAUCCI, G. M. Lexical influences on the perception of sarcasm. In: WORKSHOP ON COMPUTATIONAL APPROACHES TO FIGURATIVE LANGUAGE, 2007, Rochester. *Proceedings...* New Brunswick: Association for Computational Linguistics, 2007, p. 1-4.

KUMON-NAKAMURA, S.; GLUCKSBERG, S.; BROWN, M. How about another piece of pie: the allusional pretense theory of discourse irony. *J Exp Psychol Gen*, 124(1), p. 3-21, 1995.

LEECH, G. *Corpora* and theories of linguistic performance. In: SVARTVIK, J. (Ed.). *Directions in* Corpus *Linguistics*. Berlin: Mouton de Gruyter, 1992, p. 105-122.

LIU, B. Sentiment Analysis and Subjectivity. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). *Handbook of natural language processing*. Boca Raton: CRC Press, 2010, p. 627-666.

_____. Sentiment analysis and Opinion Mining. *Synthesis lectures on human language technologies*, v. 5, n. 1, p. 1-167, 2012.

OLIVEIRA, L. S.; CAMPOS, G. O.; DA SILVA, R. S. Mineração de dados e análise de opinião em redes sociais – Um estudo de caso sobre as Olimpíadas 2012 utilizando o Twitter, 2012.

PAUMIER, S. *Unitex*. Manuel d'utilisation. 2002.

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, v. 2, n. 1-2, p. 1-135, 2008.

REYES, A. From humor recognition to irony detection: the figurative language of social media. *Data & Knowledge Engineering*, v. 74, p. 1-12, 2012.

REYES, A.; ROSSO, P.; VEALE, T. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, v. 47, n. 1, p. 239-268, 2013.

RILOFF, E.; QADIR, A.; SURVE, P.; DE SILVA, L.; GILBERT, N.; HUANG, R. Sarcasm as contrast between a positive sentiment and negative situation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. p. 704-714. 2013

SANTOS, D. Disponibilização de *corpora* de texto através da WWW. In: MOTA, M. A.; MARRAFA, P. *Linguística Computacional*: Investigação Fundamental e Aplicações. Lisboa: Colibri, 1999.

SALEM JR, J. A. New Dictionary of the History of Ideas. *Reference & User Services Quarterly*, v. 45, n. 1, p. 86-87, 2005.

SEARLE, J. R. *The philosophy of language*. London: Oxford University Press, 1971.

SILVA, N. R.; LIMA, D.; BARROS, F. Sapair: Um processo de Análise de Sentimento no nível de característica. In: INTERNATIONAL WORKSHOP ON WEB AND TEXT INTELLIGENCE (WTI'12), 4., Curitiba, 2012. *Proceedings*... [s/l]: [s/n], 2012.

SINCLAIR, J. *Corpus* and text-basic principles. In: WYNNE, M. (Ed.). *Developing linguistic* corpora*: a guide to good practice*. Oxford: Oxbow Books, 2005, p. 1-16.

SMARSARO, A. *Descrição e formalização de palavras compostas do português do Brasil para elaboração de um dicionário eletrônico*. Tese (doutorado). Programa de Pós-Graduação em Estudos da Linguagem, Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2004.

SPERBER, D.; WILSON, D. Irony and the use-mention distinction. *Philosophy*, v. 3, p. 143-184, 1981.

_____. *Relevance*: Communication and cognition. Cambridge: Harvard University Press, 1986.

UTSUMI, A. A unified theory of irony and its computational formalization. In: COLING 96. Proceedings of the 16th conference on computational linguistics. 1996.

VANIN, A. A. et al. Some clues on irony detection in tweets. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 22., 2013. *Proceedings*... New York: ACM, 2013, p. 635-636.

VEALE, T.; HAO, Y. Detecting Ironic Intent in Creative Comparisons. In: EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE, 19., Lisboa, 2010. *Proceedings*... Amsterdam: IOS Press, 2010, p. 765-770.

WIEBE, J.; WILSON, T.; CARDIE, C. Annotating expressions of opinions and emotions in language. Language resources and evaluation. *Journal of Pediatric Psychology*. p. 167-178. 2005.

O discurso dos deputados na votação do *impeachment*: a LC combinada à ACD

Deputies' speeches in the impeachment drive: a CL approach to CDA

Rozane Rodrigues Rebechi

Resumo: Em abril de 2016, o Congresso Nacional votou a admissibilidade do processo de *impeachment* da então presidente Dilma Rousseff. Antes mesmo do término da sessão, a mídia e as redes sociais publicaram comentários sobre a escolha lexical dos congressistas ao justificarem seus votos, em geral associando palavras relacionadas a Deus, família e nação com os discursos pró-*impeachment*. Utilizando a metodologia da Linguística de *Corpus*, combinada à Análise Crítica do Discurso, este capítulo visa à análise das transcrições das falas dos deputados a fim de se confirmar ou não a impressão do público e da imprensa. Os resultados apontaram para uma escolha lexical estatisticamente coincidente das palavras mencionadas, especialmente nos dois grupos com maior número de votantes.

Palavras-chave: *Impeachment*. Discurso político. Linguística de *Corpus*. Análise Crítica do Discurso.

Rozane Rodrigues Rebechi – Professora adjunta na Universidade Federal do Rio Grande do Sul, doutora pela Universidade de São Paulo – rozanereb@gmail.com.

Abstract: In April 2016, Brazil's Lower House of Congress voted the impeachment drive of then president Dilma Rousseff. The session was not even over when internet users and mass media started commenting on the vocabulary the legislators used, in most cases associating words related to God, family and nation to the pro-impeachment speeches. By combining *Corpus* Linguistics and Critical Discourse Analysis, this chapter aims at analyzing the transcripts of the deputies' talks in order to confirm or not the impression of the general public and the media. Results pointed to a statistically similar lexical choice of the aforementioned words, especially in the two groups with the largest number of voters.

Keywords: Impeachment. Political Discourse. *Corpus* Linguistics. Critical Discourse Analysis.

1 Introdução

No dia 17 de abril de 2016, a Câmara dos Deputados votou a admissibilidade do processo de *impeachment* da então presidente do Brasil, Dilma Rousseff. Dos 513 representantes do Poder Legislativo, 367 votaram a favor do processo, 137 contra, sete se abstiveram e dois faltaram à sessão, resultando no prosseguimento do caso para o Senado. Além de escolher entre uma das três possibilidades de voto – sim, não e abstenção –, a grande maioria dos votantes optou por justificar suas escolhas, utilizando os dez segundos ao microfone a que tinham direito[1]. Antes mesmo do término da votação, que durou mais de cinco horas, as redes sociais e a mídia passaram a postar comentários de historiadores, cientistas políticos, jornalistas e público em geral, que calcularam e discutiram as palavras mais recorrentes nessas falas.

Deus, palavras relacionadas à família, à nação, ao combate à corrupção, entre outras, também deram vazão a inúmeros *memes* que viralizaram nas redes sociais. A fim de legitimar a escolha do voto, diversos congressistas iniciaram as falas com o vocábulo "pelo(a)(s)" (contração da preposição "per" + artigo definido), seguido de justificativa, como em "Pela minha família", "Por Deus", "Pelo meu país" etc. Com o intuito de ridicularizar essa recorrência, os internautas criaram *memes* com o mesmo início, "defendendo" interesses diversos, o que resultou em apelos jocosos como "Pelo Wi-Fi grátis", "Pelo emagrecimento fácil, sem dieta e academia", "Pela volta da Caverna do Dragão" etc., sempre finalizados com "eu voto sim".

A mídia também se pronunciou a respeito dos discursos dos deputados. O *site* UOL Notícias publicou uma matéria com o título *"Minha", "meus", "família"...: a lista das palavras mais citadas na sessão do impeachment*[2], em que é feita

[1] Muitos deputados extrapolaram esse tempo e continuavam se pronunciando mesmo mediante a tentativa de interrupção pelo presidente da sessão.

[2] Disponível em: https://noticias.uol.com.br/politica/ultimas-noticias/2016/04/20/familia-e-democracia-sao-citadas-mais-de-100-vezes-por-deputados-veja-outras.htm. Acesso em: 10 out. 2017.

uma análise dos discursos dos parlamentares, elencando os vocábulos mais citados. Os trechos reproduzidos a fim de ilustrar a recorrência de palavras relacionadas à família e a Deus foram extraídos dos discursos pró-*impeachment*, sendo que apenas um excerto transcreve uma fala de votante contrário ao processo, na qual critica a menção a Deus por deputados que votaram a favor da continuidade do processo.

Outras matérias foram ainda mais incisivas. Em artigo intitulado *O que os discursos dos deputados pró-impeachment revelam sobre a construção da nossa democracia*, publicado em *Carta Capital*[3], o historiador Luan Aiuá denomina "*show* de horrores" os discursos favoráveis ao processo, chamando especial atenção à "tríade família, Deus e nação", que teria permeado as falas desses "conservadores". Já *Congresso em Foco* publicou matéria com a chamada *Deputados citaram "Deus" 59 vezes na votação do impeachment*[4,5]. O artigo reproduz uma passagem do *blog* do teólogo Leonardo Boff, na qual afirma: "Dezenas de parlamentares da bancada evangélica fizeram claramente discursos de tom religioso e invocando o nome de Deus. E todos, sem exceção, votaram pelo impedimento". Com a manchete *Deus derruba a presidenta do Brasil*, seguida do *lead Deputados justificam seus votos em Deus, na moralidade e a família: o motivo real da votação é esquecido*, a matéria publicada no jornal *on-line El País*, assinada por María Martín, compara as justificativas dos votos pelo *impeachment* a um programa de auditório[6].

Diante de afirmações tão contundentes em relação à escolha lexical dos deputados durante a votação, surgiu o interesse em realizar uma pesquisa linguística mais aprofundada, não respaldada apenas pela frequência de palavras, com o intuito de confirmar – ou não – a representação dessas falas nas publicações em redes sociais e na mídia.

Por meio da combinação entre a metodologia subjacente à Linguística de *Corpus* (LC) e à Análise Crítica do Discurso (ACD), este estudo visa à análise das falas dos deputados na votação do processo de *impeachment*, a partir do levantamento quantitativo de palavras-chave e colocados, possibilitado por ferramentas computacionais.

[3] Disponível em: http://justificando.cartacapital.com.br/2016/04/19/o-que-os-discursos-dos-deputados-pro-impeachment-revelam-sobre-a-construcao-de-nossa-democracia/. Acesso em: 10 out. 2017.
[4] Disponível em: http://congressoemfoco.uol.com.br/noticias/deputados-citaram-%E2%80%9Cdeus %E2%80%9D-59-vezes-na-votacao-do-impeachment/. Acesso em: 10 out. 2017.
[5] Segundo lista de palavras do *corpus*, a palavra "Deus" ocorre 56 vezes.
[6] Disponível em: https://brasil.elpais.com/brasil/2016/04/18/politica/1460935957_433496.html. Acesso em: 17 out. 2017.

2 Linguística de *Corpus* aplicada à análise de discurso

Antes de tudo, faz-se relevante explicitar o uso da palavra "discurso" neste estudo. Se, em inglês, a polissemia de *discourse* pode ser problemática, uma vez que pode suscitar diferentes acepções – tipos de linguagens (por exemplo, discurso midiático, político etc.), linguagem em uso, unidade de língua além da sentença, entre outras (ver BAKER, 2006, para uma análise dos conceitos subjacentes a essa palavra) –, em português o problema se agrava, pois a palavra pode ainda se referir à exposição oral feita em público, ou seja, poderia ser usada para se referir às falas dos deputados na sessão da câmara. Contudo, a fim de evitar mal-entendidos, neste estudo, como em Baker (2006, p. 2), procuramos restringir "discurso" à acepção Foucaultiana, qual seja, a de um sistema de sentenças utilizadas na construção de um objeto.

Na presente pesquisa, dois tipos de discursos são analisados: (i) o discurso da mídia, que relata as falas dos deputados durante sessão de votação do *impeachment*, e (ii) as falas propriamente ditas. Em seguida, apresenta-se uma breve explanação dessas duas formas de discurso, conforme entendidas neste trabalho.

2.1 Análise Crítica do Discurso e mídia

Para Fairclough (1989), a linguagem deve ser entendida como meio de ação, uma vez que, quer seja falada ou escrita, constitui atos de fala, como promessa, declaração, advertência etc., e envolve relações de poder, nem sempre explícitas. Sobre o poder no discurso da mídia, afirma:

> O discurso dos meios de comunicação de massa [televisão, rádio, cinema, jornais] é interessante porque a natureza das relações de poder que ele estabelece muitas vezes não é clara, e há razões para entendê-lo como envolvendo relações *escusas* de poder"[7] (FAIRCLOUGH, 1989, p. 49, grifo do autor).

Enfatiza, ainda, que esse poder é construído a partir de sistematizações, ou seja, de repetições de informação nas atividades midiáticas:

> Um único texto por si só é bastante insignificante: os efeitos do poder das mídias são cumulativos, funcionando por meio da repetição de formas específicas de se lidar com causalidade e agência, formas específicas de posicionamento do leitor, e assim por diante[8] (FAIRCLOUGH, 1989, p. 54).

---

[7] "[…] *mass media discourse [television, radio, film, newspapers] is interesting because the nature of the power relations enacted in it is often not clear, and there are reasons for seeing it as involving* hidden *relations of power.*"

[8] "*A single text on its own is quite insignificant: the effects of media power are cumulative, working through the repetition of particular ways of handling causality and agency, particular ways of positioning the reader, and so forth*."

Em análises de discurso realizadas em jornais para a identificação de ideologias implícitas, Fairclough (1995) defende o conceito de "representação de discurso", uma vez que, em geral, nas publicações não se reporta de forma transparente o que foi falado ou escrito; o que se observa é uma tomada de decisão a partir de uma interpretação, e posterior representação, da informação que se deseja transmitir. Distingue, portanto, o discurso primário [*primary discourse*], ou seja, o relato propriamente dito, e o discurso secundário [*secondary discourse*], ou a representação do discurso, que é permeada por interpretação. Nesse sentido, Fairclough (1995) chama a atenção para dois "mitos": o primeiro seria a crença de que a mídia é um "espelho" da realidade, e o outro, que a própria realidade seria "transparente", de forma que poderia ser "lida" sem mediação ou interpretação. Sendo assim, entende a representação do discurso na mídia como um processo de grande importância social.

Apesar de o público não ser passivo, uma vez que o significado é criado a partir da interação entre o texto e seus leitores/ouvintes, o jornalismo os influencia com a produção de novos discursos ou com a reformulação de outros, já existentes (BAKER, 2006, p. 72). Assim, é umas das funções da ACD evidenciar (i) de que forma o grupo que detém o poder controla o discurso e (ii) como esse discurso influencia os grupos menos poderosos (VAN DIJK, 2001, apud BAKER 2006, p. 73-74). Nesta pesquisa, o discurso primário é constituído pelas falas transcritas dos deputados, e o secundário, pela representação dessas falas pela mídia, esta que, nessa relação, seria a detentora do poder, e, portanto, capaz de influenciar o leitor.

2.2 Discurso político

Para Chilton (2004, p. 3), "política" pode ser definida como "[...] luta pelo poder entre aqueles que buscam reivindicá-lo e mantê-lo e aqueles que resistem a ele"[9] e "[...] cooperação, como as práticas e instituições das quais a sociedade se utiliza para resolver conflitos de interesse em relação a questões financeiras, influência, liberdade etc."[10]. Linguagem e política estão intimamente ligadas, pois é por meio da linguagem que os atores políticos produzem efeitos. Durante a comunicação, os interlocutores sempre supõem receber tanto informações verdadeiras quanto falsas, e é justamente a expectativa do recebimento de informações verdadeiras que possibilita que o agente do discurso engane ou distorça a verdade.

A partir da análise de entrevistas com políticos, além de discursos políticos diversos, Chilton (2004) conclui que, para imprimir veracidade aos seus

[9] "[…] *a struggle for power, between those who seek to assert and maintain their power and those who seek to resist it*."
[10] "[…] *cooperation, as the practices and institutions that a society has for resolving clashes of interest over money, influence, liberty, and the like*."

enunciados, os emissores fazem uso de "evidências" como forma de legitimar seu discurso. Citando Fetzer (2002), conclui: "[...] do ponto de vista político, o que interessa é se o falante tem 'credibilidade'"[11] (CHILTON, 2004, p. 32).

Várias estratégias linguísticas são utilizadas pelos políticos com a finalidade de se "aproximar" do interlocutor e convencê-lo da "veracidade" de suas afirmações. Apelar para o patriotismo, para a causa dos menos favorecidos e para a união são algumas delas. Também usual é o político usar repetidamente o pronome em primeira pessoa do plural – "nós" –, assim como o pronome adjetivo possessivo relacionado a ele – "nosso(a)". Entre o uso estratégico da língua pelos políticos, Chilton (2004) identifica, ainda, duas vertentes: legitimação e deslegitimação. A primeira envolve pressupostos sobre a vontade dos eleitores, princípios ideológicos em geral, atitudes carismáticas e autorrepresentação positiva. Já a segunda está relacionada à representação dos opositores, em geral retratados com características negativas, por meio de acusações e ofensas.

O controle da informação pressupõe o controle do discurso, este que pode ser qualitativo ou quantitativo. O controle qualitativo pode ser realizado por meio do uso de informações falsas, omissões e negações. Já o quantitativo costuma se favorecer das generalizações. Tais estratégias puderam ser observadas nas falas dos deputados aqui analisadas, conforme veremos adiante.

2.3 LC aplicada à ACD

A ACD pode ser entendida como "um movimento acadêmico, uma forma de se fazer análise do discurso a partir de uma perspectiva crítica, em geral focada em conceitos teóricos, tais como poder, ideologia e dominação" (BAKER et al., 2008, p. 273)[12]. A área sofre críticas por dois aspectos. Em primeiro lugar, devido à sua suposta fraqueza metodológica, uma vez que, tradicionalmente, se ocupa da análise apenas de fragmentos de textos. Em segundo lugar, devido à subjetividade de seus resultados, que podem decorrer de ideias preconcebidas do analista (cf. CHENG, 2013). A LC, área de estudo que enfoca um conjunto de métodos para o estudo da língua em uso (McENERY; HARDIE, 2012, p. 1), também é criticada. Muitos acreditam que se resume a análises puramente quantitativas, pautadas unicamente por dados estatísticos. De fato, as duas áreas em questão podem apresentar lacunas, que não nos cabe tratar aqui. Acreditamos, contudo, que, ao ajudar na identificação de amostras extraídas semiautomaticamente, a partir de textos autênticos, uma metodologia subjacente à investigação da língua em uso possibilita uma análise mais objetiva do que aquela realizada a partir de

[11] "[...] *what matters, from a political point of view, is whether the speaker has 'credibility'.*"
[12] "[…] *an academic movement, a way of doing discourse analysis from a critical perspective, which often focuses on theoretical concepts such as power, ideology and domination.*"

excertos selecionados aleatoriamente, ou, ainda mais preocupante, identificados propositadamente, a fim de se "confirmarem" hipóteses prévias.

Apesar de pesquisas combinando a metodologia da LC aplicada à ACD não serem novidades, essa associação ainda não resulta em grandes números de pesquisas, ao menos se compararmos com a contribuição que a LC desempenha em áreas como lexicografia, descrição gramatical e registro (cf. PARTINGTON, 2004), tradução (BAKER, 1993; ZANETTIN, 2012), terminologia (PEARSON, 1998) e ensino de línguas (CARTER et al., 2007), para citar apenas algumas. Biber et al. (1998, p. 106) também observam que "[...] embora quase todos os estudos em discurso se baseiem na análise de textos reais, não se trata particularmente de investigações baseadas em *corpus* [...]"[13], pois não usam métodos quantitativos como ponto de partida. Para Sanderson (2008), a escassez de pesquisas que aliam LC e Discurso tem explicação:

> A combinação da Linguística de *Corpus* com a metodologia analítica do Discurso é incomum. No passado, os linguistas de *corpus* não se interessaram particularmente pelo Discurso, preferindo concentrar-se na análise lexical e morfossintática. Da mesma forma, os analistas do discurso raramente se utilizavam de *corpora*, preferindo métodos como a introspecção, a elicitação e a coleta não sistemática de evidências anedóticas. (SANDERSON, 2008, p. 59)[14]

O pesquisador conclui, portanto, que se trata de via de mão dupla: se, de um lado, o Discurso é preterido na LC, de outro, os estudos em discurso não costumam adotar a LC como metodologia. Tal associação, contudo, tem se mostrado produtiva. Nesse sentido, merecem destaque pesquisas como as de Partington (2004), Partington et al. (2004), Baker (2006, 2012), Baker et al. (2008), entre outras.

Para Baker e McEnery (2005), a análise semiautomática de *corpora* pode desempenhar papel importante, uma vez que leva os pesquisadores a identificar padrões em textos autênticos com maior objetividade e ajuda a enfatizar padrões de associação – colocações – que, em geral, extrapolam a capacidade interpretativa realizada por meio de leitura sequencial de textos. Neste estudo, buscamos o suporte da Linguística de *Corpus* para evidenciar padrões linguísticos nas falas dos deputados e sua representação na construção de discursos da mídia em massa e dos usuários de redes sociais.

---

[13] "[...] *although nearly all discourse studies are based on analysis of factual texts, they are not typically corpus-based investigations* [...]."

[14] "*The combination of corpus linguistic with discourse analytic methodology is unusual. In the past, corpus linguists have not been particularly interested in discourse, preferring to concentrate on lexical and morphosyntactic analysis. Similarly, discourse analysts have seldom worked with* corpora*, preferring methods such as introspection, elicitation and the unsystematic collection of anecdotal evidence*."

3 Metodologia

A fim de analisar as falas dos deputados no processo de *impeachment*, realizamos um levantamento semiautomático, possibilitado por ferramentas computacionais. Nesse levantamento, utilizamos, especialmente, listas de palavras-chave, colocados e linhas de concordância. Em seguida, detalharemos a compilação do *corpus* e as análises quantitativas e qualitativas realizadas.

3.1 O corpus de estudo

A análise de *corpora* orais, produzidos em contextos naturais, é escassa, se comparada à de textos escritos. Uma metodologia baseada em LC pressupõe que o *corpus* esteja em formato eletrônico, para ser, a princípio, analisado quantitativamente, por meio de ferramenta computacional. Para tanto, textos orais devem ser transcritos, e esse processo demanda tempo e, muitas vezes, investimento financeiro (cf. PARTINGTON, 2004; BAKER, 2006). A depender da pesquisa, a transcrição da fala deve considerar as características interacionais, como entonação, mudança de turno, pausa etc., como se observa em Prado (2015), que analisa interações entre pilotos e controladores de tráfego aéreo em comunicações radiofônicas realizadas durante situações anormais. Nesse caso, a pesquisadora utilizou codificações para marcar quebras prosódicas, hesitações e outras características da interação oral, processo extremamente trabalhoso – para transcrever um minuto de áudio foram necessários de trinta a quarenta minutos. Neste estudo, que visa a analisar a escolha lexical dos deputados durante seus votos no processo de *impeachment*, tal marcação não se mostrou necessária. Convém mencionar, contudo, que alguns deputados leram suas falas. Nesse caso, é esperado que o uso de marcadores discursivos, por exemplo, seja menor do que em textos falados espontaneamente (BAKER, 2006).

Desde 2007 a Coordenação de Histórico de Debates do Departamento de Taquigrafia da Câmara dos Deputados gerencia a *webpage Escrevendo a História*, que torna acessíveis "discursos memoráveis" proferidos desde 1946 em sessões no Congresso e na Câmara. Esse material pode ser disponibilizado em formato *.pdf no Diário da Câmara ou mesmo reproduzido em áudio[15]. Em geral, a íntegra das sessões é liberada algumas horas depois de seu encerramento, após a revisão e redação final das notas taquigráficas. Vale ressaltar que o método de transcrição realizado pela Câmara pode ser denominado "idealizado", uma vez que reproduz a língua escrita padrão. Também desconsidera hesitações, pausas preenchidas, repetições, entre outros marcadores próprios da língua oral. Contudo, como

[15] Informações obtidas do *site*: <http://www2.camara.leg.br/atividade-legislativa/plenario/discursos/escrevendohistoria>. Acesso em: 10 out. 2017.

já mencionado, o interesse desta pesquisa recai na escolha lexical. Sendo assim, marcadores de oralidade não se mostraram essenciais aqui.

Nesta pesquisa, utilizamos a transcrição das falas dos deputados durante a sessão que votou o processo de *impeachment* de Dilma Rousseff, conforme disponibilizada pela Câmara em formato de planilha do Google[16], formada por seis colunas: (i) nome do(a) deputado(a), (ii) partido, (iii) Estado, (iv) voto, (v) gênero e (vi) transcrição da fala de cada um. Consideramos para a análise os textos da sexta coluna, ao passo que o conteúdo das outras cinco foi mantido no cabeçalho, conforme ilustrado a seguir (Quadro 1). Os textos, salvos em formato *.txt, a fim de serem inicialmente processados por meio da ferramenta Word Smith 6 (WS) (SCOTT, 2012), foram subdivididos em três *subcorpora*, de acordo com a modalidade do voto – sim, não e abstenção. Vale ressaltar que as informações guardadas no cabeçalho (*header*) possibilitam outras pesquisas, por exemplo, sobre as palavras significativamente mais recorrentes nas falas de representantes de cada partido, se houve diferença entre as falas de homens e mulheres etc. O Quadro 1 exemplifica o modelo de cabeçalho utilizado:

Quadro 1 – Modelo de cabeçalho

```
<header>
<evento> Sessão de votação para aprovação do processo de impedimento da câmara dos deputados </evento>
<data> 17 de abril de 2016 </data>
<deputado> Abel Mesquita Jr </deputado>
<partido> DEM </partido>
<estado> RR </estado>
<voto> SIM </voto>
<gênero> M </gênero>
</header>

<transcrição> Roraima, verás que o filho teu não foge à luta! O povo brasileiro merece respeito! Por um Brasil com justiça, igualdade social e sem corrupção, por uma Roraima desacorrentada, para que possamos exercer o direito constitucional de ir e vir e por todas as famílias roraimenses, eu voto sim, Sr. Presidente.
</transcrição>
```

Fonte: Elaborado pela autora

[16] A íntegra das falas dos deputados utilizadas nesta pesquisa pode ser obtida em formato de Planilha Google em: <https://docs.google.com/spreadsheets/d/1rTFC9kbPqeHWrr2JalcvehgRXdAu03BwyBTgTL6oFC8/edit#gid=177016461>. Acesso em: 28 nov. 2017.

A Tabela 1 resume o conteúdo do *corpus* de estudo:

Tabela 1 – *Corpus* de estudo

	Nº textos	**Nº palavras (*tokens*)**
SIM	367	19.249
NÃO	137	7.836
ABSTENÇÃO	7	299
Total	**511**	**27.384**

Fonte: Elaborado pela autora

O número de votos, e, consequentemente, o número de palavras é bastante discrepante entre as modalidades de voto, o que pode ficar ainda mais claro por meio do Gráfico 1:

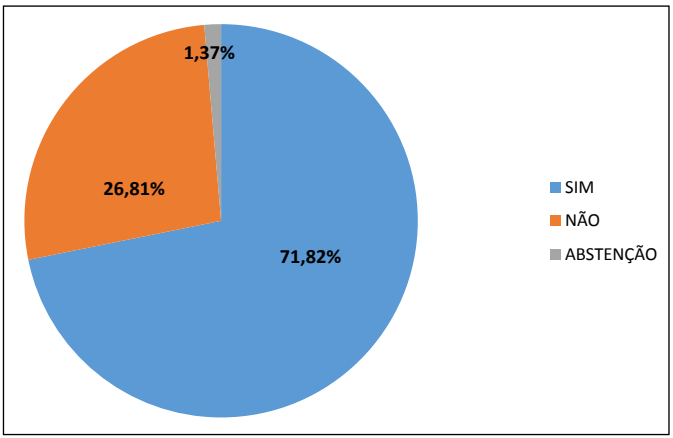


Gráfico 1 – Proporção entre os votos nas três modalidades
Fonte: Elaborado pela autora

Portanto, uma análise quantitativa baseada simplesmente em recorrência de itens lexicais, como aquelas publicadas na mídia e citadas anteriormente, não se mostraria adequada para o estudo das falas dos deputados. Assim, procedemos com o levantamento das palavras-chave, ou seja, palavras significativamente mais recorrentes no *corpus* de estudo, quando comparadas a um *corpus* de referência. Neste estudo, utilizamos como *corpus* de referência textos da língua geral[17]. Por meio da função *KW Database*, do WS, foi possível identificar as palavras-chave

[17] O *corpus* de referência utilizado é uma parte (aproximadamente 2 milhões de palavras) do *Lácio-Ref*, *corpus* que faz parte do projeto Lácio-*Web* (disponível em: <http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm>).

recorrentes nas três modalidades de voto (coluna *Texts*) – neste caso, 100% –, a que denominamos palavras-chave-chave (coluna *KW*). A coluna *Overall Freq.* apresenta o número total de ocorrências da palavra no *corpus* de estudo, conforme apresentado na Tabela 2 em ordem decrescente de frequência:

Tabela 2 – Palavras-chave-chave nos três *subcorpora*

N	KW	Texts	%	Overall Freq.	No. Ass.
1	voto	3	100	543	129
2	Presidente	3	100	494	129
3	eu	3	100	441	129
4	meu	3	100	435	129
5	Sr	3	100	384	129
6	Brasil	3	100	320	129
7	minha	3	100	312	129
8	contra	3	100	157	129
9	país	3	100	151	129
10	*impeachment*	3	100	131	129
11	respeito	3	100	109	129
12	todos	3	100	91	129
13	me	3	100	73	129
14	corrupção	3	100	66	129
15	partido	3	100	62	129
16	favor	3	100	38	129
17	porque	3	100	36	129
18	votar	3	100	36	129
19	vou	3	100	27	129
20	Bahia	3	100	24	129

Fonte: Elaborado pela autora

É preciso enfatizar que o *corpus* de estudo não recebeu etiquetagem morfossintática. Por se tratar de *corpus* de pequenas proporções, palavras grafadas de forma idêntica, porém pertencentes a diferentes categorias gramaticais, por exemplo, puderam ser identificadas manualmente. É o caso de "voto", palavra-chave-chave com maior número de ocorrências. A visualização das linhas de concordância, possibilitada pela função Concord, do WS, evidenciou que a palavra é utilizada tanto como substantivo – como em "meu voto é" – quanto como conjugação do verbo "votar" na primeira pessoa do presente do indicativo – "[eu] voto sim/não", conforme ilustrado na Figura 1:

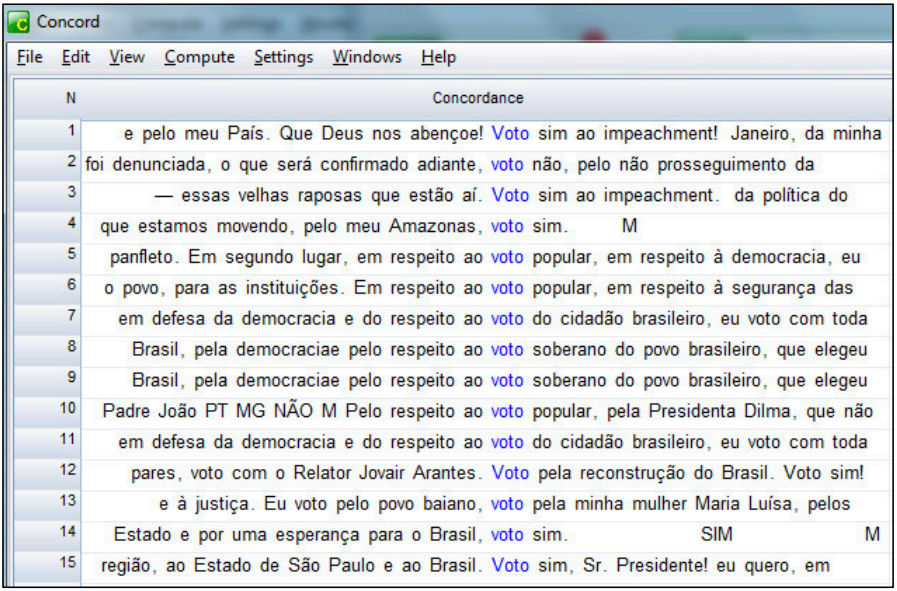


Figura 1 – Visualização parcial das linhas de concordância da palavra de busca "voto"
Fonte: Elaborado pela autora

Também recorrente nas três modalidades de voto é o vocativo "Sr. Presidente", comumente utilizado pelos parlamentares ao se dirigirem ao representante da Câmara dos Deputados[18]. Repetem-se, também, entre as palavras-chave "Brasil", "Bahia", "país", "partido" e "todos", resultando em frases como "em/a favor do Brasil", "pelo meu Brasil", "pela minha Bahia", "pelo meu País", "por todos os brasileiros" e "em respeito ao meu partido", utilizados para justificar as escolhas. Também não faltaram justificativas em nome de familiares, como em "meu pai, que me ensinou", "em nome da minha família". Votos "contra/em combate à/pelo fim da corrupção" também foram frequentes. A conjunção explicativa "porque" foi bastante utilizada para explicar o voto: "porque é constitucional/incompetente/necessário".

Devido ao tamanho reduzido do *subcorpus* "ABSTENÇÃO", o levantamento das palavras-chave-chave, ou seja, aquelas recorrentes nos três *subcorpora*, apresentou poucos resultados. Além disso, consideramos importante também analisar o entorno dessas palavras, ou seja, seus colocados. E um levantamento dos colocados das palavras estatisticamente recorrentes no *subcorpus* "ABSTENÇÃO" retornou um número insignificante de palavras lexicais. Além de palavras gramaticais – preposições, artigos e pronomes –, a única palavra de conteúdo que aparece

[18] Na época da votação do processo de *impeachment* de Dilma Rousseff, o presidente da Câmara era o então deputado Eduardo Cunha.

como colocado (da palavra-chave-chave "me") é "abstenho", como na sentença "eu me abstenho [de votar]". Portanto, decidimos dar continuidade na análise considerando somente os *subcorpora* "SIM" e "NÃO".

3.2 Palavras-chave

A fim de revelar as diferenças e semelhanças nos *subcorpora* "SIM" e "NÃO", partimos para o levantamento das palavras-chave exclusivas de cada um deles – apresentados nas duas primeiras colunas da Tabela 3, a seguir –, assim como das palavras-chave recorrentes nos dois *subcorpora* – apresentadas na terceira coluna. Entre parênteses, encontra-se o número de ocorrências de cada palavra. Na coluna "SIM" e "NÃO", os números em parênteses se referem, respectivamente, a cada modalidade de voto:

Tabela 3 – Palavras-chave nos *subcorpora* "SIM" e "NÃO"

SIM	NÃO	SIM e NÃO
abençoe (11), abraço (7), acima (10), agora (13), agricultores (7), Amazonas (9), amigos (14), amor (13), Arantes (7), Campos (9), Catarina (12), certeza (8), colegas (7), crescimento (10), dar (11), desempregados (10), dias (13), do (524), economia (9), eleitores (26), em (289), emprego (12), especial (15), especialmente (7), Espírito (9), esposa (18), fazer (20), fez (9), filha(s) (19), fim (20), fui (9), futuro (44), gerações (8), Goiás (12), gostaria (8), Grande (33), Grosso (10), honra (11), Janeiro (22), João (9), Jovair (7), jovens (8), mãe (13), Mato (10), me (59), melhor(es) (35), memória (10), mil (9), mim (20), mineiros (10), momento (40), Moro (8), mudança (12), mudar (10), muita (9), nacional (12), natal (7), neto(s) (23), Norte (9), nova(o)s (23),	aceito (3), agrária (7), aí (5), aquilo (4), ausência (3), bolsa (3), Brizola (3), cadeira (6), Câmara (3), campo (5), canalhas (4), cidadão (3), cinco (5), coerência (4), companheiros (8), conseguiu (4), contas (6), convicção (3), coragem (4), cumprir (7), da (144), daqueles (6), defender (6), deixar (3), democraticamente (3), democrático (5), deram (3), deveria (5), direito(s) (22), ditadura (7), elegeu (3), eleição(ões) (7), eleita (3), em (170), esse (27), estiveram (5), Exa. (6), fácil (5), falar (9), farsa (10), ficar (3), fizeram (3), fraude (3), golpistas (10), Governador (4), hipocrisia (8), história (13), homens (7), honesta (9), honrada (8), ilegítimo (4), injustiça (3), instituições (3), isso (29), jurei (7), lava-jato (5), legalidade (3), legitimamente (3), legitimidade (4), liberdade (6), luta (18), lutando (3), lutar (4), lutaram (7), mãos (6), mesa (3), Michel (13), movimentos (5), muitos (7), não (270), nas (24),	à (100) (53), ao(s) (207) (92), aqueles (13) (6), aqui (85) (36), Bahia (13) (9), Brasil (272) (46), brasileiro(a)(s) (162) (58), Casa (39) (26), cidade (44) (9), cometeu (13) (9), consciência (15) (4), Constituição (33) (38), contra (65) (89), corrupção (47) (17), crime(s) (27) (28), Cunha (8) (29), defesa (15) (49), democracia (14) (98), deputado(a)(s) (86) (41), desta(e) (30) (11), Deus (49) (7), dia (13) (8), dignidade (11) (6), digo (16) (5), Dilma (58) (45), direito (7) (12), dizer (31) (16), e (632) (251), é (238) (139), Eduardo (12) (23), em (289) (170), esperança (52) (4), esta(e) (80) (35), está (34) (29), Estado (137) (22), estamos (15) (5), estão (29) (20), estou (14) (4), ética (8) (3), eu (303) (129), família(s) (125) (11), favor (31) (5), federal(ais) (24) (12), filho(s) (60) (7), fora (26) (6), Gerais (31) (3), golpe (8) (87), Governo(s) (46) (4), hoje (48) (13), homenagem (32) (28), *impeachment* (92) (36),

o (548), Oeste (7), oportunidade (11), orgulho (11), pai(s) (29), Paraíba (9), Paraná (19), Paulo (37), pedir (9), possa (8), PRB (7), precisamos (11), princípios (7), PT (23), pública (12), região (19), representando (10), retomada (8), Rio (40), Rousseff (14), Santa (16), Santo (9), saúde (10), Sul (18), suplente (7), tão (12), temos (14), tenha (8), terra (10), toda(o) (38), viva (14), vontade (9)	nenhum (8), Neves (5), nunca (5), oposição (3), ouvi (7), pares (3), passaram (4), PCdoB (5), PDT (5), plenário (4), pobres (8), popular (12), porque (12), posição (8), posse (3), presidência (3), presidindo (4), primeiro (10), processo (24), quem (8), querem (13), quilombolas (4), razão (5), reconheço (3), reforma (10), respeitando (4), respeitar (3), resultado (6), réu(s) (8), sei (4), sem-terra (3), senhor(es) (8), sentado (3), séria (3), sertão (3), sessão (6), soberania (5), sociais (9), solução (6), Suíça (3), Supremo (6), Tancredo (3), tem (19), Temer (23), ter (10), tirar (7), tomei (3), trabalhador(a)(s) (es) (45), traidor (7), Tribunal (9), universidades (3), urnas (5), V.[Vossa] (6), venho (3), vi (4), vice-presidente (3), vocês (8), vota (3)	impedimento (9) (4), justiça (8) (4), juventude (8) (6), liberdade (10) (6), Lula (9) (7), maioria (17) (6), mandato (11) (6), meu(s) (461) (73), milhares (10) (5), milhões (28) (11), Minas (36) (4), minha(s) (275) (48), mulher(es) (9) (18), nação (31) (3), nesta(e) (62) (22), ninguém (14) (3), nome (105) (30), nós (61) (21), nossa(o)(s) (115) (29), país (115) (34), Pará (11) (4), parlamentares (13) (5), Partido (48) (9), pela(o)(s) (734) (151), Pernambuco (9) (8), política (24) (12), população (20) (11), por (242) (87), povo (192) (55), Presidenta (8) (22), Presidente (371) (119), que (529) (259), querida(o) (67) (8), quero (45) (26), República (20) (7), respeito (68) (39), responsabilidade (35) (13), rua(s) (27) (24), sim (393) (8), sou (19) (22), Sr(a)(s) (377) (102), tenho (19) (3), todos (72) (17), vai (21) (15), vamos (14) (5), vida (25) (14), votamos (7) (14), votando (9) (3), votar (22) (11), voto (384) (155), votos (9) (13), vou (16) (9)

Fonte: Elaborado pela autora

Fazer a análise de cada palavra-chave dos dois *subcorpora* fugiria ao escopo – e ao limite! – deste capítulo. Decidimos, portanto, enfocar as palavras que geraram maior polêmica na mídia e nas redes sociais, quais sejam, aquelas relacionadas a "Deus", "família" e "nação". Julgamos relevante, antes de tudo, verificar até que ponto os resultados levantados em nosso *corpus* de estudo coincidem com as conclusões de Chilton (2004). Para tanto, iniciamos a análise quantitativa a partir das palavras-chave presentes em "SIM" ou "NÃO".

3.2.1 "SIM" ou "NÃO"

Seguindo a dicotomia das estratégias de legitimação/deslegitimação identificada por Chilton (2004), observamos nos dois *subcorpora* diferentes formas de autorrepresentação positiva e representação negativa do opositor. "Amor",

"crescimento", "honra" e "orgulho" são algumas das escolhas lexicais dos votantes pró-*impeachment* para se referir a si mesmos, como podemos observar na fala transcrita a seguir (grifo nosso):

(1) "Sr. Presidente, quanta **honra** o destino me reservou de poder da minha voz sair o grito de esperança de milhões de brasileiros."

Já os eleitores contrários ao *impeachment* utilizaram, entre outras, palavras como "convicção", "coragem", "honesta" e "honrada" para se referir a si mesmos e à então presidente:

(2) "Defender a Constituição em momentos contra majoritários é para quem tem **coragem**."

Em relação à deslegitimação, é possível observar, do lado daqueles que apoiavam o processo, um vocabulário de crítica e acusação ao governo da época e de expectativa de progresso, caso o *impeachment* se concretizasse:

(3) "[...] pensando também nos 10 milhões de brasileiros que estão **desempregados** [...]."
(4) "[...] pelo meu querido povo mineiro e pela **retomada** do **crescimento** do Brasil [...]."

"Fraude", "farsa", "hipocrisia" e "golpistas" são exemplos de insultos dos governistas representando a situação, em relação a seus opositores:

(5) "E durmam com essa, **canalhas**!"

Identificamos, portanto, que muitas das estratégias identificadas por Chilton (2004), em sua análise de discursos e entrevistas envolvendo políticos ingleses, norte-americanos, entre outros, também foram reconhecidas de forma sistemática nas falas dos deputados brasileiros, tanto contrárias quanto favoráveis ao processo em votação.

Passemos, então, à análise de palavras estatisticamente recorrentes nas duas modalidades de votos.

3.2.2 "SIM" e "NÃO"

Entre as palavras-chave que se repetem na coluna "SIM e NÃO", encontram-se palavras funcionais, tais como preposições, artigos, conjunções etc. Em geral, essas palavras são preteridas pelas lexicais, ou de conteúdo, na análise do

corpus. Contudo, duas categorias merecem nossa atenção. São elas (i) o pronome em primeira pessoa do plural "nós" e seu respectivo pronome adjetivo possessivo ("nosso(a)(s)") e (ii) a combinação da preposição "per" e o artigo definido "o/a", resultando em "pelo(a)(s)", com significado de "em respeito a", "em nome de". O primeiro caso coincide com a observação de Chilton (2004), que afirma que, no discurso político, é comum a estratégia de aproximação com o interlocutor por meio desse recurso linguístico. Vejamos as declarações abaixo, que representam, respectivamente, votos "SIM" e "NÃO":

(6) "[...] em favor de melhorar a economia do **nosso** País [...]."
(7) "E golpe, **nós** não podemos votar por ele."

Legitimar o voto associando-o a outrem também pôde ser observado nas duas modalidades de voto:

(8) "**Pelo** Brasil, **pelo** meu Estado e **pela** honra da minha família, eu voto sim."
(9) "**Por** aquela trabalhadora que conseguiu ter uma carteira assinada; **por** aquele trabalhador que conseguiu colocar seu filho em Harvard ou no MIT; **pelo** trabalhador rural que recebeu energia elétrica na sua casa; **pelo** fim da hipocrisia, meu voto é não, Sr. Presidente."

Para Chilton (2004), o controle político está relacionado ao controle da informação. Esse controle pode ser qualitativo ou quantitativo. No segundo caso, o falante fornece dados estatísticos que (supostamente) comprovam a informação apresentada. Nas falas dos deputados que votaram a favor e contra o processo, observamos recorrência de palavras utilizadas para se referir a grandes quantidades. Vejamos os excertos (10), de um voto a favor do processo, e (11), contra:

(10) "Quero fazer homenagem aqui aos brasileiros de bem, àqueles **milhões** que foram às ruas para reivindicar mudanças [...]."
(11) "[...] em respeito aos **milhares** e **milhares** de brasileiros e brasileiras que votaram em Dilma [...]".

3.2.3 Tríade "Deus", "família" e "nação"

Voltemo-nos, agora, às palavras que deram vazão a diversas matérias jornalísticas e *posts* em redes sociais, em especial de crítica às falas pró-*impeachment*, a saber "Deus", "família" e "nação". Por meio do levantamento das palavras-chave significativamente recorrentes nos *subcorpora* "SIM" e "NÃO", é possível observar que (i) palavras relacionadas a familiares – "filhos" e "netos" –, (ii) palavras que se

referem ao território nacional – nação, país, Brasil –, e (iii) "Deus" recorrem de forma estatisticamente significativa nos dois *subcorpora*.

Contudo, é preciso ter cautela para não incorrermos nas mesmas generalizações feitas pela mídia e pelos internautas, que simplesmente calcularam o número de vezes que determinada palavra foi pronunciada e a associaram a uma das modalidades de voto, em geral àquela favorável ao *impeachment*. Afinal, frequência, apenas, não constitui análise confiável em um *corpus* como o deste estudo, em que não há balanceamento do número de palavras nos *subcorpora* que o compõem. Portanto, a fim de analisar as palavras que formam a "tríade" tão discutida, recorremos, quando possível, à análise dos colocados dessas palavras, a fim de se constatar ou não diferenças nessas modalidades de voto.

3.3 Colocação

Em conformidade com McEnery e Hardie (2012, p. 123), neste estudo colocação (*collocation*) é entendida como "padrões de coocorrência observados em dados do *corpus*"[19]. Várias são as ferramentas computacionais que medem o grau de proximidade entre dada palavra de busca e aquelas em seu entorno. Alguns exemplos são WS e AntConc (ANTHONY, 2016)[20]. Neste estudo, optamos pela ferramenta GraphColl (BREZINA et al., 2015)[21], que permite a análise de redes de colocados. Segundo Brezina et al. (2015, p. 141), "Colocados de palavras não ocorrem de forma isolada, mas fazem parte de uma rede complexa de relações semânticas que acaba revelando seu significado, assim como a estrutura semântica de um texto ou *corpus*"[22].

Assim como outros *softwares* de análise textual, o GraphColl oferece diferentes testes estatísticos para medir a atração entre palavras. Contudo, como em Baker (2016), optamos pelo MI (*mutual information*), uma vez que esse teste privilegia as relações entre palavras lexicais, preterindo as combinações com palavras gramaticais, que ocorrem com maior frequência. O MI considera a razão Observado/Esperado para medir a força de associação entre a palavra de busca e o colocado (HUNSTON, 2002), ou seja, calcula a frequência de coocorrência das palavras no *corpus* dentro de dada janela (*span*) (razão Observado), e a compara à frequência esperada de coocorrência, considerando o tamanho do *corpus* e a

[19] "[...] *co-occurrence patterns observed in* corpus *data.*"
[20] Para uma análise do levantamento de colocados por essas ferramentas, sugerimos a leitura de Baker (2016).
[21] Para detalhes sobre o potencial de uso da ferramenta, ver Brezina et al. (2015) e Baker (2016). Para informações detalhadas sobre os testes estatísticos disponíveis no GraphColl para medir a associação entre as palavras, recomendamos a leitura de Brezina et al. (2015).
[22] "*Collocates of words do not occur in isolation, but are part of a complex network of semantic relationships which ultimately reveals their meaning and the semantic structure of a text or* corpus."

frequência relativa de cada uma das palavras, caso ocorressem aleatoriamente no *corpus* (razão Esperado) (cf. CLEAR, 1993). Portanto, utilizando o teste estatístico MI (*Mutual Information*) (Stat: 03 – MI), ajustamos a ferramenta para identificar como colocados palavras que ocorressem ao menos cinco vezes em uma janela de cinco à direita e cinco à esquerda da palavra de busca (Span: 5 < > 5).

A título de ilustração, apresentamos, na Figura 2, o levantamento de colocados da palavra de busca "*Impeachment"* no *subcorpus* "SIM":

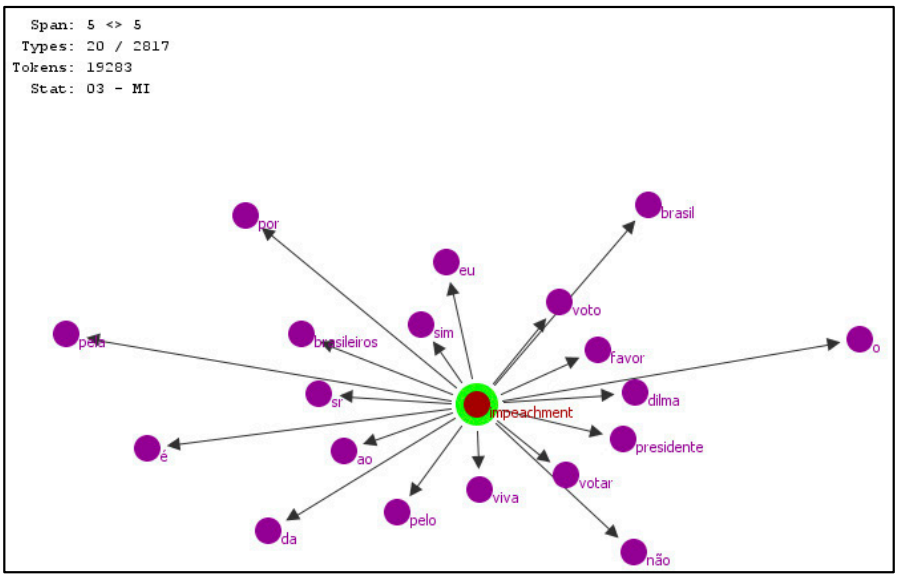


Figura 2 – Colocados de "*Impeachment"* no *subcorpus* "SIM"
Fonte: Elaborado pela autora

A figura apresenta a palavra de busca (*node word*) em destaque, ligada aos seus colocados por setas. O tamanho das linhas que separam a palavra de busca de seus colocados indica a proximidade entre eles: quanto mais curta a linha, mais forte a ligação entre eles.

A ferramenta também apresenta os colocados em formato de tabela, conforme apresentado a seguir:

Tabela 4 – Colocados de "*impeachment*" no *subcorpus* "SIM"

impeachment

Freq: 87, Edges: 0 in, 18 out

Dir	Type	Stat	Freq (within)	Freq (all)
out	Sim	6,654595	5	11
out	Voto	6,289598	6	17
out	ao	5,792098	39	156
out	votar	5,654595	5	22
out	favor	5,645257	7	31
out	Dilma	5,634557	13	58
out	sim	5,191949	63	382
out	Presidente	5,143360	59	370
out	Sr	5,108283	47	302
out	Por	4,866099	5	38
out	pelo	4,758378	37	303
out	eu	4,553312	25	236
out	voto	4,520390	38	367
out	brasileiros	4,470170	5	50
out	da	3,929151	20	291
out	é	3,653334	13	229
out	não	3,582645	6	111
out	o	3,198574	21	507

Fonte: Elaborado pela autora

Agora vejamos o resultado da busca com a mesma palavra, no *subcorpus* "NÃO" (Figura 3):

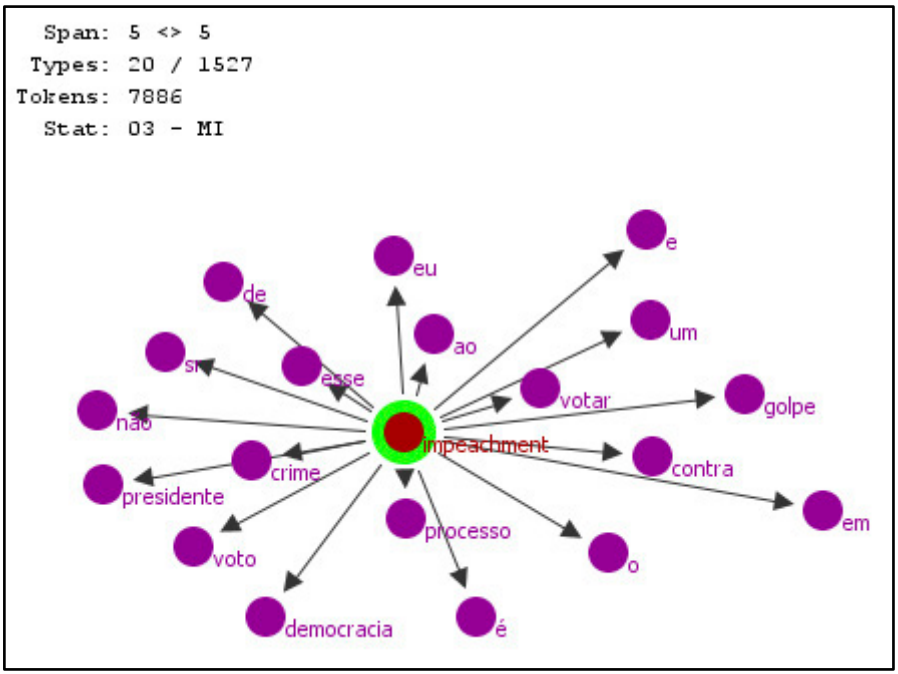


Figura 3 – Colocados de "*impeachment*" no *subcorpus* "NÃO"
Fonte: Elaborado pela autora

Não surpreende que no *subcorpus* dos votos favoráveis ao impedimento o colocado mais próximo de "*impeachment*" seja "sim", seguido de "presidente", "sr", "voto" etc. Expandindo a rede de ligações, observamos relações tais como "sr presidente voto sim ao *impeachment*". Semelhante levantamento no *subcorpus* "NÃO" mostrou que os principais colocados de "*impeachment*" são "não", "o", "de", "voto" etc., e que "não" se liga fortemente a "voto", "votar", "é"etc. No *subcorpus* "NÃO", coocorrem com "golpe" as palavras "contra", "não", "o", "voto" etc., formando relações tais como "contra o golpe eu voto não". Já no *subcorpus* "SIM", os colocados mais significativos com a mesma palavra de busca são "há" e "não", formando, entre outras, a relação "não há golpe".

Por meio dos colocados, é possível constatar, portanto, a posição de cada uma das modalidades de voto, representadas, aqui, pelos *subcorpora* "SIM" e "NÃO". Contudo, até aqui não há novidade em relação aos resultados obtidos com as outras ferramentas. Observamos, inclusive, colocados idênticos nos dois levantamentos. Nesses casos, vale a pena fazer uma busca em segundo nível, a fim de identificar as redes de colocações. Vejamos, a seguir, a Figura 4:

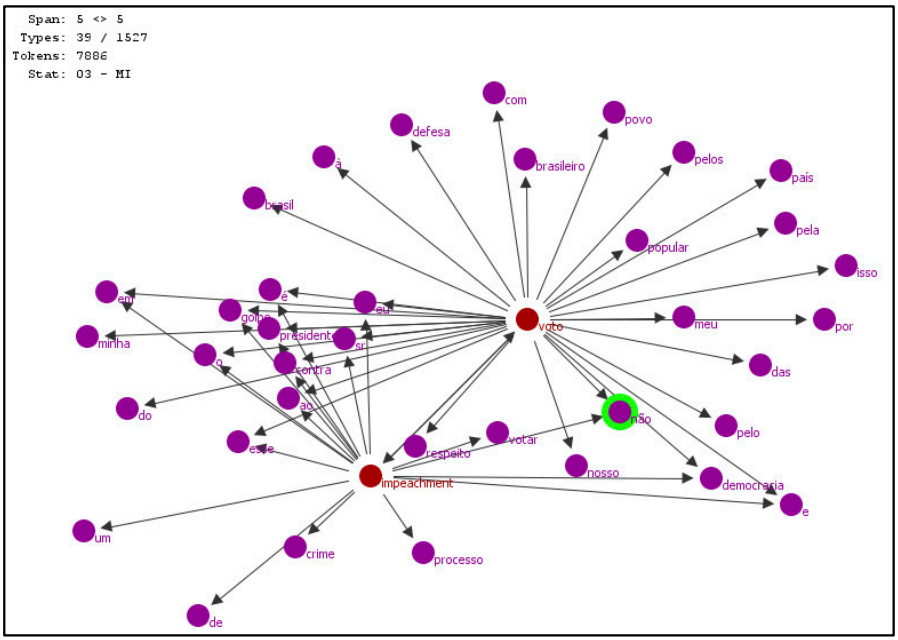


Figura 4 – Redes de colocados com a palavra de busca "*impeachment*" e seu colocado mais próximo, "voto", no *subcorpus* "NÃO"
Fonte: Elaborado pela autora

No *subcorpus* "NÃO", verificamos a relação entre "*impeachment*" → "voto" → "não". Busca semelhante no *subcorpus* "SIM" revelou a rede "*impeachment*" → "voto" → "sim". Portanto, podemos afirmar que o simples levantamento das palavras-chave recorrentes em cada *subcorpus* não esclarece a escolha lexical. Afinal, conforme evidenciado na Tabela 3, palavras relacionadas à chamada "tríade" – "família", "nação" e "Deus" – aparecem entre as palavras-chave tanto no *subcorpus* "SIM" quanto no "NÃO", o que nos levou a considerar importante o contexto de uso dessas palavras, antes de tecermos generalizações sobre elas. Em seguida, essas palavras serão analisadas mais detalhadamente.

3.3.1 "família"

Justificar a escolha em nome de familiares também constitui estratégia recorrente nas duas modalidades de voto, conforme pode ser observado nos excertos (12) e (13), extraídos, respectivamente, dos *subcorpora* "SIM" e "NÃO" (grifo nosso):

(12) "**Pela família** e **pela** inocência das crianças em sala de aula, que o PT nunca teve...".
(13) "**Em respeito à** minha **família** e à Constituição e **por** uma democracia plena no nosso País, eu voto não."

A análise dos colocados da palavra "família" mostrou que, entre aqueles lexicais, encontram-se, no *subcorpus* "SIM", "amigos", "esposa", brasileiro(a), "filhos", "pai", "respeito", "Deus", "Estado", "sim", "Brasil", "povo" e "voto". Já no *subcorpus* "NÃO", os colocados de "família" são apenas "minha" e "Bolsa", este compondo "Bolsa Família", programa social que visa a suprir as necessidades básicas de famílias de baixa renda. Também recorrente nos dois *subcorpora*, "filho(s)" forma colocação com "pelo" e "meu" no *subcorpus* "SIM". Já no *subcorpus* "NÃO", o número de ocorrências da palavra de busca não é suficiente para revelar colocados a partir dos ajustes estabelecidos.

3.3.2 "nação"

Em conformidade com Chilton (2004), observou-se, nos dois *subcorpora*, recorrência de palavras utilizadas como forma de se apelar ao patriotismo – "nação", "país" e "Brasil" – além de "pátria", que, com apenas oito ocorrências não aparece nas listas de palavras-chave. Os excertos (14) e (15) ilustram, respectivamente, os votos "SIM" e "NÃO" (grifo nosso):

(14) "É nessa direção, com respeito ao povo de São Paulo e por amor à **Nação** brasileira, que eu voto sim.
(15) "Em defesa da minha **Nação**, do Nordeste, do Piauí, da minha cidade de Oeiras, mas, principalmente, pelo combate à corrupção representada por Eduardo Cunha e Michel Temer, eu digo não a esta corrupção ridícula que envergonha o meu País".

Quanto aos colocados lexicais de "nação", observamos, no *subcorpus* "SIM", "mudar", "Deus", "Brasil" e "voto". Já no *subcorpus* "NÃO", não houve retorno de colocados com essa palavra de busca. Em relação à palavra "Brasil", com maior número de ocorrências entre as palavras que se referem a "nação", observamos "desistir" como principal colocado no *subcorpus* "SIM", formando uma rede de colocados com "não", "vamos" e "do". Já no *subcorpus* "NÃO", os principais colocados são "voto", "não", "democracia" e "contra", sendo que o principal é "quero", que forma a rede de colocados "dizer", "eu", "Presidente", "que" e "não".

3.3.3 "Deus"

A linguagem do discurso político costuma se entrelaçar com crenças religiosas (cf. CHILTON, 2004). Corroborando essa conclusão, "Deus" aparece como palavra-chave nos dois *subcorpora*. Vejamos dois excertos, (16) e (17), que representam, respectivamente, as modalidades "SIM" e "NÃO" (grifo nosso):

(16) "Que **Deus** abençoe o nosso país."
(17) "[...] eu rogo a **Deus** que ilumine os caminhos da Paraíba e os caminhos do Brasil."

A análise dos colocados, contudo, revela diferenças entre o uso da palavra "Deus" entre aqueles favoráveis e contrários ao processo. Observamos, no *subcorpus* "SIM", "abençoe", "senhor", "feliz", "nação", "país", "família", "sim", "voto", "Presidente", "povo" e "Brasil", só para mencionar os colocados lexicais de "Deus". Já em relação ao *subcorpus* "NÃO", o número baixo de ocorrências impossibilitou a busca por colocados, ao menos a partir dos ajustes feitos na ferramenta. Portanto, observemos as linhas de concordância (Quadro 2):

Quadro 2 – Linhas de concordância com a palavra "Deus" no *subcorpus* "NÃO"

N	Concordance		
1	e dando razão a V.Exa. quando pediu a	**Deus**	que tenha misericórdia deste País, e tem que ter
2	e Srs. Deputados, primeiro, eu rogo a	**Deus**	que ilumine os caminhos da Paraíba e os caminhos
3	quarto e ao sétimo mandamentos da lei de	**Deus**.	Quero dizer também, colegas Deputadas e
4	tão curto, eu ouvi tantas vezes o nome de	**Deus**	ser usado em vão, como se fosse um panfleto.
5	coisas mais diversas, inclusive o nome de	**Deus**.	Não aludem ao crime de responsabilidade, que
6	Meu	**Deus**!	Quanta hipocrisia! Não é Dilma que tem que sair do
7	entes, em primeiro lugar, eu oro para que	**Deus**	abençoe a nossa querida Nação, o Brasil. Em segundo

Fonte: Elaborado pela autora

Analisando-se os contextos de uso, verificamos que, das sete ocorrências da palavra "Deus" no *subcorpus* "NÃO", a palavra é utilizada (i) para criticar as falas daqueles que apoiaram o *impeachment* (linhas 1, 3, 4 e 5), (ii) como interjeição (linha 6) e (iii) para invocar ajuda divina (linhas 2 e 7).

4 Discussão

Conforme exposto na introdução deste capítulo, o que motivou a pesquisa foi o interesse em comparar as informações publicadas na mídia e nas redes sociais sobre as escolhas lexicais dos deputados durante votação do processo de

impeachment de Dilma Rousseff, em geral com críticas à recorrência de palavras relacionadas a Deus, nação e família na justificativa dos votos favoráveis ao prosseguimento do processo para o Senado.

Segundo os ajustes feitos na ferramenta para o levantamento das palavras que ocorrem com frequência significativamente mais alta no *corpus* de estudo do que no *corpus* de referência, observamos que palavras relacionadas à chamada "tríade" "família", "Deus" e "nação", apenas para mencionar aquelas que mais chamaram a atenção da mídia e dos internautas, foram recorrentes nas falas dos deputados que votaram a favor e contra o processo de *impeachment*, durante a votação realizada em 17 de abril de 2016, ainda que não ocorressem estatisticamente na mesma proporção: "Deus" ocorre em proporção de 0,25% no *subcorpus* "SIM" e 0,09%, no "NÃO". Além disso, a análise dos contextos de uso da palavra mostrou que não foi usada com o mesmo conceito nas duas modalidades de voto. Em todas as ocorrências da palavra no *subcorpus* "SIM", observamos o apelo para que Deus beneficie aqueles a quem o político alega defender, corroborando as conclusões de Chilton (2004) em sua análise de discursos políticos proferidos por líderes de diferentes nações. Já no *subcorpus* "NÃO", a palavra raramente foi utilizada com a mesma intenção.

Já as palavras relacionadas a território – "Brasil", "nação" e "país" – totalizam 2,17% das palavras do *subcorpus* "SIM" e 1,05% do "NÃO", ainda em conformidade com Chilton (2004), que identifica no discurso político o uso estratégico de referências patriotas. Justificativas associadas a familiares também recorrem nos dois *subcorpora*. Portanto, excluindo-se as palavras vinculadas aos polos contrários, quais sejam (i) o de autorrepresentação por meio de palavras positivas e (ii) o de desqualificar o opositor com palavras negativas, em geral observamos escolhas lexicais semelhantes nos dois *subcorpora*, evidenciando "[...] divergências entre o que o falante professa e aquilo em que realmente parece acreditar"[23] (PARTINGTON, 2004).

Para Van Dijk (2001 apud BAKER, 2006, p. 73-74), duas questões essenciais permeiam a ACD: a primeira diz respeito a como os grupos mais poderosos controlam o discurso público; a segunda, a como esse discurso influencia a forma de pensar e agir dos grupos menos poderosos. A mídia tem papel preponderante na produção e reprodução de discursos (BAKER, 2006). Assim, a representação do discurso na mídia deve ser entendida como um importante processo ideológico (FAIRCLOUGH, 1995) e, como tal, deve ser revelado. Ainda que se declare neutra, cada publicação tende a privilegiar algum lado. Em relação à política, em geral as publicações se posicionam a partir de ideologias relacionadas à direita ou à esquerda. Sendo a então Presidente da República, Dilma Rousseff, representante de partido de esquerda – PT (Partido dos Trabalhadores) –, não surpreende que

[23] "[…] *divergencies between what a speaker professes and what they really seem to think*."

jornais e revistas, impressos ou *on-line*, tenham se posicionado de forma contrária ou favorável ao impedimento. Vale ressaltar, contudo, que este estudo não pretende abarcar questões ideológicas das dicotomias políticas, nem mesmo das publicações utilizadas para justificar este estudo, mas simplesmente analisar o discurso da mídia e do público após o processo, no que tange especificamente às menções às escolhas lexicais utilizadas durante as falas dos deputados.

Ora, uma vez que o resultado da votação contentou aqueles favoráveis ao processo, é compreensível que as falas dos deputados que permitiram a continuidade do processo não tenham recebido grande atenção nas discussões pós-sessão. Já àqueles contrariados com o resultado restou a chance de criticar essas falas. No entanto, o posicionamento da mídia tem consequências. Os jornalistas exercem influência nos leitores ao produzir seus próprios discursos ou reformular outros já existentes (BAKER, 2006). No entanto, conforme enfatiza Fairclough (1989), os efeitos produzidos pela mídia são cumulativos: só atingem o objetivo desejado por meio da repetição. Além disso, os interlocutores não são passivos, mas interagem com o conteúdo a partir de sua posição ideológica.

5 Considerações finais

A análise semiautomática das falas dos deputados mostrou que, diferentemente do que a mídia publicou – e as redes sociais endossaram –, a tríade "nação", "Deus" e "família" não ocorre significativamente com maior frequência nos discursos pró-*impeachment*, mas sim em maior quantidade nesses discursos, uma vez que os votos favoráveis ao processo foram 2,68 vezes maiores do que aqueles contrários a ele. Além disso, é importante que se analise também o entorno das palavras, ou seja, os colocados, estes que podem evidenciar as reais diferenças entre os discursos.

A análise quantitativa, baseada em dados estatísticos, aliada à manual, possibilitada pela ACD, ajuda a revelar dados que poderiam ficar restritos apenas ao quantitativo ou à interpretação (tendenciosa) do analista. A metodologia da LC aplicada à ACD possibilita uma análise mais objetiva, a partir de dados revelados por meio de ferramentas computacionais. Contudo, não podemos afirmar que essa análise esteja totalmente livre de subjetividades. O analista sempre faz escolhas. Neste estudo, por exemplo, selecionamos as manchetes dos jornais que justificam nossa pesquisa, priorizamos a análise de algumas palavras-chave e colocados, em detrimento de outros, e escolhemos os excertos do *corpus* para ilustrar os resultados quantitativos obtidos.

A partir dos resultados evidenciados por meio da metodologia explicitada, não temos a pretensão de tecer generalizações sobre o discurso político, tão somente analisar as falas dos deputados durante o processo de *impeachment* de Dilma

Rousseff. Aventamos, para o futuro, a comparação com as falas dos deputados durante o processo de *impeachment* de Fernando Collor, a fim de verificar se houve diferença significativa em relação às falas analisadas neste estudo, uma vez que naquele caso o então presidente não representava ideologias de esquerda. Outra possibilidade, também, seria comparar as falas dos deputados às dos senadores, que deram continuidade ao processo. Enfim, várias são as possibilidades de análise do discurso político, que, naturalmente, não encerra com o término deste capítulo.

Referências

ANTHONY, L. *AntConc (3.4.4)* [Computer Software]. Tokyo: Waseda University, 2016. Disponível em: <http://www.laurenceanthony.net/software.html>. Acesso em: 23 out. 2017.

BAKER, M. *Corpus* Linguistics and Translation Studies: implications and applications. In: BAKER, M.; FRANCIS, G.; TOGNINI-BONELLI, E. (Org.). *Text and technology*: in Honour of John Sinclair. Amsterdam: John Benjamins, 1993, p. 233-250.

BAKER, P. *Using* corpora *in Discourse Analysis*. London: Continuum, 2006.

_____. Acceptable bias? Using *Corpus* Linguistics methods with Critical Discourse Analysis. *Critical Discourse Studies*, v. 9, n. 3, p. 247-256, ago. 2012.

_____. The shapes of collocation. *International Journal of* Corpus *Linguistics*, v. 21, n. 2, p. 139-164, 2016.

BAKER, P.; McENERY, T. A *corpus*-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics*, v. 4, n. 2, p. 197-226, 2005.

BAKER, P.; GABRIELATOS, C.; KHOSRAVINIK, M.; KRZYŻANOWSKI, M.; McENERY, T.; WODAK, R. A useful methodological synergy? Combining Critical Discourse Analysis and *Corpus* Linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, v. 19, n. 3, p. 273-306, 2008.

BIBER, D.; CONRAD, S.; REPPEN, R. Corpus *Linguistics*: investigating language structure and use. Cambridge: Cambridge University Press, 1998.

BREZINA, V.; McENERY, T.; WATTAM, S. Collocations in context: a new perspective on collocation networks. *International Journal of* Corpus *Linguistics*, v. 20, n. 2, p. 139-173, 2015.

CARTER, R.; McCARTHY, M.; O'KEEFFE, A. *From* corpus *to classroom*: language use and language teaching. Cambridge: Cambridge University Press, 2007.

CHENG, W. *Corpus*-based linguistic approaches to Critical Discourse Analysis. In: CHAPELLE, C. (Org.). *The Encyclopedia of Applied Linguistics*. London: Blackwell, 2013, p. 1-8.

CLEAR, J. From Firth principles: computational tools for the study of collocation. In: BAKER, M.; FRANCIS, G.; TOGNINI-BONELLI, E. (Org.). *Text and technology*: in honour of John Sinclair. Amsterdam: John Benjamins, 1993, p. 271-292.

CHILTON, P. *Analysing Political Discourse*: theory and practice. London: Routledge, 2004.

FAIRCLOUGH, N. *Language and power*. London: Longman, 1989.

_____. *Critical Discourse Analysis*: the critical study of language. London: Longman, 1995.

FETZER, A. Put bluntly, you have something of a credibility problem: sincerity and credibility in political interviews. In: CHILTON, P.; SCHÄFFNER, C. (Org.). *Politics as Text and Talk*. Amsterdam: John Benjamins, 2002, p. 173-201.

HUNSTON, S. Corpora *in Applied Linguistics*. Cambridge: Cambridge University Press, 2002.

McENERY, T.; HARDIE, A. Corpus *linguistics*: method, theory and practice. Cambridge: Cambridge University Press, 2012.

PARTINGTON, A. *Corpora* and discourse, a most congruous beast. In: PARTINGTON, A.; MORLEY, J.; HAARMAN, L. (Org.). Corpora *and discourse*. Bern: Peter Lang, 2004, p. 9-18.

PARTINGTON, A.; MORLEY, J.; HAARMAN, L. (Org.). Corpora *and discourse*. Bern: Peter Lang, 2004.

PEARSON, J. *Terms in context*. Amsterdam: John Benjamins, 1998.

PRADO, M. *Levantamento dos padrões léxico-gramaticais do inglês para aviação*: um estudo vetorado pela Linguística de *Corpus*. 2015. 133 f. Dissertação (mestrado em Estudos Linguísticos e Literários em Inglês). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2015.

SANDERSON, T. Corpus*, culture, discourse*. Tübingen: Gunter Narr Verlag, 2008.

SCOTT, M. *Word Smith Tools version 6.0*. Oxford: Oxford University Press, 2012.

VAN DIJK, T. Critical Discourse Analysis. In: SCHIFFRIN, D. T.; HAMILTON, H. E. (Org.). *The Handbook of Discourse Analysis.* London: Blackwell, 2001, p. 352-71.

ZANETTIN, F. *Translation-driven* corpora: *corpus* resources for Descriptive and Applied Translation Studies. Manchester: St. Jerome, 2012.

Hierarchical clustering of aspects for opinion mining: a *corpus* study

Clusterização hierárquica de aspectos para mineração de opinião: um estudo de córpus

Francielle Alves Vargas
Thiago Alexandre Salgueiro Pardo

Abstract: This paper consists of an empirical study on the problem of clustering and hierarchically organizing opinion aspects in product reviews in order to support aspect-based opinion mining applications. We performed a *corpus* study for characterizing and understanding the involved tasks, looking for linguistic patterns and convergences and divergences across domains. The process has been manually performed and resulted in reference data for future developments and evaluation of automatic methods in the area.

Keywords: Natural Language Processing. Opinion Mining. *Corpus* Linguistics.

Francielle Alves Vargas – Msc in Computer Science and Mathematics Computational (ICMC), University of São Paulo (USP) – francielleavargas@usp.br.
Thiago Alexandre Salgueiro Pardo – Professor and researcher at Institute of Mathematical and Computer Sciences (ICMC) in University of São Paulo (USP), Phd in Computer Science and Mathematics Computational, University of São Paulo (USP) – taspardo@icmc.usp.br.

Resumo: Este artigo consiste em um estudo empírico sobre o problema de agrupamento e organização hierárquica de aspectos a partir de revisões de usuários sobre produtos na *web*, a fim de apoiar aplicações de mineração de opinião baseada em aspectos. Realizamos um estudo de córpus para caracterização e compreensão das tarefas envolvidas, buscando padrões linguísticos, além de convergências e divergências entre os domínios. O processo foi realizado manualmente e resultou em dados de referência para pesquisas futuras e avaliação de métodos automáticos na área.

Palavras-chave: Processamento de Linguagem Natural. Mineração de Opinião Baseada em Aspectos. Linguística de *Corpus*.

1 Introduction

The expansion of the social networks and e-commerce services resulted in the growth of on-line reviews in the web. Websites as Amazon and Buscape encourage users to write reviews for products, where users may do objective or subjective descriptions for a product and its aspects or properties. Subjective descriptions are characterized by a personal language, with opinions, sentiments, emotions and judgments. The research area in charge of identifying, extracting and summarizing subjective information in texts is called opinion mining or sentiment analysis (PANG et al., 2002). According to Zhao e Li (2009), this area is different from the traditional text mining area, which is mostly based on objective topics rather than on subjective perceptions. Many searches used reviews of movie, book, and electronic product domains (HU et al., 2004), because these domains have relevance to both companies and consumers. For companies, it is important to evaluate their reputation, acceptability and evaluation of their products. For consumers, summarizing reviews from other users makes it easier to make decisions at the time of purchase. Therefore, providing relevant subjective content in reviews, among these various domains, is an important task in the current context, because it provides a better utilization of those data, for this purpose consumers and for private and governmental organizations.

2 Background

Opinion mining or sentiment analysis is the field of research in intersection between linguistics and computation, responsible for proposing methods of analysis, processing, summarization and classification for large volumes of data, mostly text type, for the extraction of subjective content. According to Liu (2012), there are three main granularity levels of analysis for opinion mining. These are: (i) document level, (ii) sentence level and (iii) aspects level. At the document level, the set of the opinions expressed in the document is accounted. For example, a document composed by subjective content may be classified as positive, negative

or neutral, according to the accounting of relevant content that express sentiment. According to Liu (2012), these contents are usually expressed through adjectives. In Figure 1, we present a set of reviews on a smartphone.

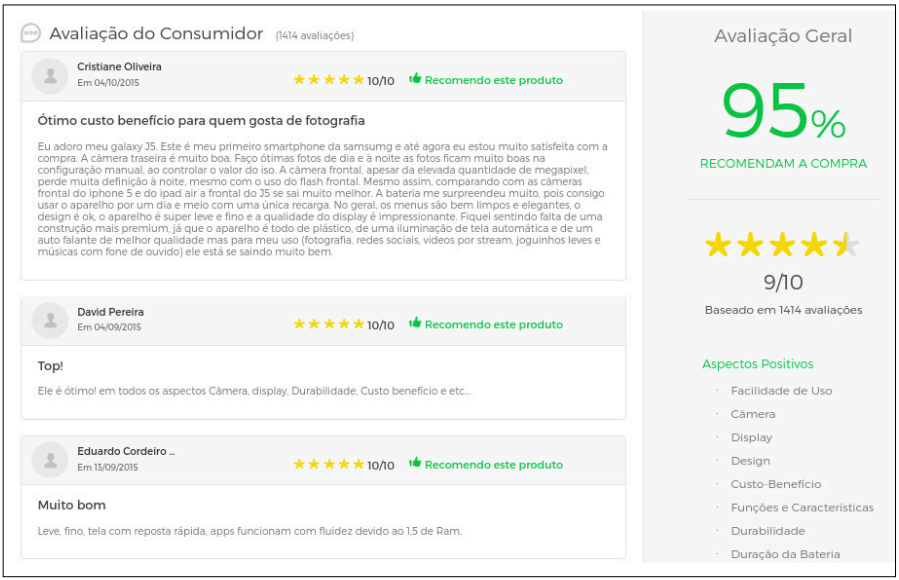


Figure 1 – Set of reviews on a smartphone and extracted from Buscapé

Note that 1.414 reviews have been issued on the smartphone product. Each of the 1.414 reviews refers to a document for opinion mining systems. Therefore, at document level, a positive, negative or neutral score is issued for each document. In this level of granularity, it is not possible to know precisely what the user liked or did not like. At the sentence level, the objective is to determine the opinion expressed in each sentence of the document. Therefore, a set of documents is segmented into sentences, and then a score is issued for each of the sentences. For example, in a document composed by $\boldsymbol{X}$ sentences, for each of those sentences there will be a positive, negative or neutral score. See again the Figure 1. In the third review, the user issues the following evaluation "Leve, fino, tela com resposta rápida, apps funcionam com fluidez devido ao 15 de Ram". Note that at this level, it is still not possible to know accurately the properties of the product evaluated by the user. To solve this problem, Liu (2012) argues that it is necessary a deeper level of analysis: aspect-based opinion mining. Aspects represent properties or parts of entities that are evaluated by users, in reviews, such as comments on websites and blogs on the web (LIU, 2012). However, the distinction between "attributes" and "aspects" is not clear in the literature. Mostly, they are used with synonyms.

For example, in the review, "A qualidade da imagem da câmera é excelente", the term "qualidade da imagem" is an attribute of the "imagem" aspect. Within this paper scope, we limit our contributions to the aspect level.

For Bhuiyan et al. (2009) opinion mining surveys may be divided into two main tasks: the *sentiment classification* and the aspect-based opinion mining. The sentiment classification consists in to recognize the general sentiment present in a document or a sentence. Typically, this task is simplified, classifying a document or a sentence into 3 classes: positive, negative or neutral (AVANCO; NUNES, 2014). Aspect-based opinion mining is usually focused on the three tasks: (i) aspects identification, (ii) polarity identification, and (iii) summarization (LIU, 2012). In Figure 2, we illustrate these tasks, and then we describe .

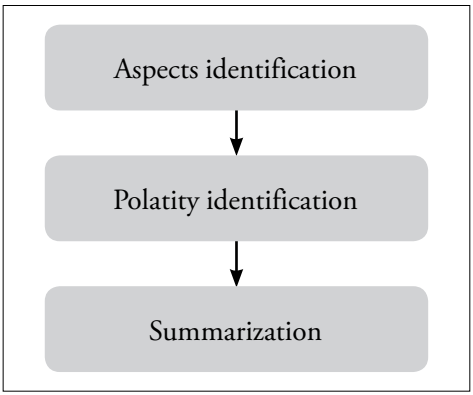


Figure 2 – Aspect-based opinion mining tasks

Aspect identification: features evaluated by users are extracted on the target of the opinion. For example, in the review "The Iphone 6 screen is amazing", the aspect evaluated is "screen"'.

Polarity identification: the sentiment associated with the aspects are extracted. For example, in the review "The camera battery is bad", the sentiment expressed about "bateria" aspect is negative, so the polarity of this aspect, considering this review is negative.

Summarization: the most relevant content is displayed through summaries, usually *extractive summaries*, which display the summarized content through the sentences clustering; or the *abstractive summaries*, which not only select the most relevant sentences of the source texts, but analyze the document and automatically generate new sentences. This approach attempts to produce new texts from the original fragments identified as relevant.

In addition to the tasks of aspects identification, polarity identification and summarization, according to Taboada (2016), another task of opinion mining

responsible for determining whether a text, or most of them, is subjective or objective. According to the author, textual content may contain objective information (facts, actions) or subjective information (perceptions, opinions, sentiments). Moreover, subjective texts express a positive or negative view and this direction of opinion - whether positive or negative - is also known as semantic orientation.

Aspect-based opinion mining, according to Liu (2012), represents a "delicious challenge". Natural languages are very rich and allow to express subjectivity in different ways. Not every opinion is directly expressed and not every aspect appears in a explicit way in the text. For example, in "The camera is expensive", the evaluated aspect is "price", but it is implicit, not being explicitly said in the sentence and, therefore, must be inferred from the context. Therefore, the aspects may be found explicitly and/or implicitly. Explicit aspects are explicit evaluations of one or more properties of the object / target of opinion. For example, the review showed in Figure 3 (kept in Portuguese, the original language), *"Amazing price-benefit relation, it has good digital camera, inclusive for video. Good memory space. What I liked: I received calls up to the riverside. What I did not liked: the sound sometime is low."* The review aspects are: "*price-benefit*", "camera", "video", "memory space", "sound" and "signal", however, the "signal" aspect is implicit. The users used the expression "r*I received calls up to the riverside*" to evaluate the "signal" aspect of a smartphone.

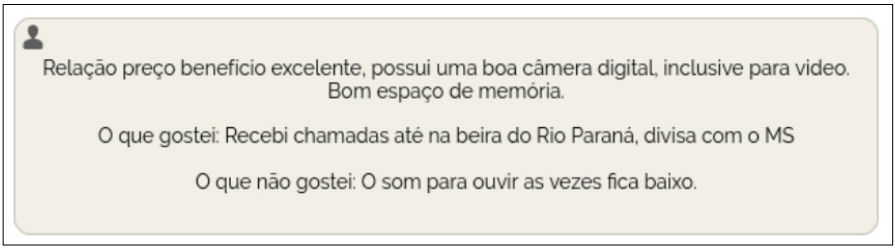


Figure 3 – Review about a smartphone

Another challenge consists in identifying different aspects that refers to the same object attribute/property. Users recurrently make reference to services or products attributes/properties using different terms. For example, consumers may use the terms "value", "cost", "price" and "investment" to designate the price of a smartphone, or to employ the terms "screen" "glass" and "display" to qualify a same smartphone property. Furthermore, users may employ implicit aspects cues. For example, the expressions "*I received calls up to the riverside*" and "It working anywhere", all these were employed to evaluate the signal property of a smartphone. Another example is the term "compatibility", which was employed to evaluate the operating system of a smartphone. Concomitant, the terms

"program", "system" and "application" were also employed. In addition, there is a significant portion of proper nouns applied to refers to the same property of the object. For example, "edward", "edward cullen", "noelle page", "larry" and "bella" are employed to evaluate the "protagonist"' and the terms "josé saramago" and "thalita rebouças" were employed to evaluate the "author" in the book domain. In the camera domain, the terms "sony", "nikon", "fuji" and "benq" are applied to evaluate the "brand" of a camera. Therefore, in opinion mining systems, the aspects clustering is very important, because this task provides the results the results over the veritable properties evaluated by users.

Explicit and implicit aspects clustering task has great relevance for opinion mining systems. However, this task is not trivial. For example, the speakers of a natural language may employed distinct lexical items to refer to the same object in the world. The words "price", "cost", "price-benefict", "value", "cheap", "expensive", may be employed to refer to the same smartphone propriety, for example. To illustrate how this phenomenon affects reviews, see the diagram shown in Figure 4. In this figure, we presented a portion of the groups of aspects identified in the smartphone domain.

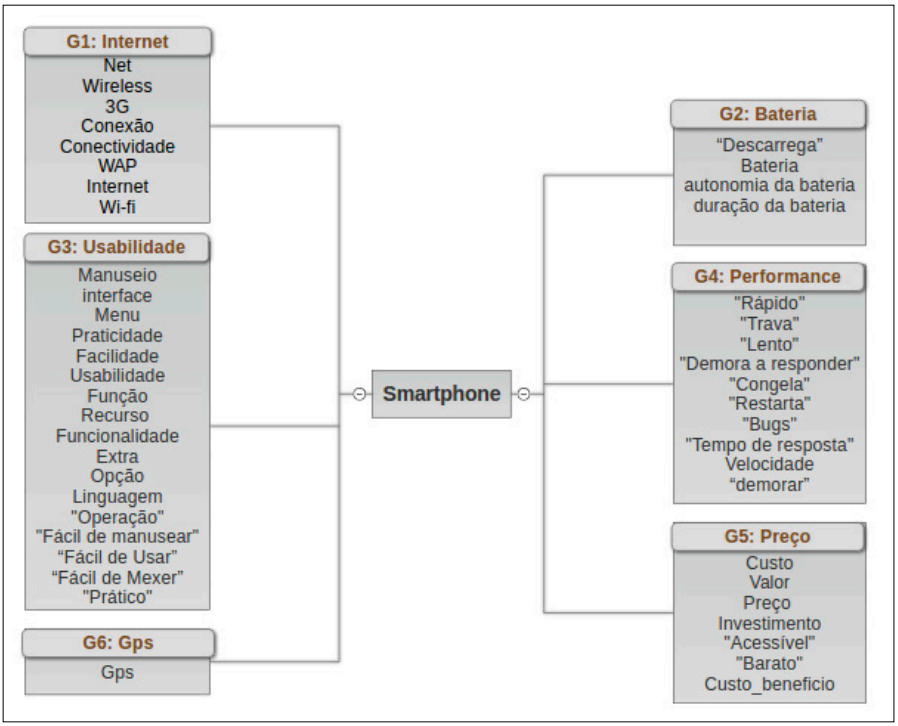


Figure 4 – Some smartphone domain aspects groups

In G1 group, the object property is "internet". Note that users may use the terms "3g", "wifi"' and "wireless", which are internet connection types, to evaluate this property of the device. Note also that these terms, because they represent the domain specificities, are not always found in linguistic-computational resources of the language as wordnets[1]. As well as these terms, other terms may be found, for example, "net", "internet", "connection" and "connectivity" to evaluate the same ownership of the mobile device. In G2 group, we find a recurring phenomenon in the *corpus* of user reviews: *aspect attributes*. According to Liu, (2012), aspects have attributes that have aspect properties. For example, the expressions "battery life" is a property of the "battery" aspect. In this case, there is an intrinsic relation between the lexical units. It is also a relation of substring[2]. The G3 group consists of aspects used to evaluate the "usability" property of a smartphone. See how the aspects clustering task is not simple, since many of them are terms that denote vageza (for example, "option"', "function", "resource", "extra"'). According to Zipf (1949), most words have multiple definitions, however, more frequent words tend to be more ambiguous. Still on the items of G3 group, we may see expressions indicative of implicit aspects. For example, the terms "It easy use" and "It easy handle", in addition to the terms "operation" and "practice" are used to designate the aspect of the smartphone. Notice the difficulty for items clustering from distinct nature (verbs, nouns, adjectives) in the same group. In G4 group, users evaluated the performance property of the smartphone. The term "bugs", derived from foreignism (in portuguese language) and the terms "response time"' and "take time to respond" are indicative of aspects implicit and they are used to evaluate the "performance" propriety. In G5 group, it is interesting to observe 2 behaviors in particular. The first behavior consists of the terms "accessible" and "cheap", which are terms applied to indicate implicit aspects. See that the terms "accessible" and "cheap" are terms highly ambiguous, and an inference mechanism in the domain is required for correct interpretive correspondence of these items. The second behavior is represented by the term "investimento". We observe a semantic neologism to which the added value "cost" or "price" is inserted. Finally, the G6 group represents the unit groups in the user review *corpus*. The unit groups represent unique units without semantic correspondence and may be localized in the content plan in reviews. For example, we did not find in the *corpus* another similar aspects with the "gps"' aspect of the smartphone, so this aspect constitutes a unit group. Therefore, the aspects clustering task may be defined by the

[1] Wordnets are large lexical database of a language in which nouns, verbs, adjectives and adverbs, for example, are grouped into synonym synsets, each expressing a distinct concept (MILLER et al., 1990).

[2] Substring is a string that appears within words in the text. For example, the string "ando" is a substring of "walking".

recognition of correlated aspects semantically, in other words, all of which have interpretive correspondence in a given domain.

Another significant challenge for opinion mining, according to Yu et al. (2011), is that the product reviews are numerous and disorganized. For example, at the *Buscape.com*, the *Smartphone Samsung Galaxy J5 SM-J500M* product has 931 reviews and, for each review, several aspects are evaluated. Thus, consumers will hardly learn all the other consumers' opinions about the product. According to the author, the hierarchical organization of aspects in product reviews would allow a better structuring of this data, so that it becomes intelligible for both machines and humans. A example of hierarchical organization of aspects in product reviews is shown in Figure 5. This work was proposed by Yu et al. (2011). The authors proposed a method based in linguistic and statistical knowledge in order to provide hierarchical organization of aspects from product reviews. There were considered 9.245 reviews on an iPhone 3G. Note that the hierarchical organization of the evaluated aspects about the iPhone 3G product is clear, unambiguous and may be easily understood for other consumers.

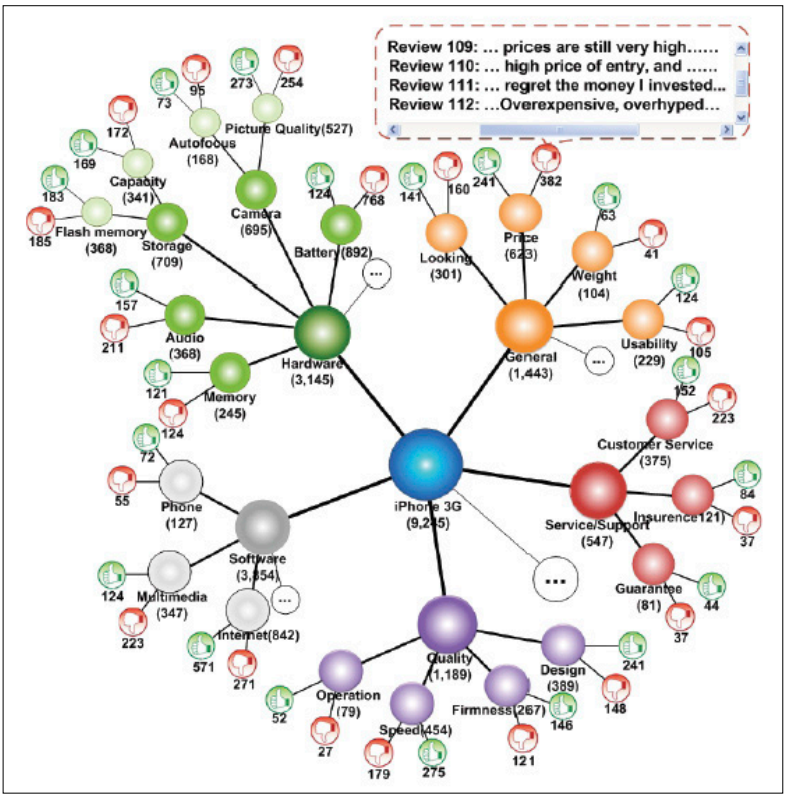


Figure 5 – Hierarchical organization of aspects (YU et al., 2011)

In this scenario, due to all the challenges, we present a *corpus* study of opinion aspects, regarding both their clustering and hierarchical organization. We analyze reviews for books and electronic products, looking for linguistic patterns and convergences and divergences across these domains. We expect that such investigation may help characterizing the involved tasks and provide valuable reference data for future developments and evaluations in the area. The rest of this paper is organized as follows. Section 3 describes the dataset and the analysis method. Section 4 presents the achieved results and learned lessons. Some final remarks are made in Section 5.

3 *Corpus* study

3.1 The dataset

The dataset overview is shown in Table 1. According to Zhao e Li (2009), most of the existing opinion mining initiatives are based on product reviews because reviews usually focus on specific products and contain little irrelevant information. Therefore, we randomly selected 60 smartphone and 60 camera reviews from the Buscape *corpus* (HARTMANN et al., 2007) and 60 book reviews from the ReLi *corpus* (FREITAS et al., 2012). We manually analyzed the data behavior in each domain. The Buscape *corpus* is composed of product reviews in Portuguese language for cameras, notebooks, telephones, TVs etc. In this *corpus*, the reviews are partially structured, with sections for "overall impression", "what I liked" and "what I did not like". For example, see the following camera review: *"Amazing, even today everyone is impressed by its size and beauty beyond perfect pictures that can be taken up 6.3 megapixels! What I liked: slim, practice and light. What I did liked: None!*" . One may note that several aspects were evaluated in this review, but some are not explicit. For example, the terms "beauty", "slim", "practice" and "light" are *clues* that indicate the implicit aspects "design", "size", "usability" and "weight", respectively. The ReLi *corpus* consists of book reviews, that are also in Portuguese language. As an example, one may find the following review: *"Amazing book, very different of what I imagined. Despite being old, it is good reading with the very modern language."*. We chose to select only 60 reviews for each domain, mainly because ours is a study carried out manually, from the qualitative and quantitative approach. Our main hypothesis is that for each domain there are different linguistic behaviors and phenomena.

Table 1

Domain	Reviews	Tokens	Types
Book	60	35.771	1.577
Smartphone	60	6.077	1.496
Camera	60	3.887	1.060

In the book domain, according to the Table 1, there was a significant spike in the number of tokens when compared to smartphone and camera domains. In this domain, we characterize an expressive number of irrelevant content. There were identified 52,01% relevant content and 47,98% irrelevant content. Smartphone and camera domains, the irrelevant content was not statistically significant.

3.2 The analysis method

In this work, the main purpose is to investigate the clustering and hierarchical organization of opinion aspects. Our main motivation with this *corpus* study was to understand the characteristics and challenges in the process of recognizing groups of aspects and the semantic organization of these groups. Our goal is to propose linguistically motivated solutions for opinion mining systems. We have selected 3 distinct domains: smartphone, camera and book in order to understand the convergence and divergence of behavior between domains. The empirical analysis was performed manually and will serve as a reference (human) for the evaluation of the proposed automatic methods, as well as a resource for the future research. In our study, we presented several quantitative and qualitative data on the aspects clustering task, besides some empirical evidence that the linguistic behavior varies between domains and that these variations have strong linkages with the knowledge specificities of a domain and with the profiles of the writer/user which produces the content. Figure 6 illustrates the clustering process, which was manually performed.

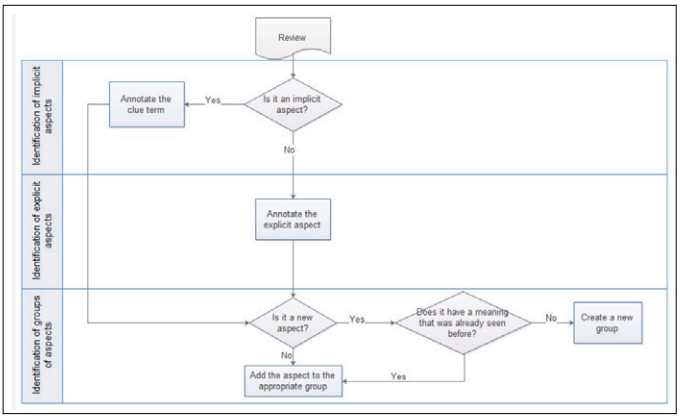


Figure 6 – Clustering product aspects

For the 180 reviews, a human labeled the implicit and explicit aspects. In the identification of implicit aspects were labeled the clue terms that indicated the aspects. For example, in "*This camera is expensive*", the evaluated aspect is "price", but it is implicit. The term "expensive" is the clue term. The identification of explicit aspects were directly labeled the aspects. For example, in "*The history of the book is bad*", "history" is an explicit aspect. In the last stage, the aspects were clustered the that had similar meaning but with different wording, in order to identify groups. For example, the "cost", "value", "price" and "investment" aspects form an unique group.

We also modeled the progression of this process of clustering product aspects, looking for a "learning curve" (shown in Appendix 1). Our objective was to measure the behavior of the aspects clustering task in order to identify the *stabilization point* for identifying new groups of aspects in a domain. The curves are shown in Figures 7, 8 and 9. The ***X*** axis of the learning curves represents the number of reviews analyzed and the ***Y*** axis the number of new groups identified. For example, The ***X*** axis, the number 1, shown in Figure 7, there were recognition 8 groups of aspects, as shown by ***Y*** axis. After the analysis of the first 10 reviews, 33 groups of aspects were recognized, and so on. For smartphone, digital camera and book domains, an average of 40 reviews are required for learning groups of representative aspects of the domain.

Once clustering was ready, the obtained groups were manually organized in hierarchies (one for each domain, shown in Appendix 2). We compared our obtained hierarchies with other available hierarchies in the area. We also identified the groups of aspects with the highest number of evaluations in the smartphone, camera and book domains, looking for a "prototypical groups" (shown in Appendix 3). For example, for the smartphone domain, some groups of prototypes are: "smartphone", "usability", "design", "value", "battery", "brand" etc.

4 Results

As explained before, we manually analyzed the product reviews and could observe some very interesting things. The results demonstrated that product reviews may contain portions of irrelevant information, i.e., information that is not directly related to the opinions about the products. The book domain showed 47.98% of irrelevant content, when users comment about the books but do not express any opinion or sentiment. However, for smartphone and camera domains, there was no significant value of irrelevant content.

We could notice that the user profile influences the review informational status[3]. We observed that the smartphone and camera domains present more aspects and groups of aspects than the book domain, as shown in Table 2. Note that the total aspects number in each domain and the average number of aspects in reviews seems to us to be relevant empirical evidence for the relationship between user profile and informational status. Smartphones and cameras are popular technological products and their aspects are more easily identified by non-expert users. Therefore, "expert users" have a greater level of information, in other words, these users have more knowledge about the domain, which allows them to evaluate a larger number of aspects of the evaluated entity. In the book domain, the users often are "just readers" and non-expert users in literature or literary critic. Therefore, they usually do not care about the book technical aspects (such as "size" or "paper type"). These users have been able to evaluate a limited number of product aspects, generally prototypical aspects of the books. It is also interesting that the vocabulary in book reviews are not uniform, as the users do not share the same extralinguistic variables, such as age, gender, education and social level. More "adult" books have more sophisticated reviews, with better language, while the reviews of "teenager" books are more often marked by the orality and informal language. These results demonstrate how complex the tasks of opinion mining are, especially aspect-based opinion mining. Opinion mining systems that does not consider the linguistic behavior and/or domain specificities as a processing criterion, for example, incurs the risk of classifying aspects that were not evaluated by the user, so they will return a result that is not in accordance with the reality presented in the review. In addition, we noticed in reviews content that had opinion/sentiment explicitly, they was accompanied mainly by psychological verbs, as occurs, for example, in "I found the history a little stopped", "I loved the book" and "Although I did not like the book", without necessarily having adjectives.

[3] According to Koch (2009), the informativeness of a text is associated to its ability to present new and unexpected information.

Table 2

	Smartphone	Camera	Book	Average
Total number of aspects	459	342	323	374,66
Unique aspects	180	132	103	138,33
Explicit aspects	392	289	298	326,33
Implicit aspects	67	53	45	55,00

Overall, 87.08% of the aspects are explicit and 12.91% are implicit in the domains. Furthermore, a product review is composed of, on average, 6 aspects, and it may have at least 1 implicit aspect (see Table 3). We also identified product reviews with the maximum of 20 aspects and the maximum of 5 implicit aspects.

Table 3

	Smartphone	Camera	Book	Average
Average number of aspects	7,65	5,70	5,38	6,24
Average number of explicit aspects	6,53	4,81	4,96	5,43
Average number of implicit aspects	1,11	0,91	0,85	0,95
Maximum number of aspects	20	20	15	18,33

We also perform a mapping of the grammatical classes of the terms indicative of implicit aspects. We divide the indicative terms of aspects into 2 classes, *nominal* and *verbal,* in order to measure the proportion of each one of these classes in the analyzed domains. In the "nominal" class , we framed "non-verbal" lexical items, in other words, it is nouns, adjectives, adverbs etc. In the verbal class, verbal lexical items were framed, in other words, it is verbs. In the smartphone domain, 73,68% are nominal implicit aspects and 26,31% are verbal implicit aspects. In the camera domain, 69,56% are nominal implicit aspects and 30,43% are verbal implicit aspects. Lastly, in the book domain, 50% are nominal implicit aspects and 50% are verbal implicit aspects.

Regarding the clustering step, we identified, on average, 3,08 explicit aspects and 0,77 implicit aspects per group. Some groups presented the maximum of 19 aspects, as shown in Table 4. In these groups (those that are not unitary, i.e., that contain more than one aspect), the predominant relation between 2 aspects is of the *is-a* / hypernym (or hyponym, dependending of the direction of the relation) type (e.g., between the aspects "equipment" and "product"), followed by synonym ("price" and "cost") or identity (when there is a single aspect without a

direct corresponding synonym in the group), part-of / metonym (or holonym) ("key" and "keyboard"), deverbal construction ("reflect" and "reflection") and coreference ("manufacturer" and "brand"). The remaining cases are formed by unitary groups, with only one aspect (without relations, therefore). Table 5 shows the distribution of these relations.

Table 4

	Smartphone	Camera	Book	Average
number of groups of aspects	**48**	**37**	**21**	**35,33**
avg number of aspects in a group	3,75	3,56	4,29	3,86
avg number of explicit aspects in a group	2,85	2,78	3,62	3,08
avg number of implicit aspects in a group	0,89	0,88	0,86	0,87
maximum number of aspects in a group	15	19	17	17

Table 5

	Smartphone	Camera	Book	Average
is-a / hypernym	**45,00%**	**37,12%**	**46,60%**	**42,90%**
synonym / identity	23,88%	18,93%	26,21%	23,00%
part-of / metonym	8,91%	15,18%	7,76%	10,61%
deverbal construction	5,55%	6,81%	9,70%	7,35%
coreference	6,66%	8,33%	0,00%	4,99%
no relation (unitary groups)	10,00%	13,63%	9,73%	11,12%

We found several challenges in the analysis: *(i) the inherent ambiguity of the natural languages*, occurring, for example, for the terms "function", "resource" and "application", that are used to refer to the same smartphone application; *(ii) the specificities of the domain*, as each domain requires specific background knowledge; *(iii) the implicit aspects,* as the implicit aspect identification task is not always intuitive; *(iv) the aspects outside the domain,* as the terms "delivery", "technical assistance" and "SAC", which, although have been evaluated, are not directly related to the products. Our study also showed that it is necessary the analysis of 40 reviews, on average, to learn/identify most of the relevant aspects in a given domain. The "learning curves" (shown in Appendix 1), represent the learning behavior of groups of aspects for the analyzed domains, that is the amount of new groups of aspects learned at each review. We also hierarchically organized the identified groups of aspects (see in Appendix 2) and compared our hierarchies

with the hierarchies proposed by Condori (2015), Acir et al. (2006) and Goulart and Montardo (2007). In the hierarchies in the literature, the relations of the type *is-a* are more often used. However, we observed that reviews are predominantly composed by *part-of* relations. Furthermore, the hierarchies in the literature do not represent all the domain specificities.

5 Final remarks

As shown above, clustering product aspects and building their hierarchical organizations are not simple tasks. There are several challenges to overcome. The results demonstrated that product reviews may contain a significant portion of irrelevant content and that informational status may be influenced by the user profile. The vocabulary in book reviews is not uniform, as the users do not share the same extralinguistic variables, such as age, gender, education and social level, which results in varied writing behavior. In addition, it was found that, for a good domain coverage, at least 40 reviews are required, on average. We also observed that, on average, some domains may have more identifiable aspects. The aspect groups and the hierarchies will be made available for research purposes. We expect that automatic methods for opinion mining may be trained and/or evaluated over such datasets.

Acknowledgments

The authors are grateful to FAPESP and CAPES for supporting this work.

References

ACIR, S.; ZHANG, D.; SIMOFF, S.; DEBENHAM, J. Recommender system based on consumer product reviews. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, 2006. *Proceedings*... Washington: [s/n], 2003, p. 719-723.

AVANCO, L.; NUNES, G. M. V. Lexicon-based Sentiment Analysis for reviews of products in Brazilian Portuguese. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS. *Proceedings*... São Carlos: [s/n], 2014, p. 277-281.

BHUIYANET, T.; XU, Y.; JOSANG, A. state-of-the-art review on Opinion Mining from online customers feedback. In: ASIA-PACIFIC COMPLEX SYSTEMS CONFERENCE, 9. *Proceedings*... Tokyo: [s/n], 2009, p. 385-390.

CONDORI, R. E. L. *Sumarização automática de opiniões baseada em aspectos*. Dissertação (mestrado em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2015.

FREITAS, C.; MOTTA, E.; MILIDIU, R.; CESAR, J. Vampiro que brilha... rá! desafios na anotação de opinião em um córpus de resenhas de livros. In: ENCONTRO DE LINGUÍSTICA DE *CORPUS*, 11. *Anais*... São Carlos: [s/n], 2012.

GOULART, R. R. V.; MONTARDO, S. P. Os mecanismos de busca e suas implicações em Comunicação e Marketing. In: CONGRESSO NACIONAL DE HISTÓRIA DA MÍDIA, 5. *Anais*... São Paulo: [s/n], 2007, p. 478-514.

HARTMANN, N.; AVANÇO, L.; BALAGE, P.; DURAN, M.; NUNES, M. D. G. V.; PARDO, T.; ALUÍSIO, S. A large *corpus* of product reviews in Portuguese: tackling out-of vocabulary words. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 9. *Proceedings*... Reykjavik: [s/n], 2007, p. 3865-3871.

KOCH, I. G. V. *Introdução a Linguística Textual*. 2. ed. São Paulo: Martins Fontes, 2009.

LIU, B. *Sentiment Analysis and Opinion Mining*. 1. ed. San Rafael: Morgan & Claypool Publishers, 2012.

MILLER, G. A., BECKWITH, R., FELBAUM, C., GROSS, D. and MILLER, K. WordNet: An on-line lexical database. *International Journal of Lexicography*, v. 3, p. 235-244, 1990.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment classification using machine learning techniques. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. *Proceedings*... Vol. 10. Stroudsburg: [s/n], 2002, p. 79-86.

TABOADA, M. Sentiment Analysis: an overview from Linguistics. *Annual Review of Linguistics*, v. 2, p. 325-347, 2016.

YU, J.; ZHA, Z.; MENG, W.; WANG, K.; CHUA, T. Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. *Proceedings*... Stroudsburg: [s/n], 2011, p. 140-150.

ZHAO, L.; LI, C. Ontology based Opinion Mining for movie reviews. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE SCIENCE, ENGINEERING AND MANAGEMENT, 3. *Proceedings*... Berlin: Springer-Verlag, 2009, p. 204-214,

ZIPF, G. K. *Human behavior and the principle of least effort*. 1. ed. Cambridge: Addison-Wesley Press, 1949.

Appendix 1

We present below the learning curves for the identification of groups of aspects. As an illustration of how to interpret these graphics, in Figure 2, one may see that, after have analyzed 2 smartphone reviews, we could identify 10 groups of aspects; after 60 reviews, we end up with 48 groups.

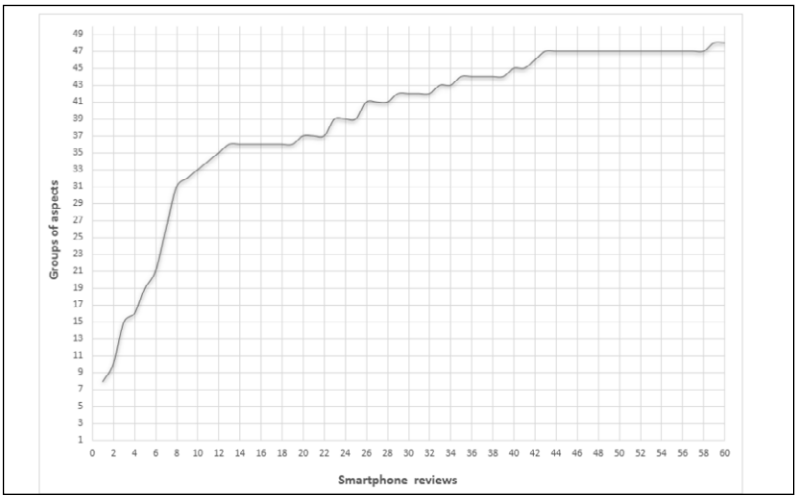


Figure 7 – Learning curve for the smartphone domain

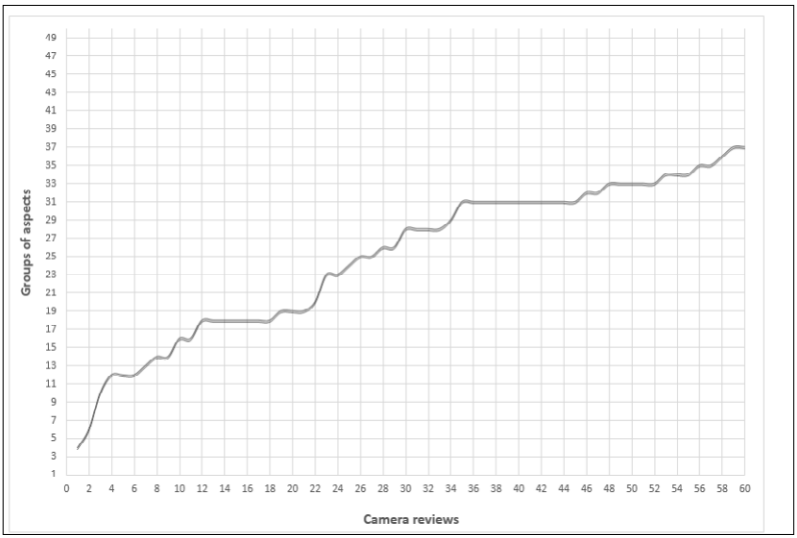


Figure 8 – Learning curve for the camera domain

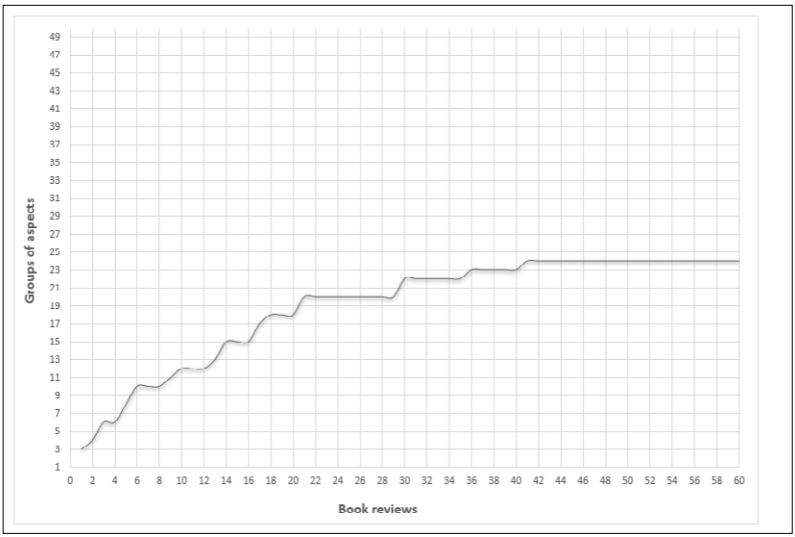


Figure 9 – Learning curve for the book domain

Appendix 2

We present below the hierarchies obtained for the smartphone, camera and book domains, where each circle represents a group of aspects. For each group, we show only the most representative word. We show them in Portuguese because the *corpus* is in this language.

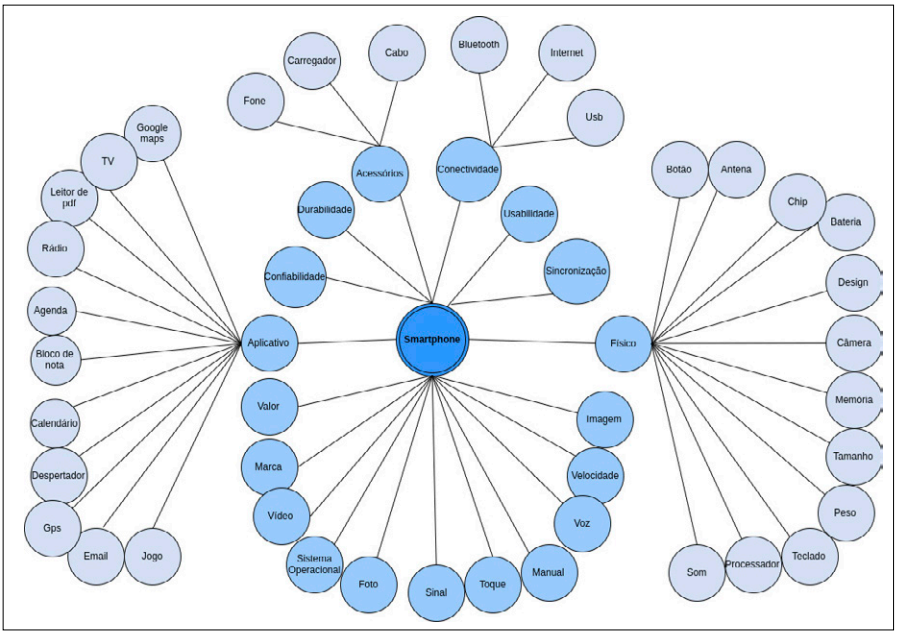


Figure 10 – Hierarchy for the smartphone domain

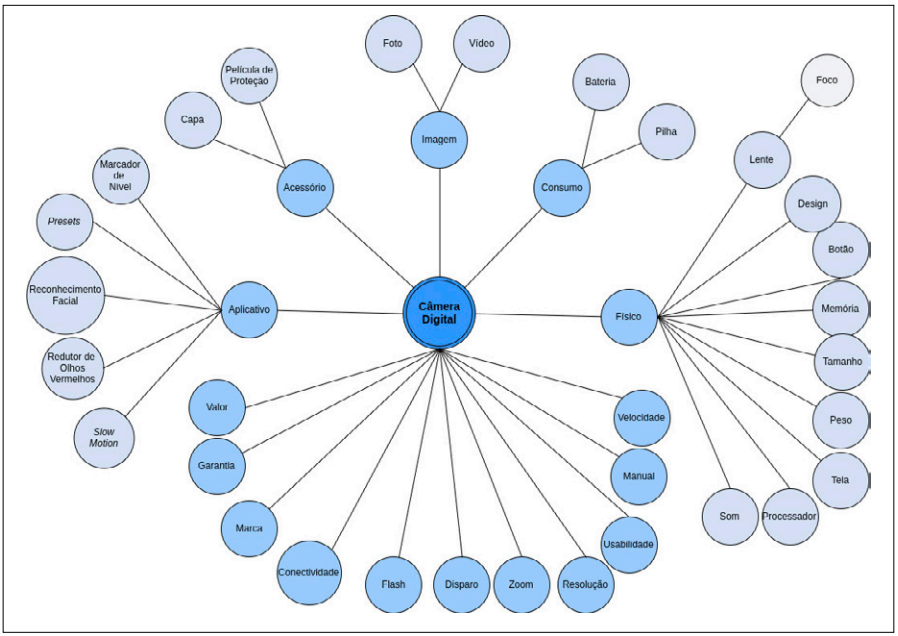


Figure 11 – Hierarchy for the camera domain

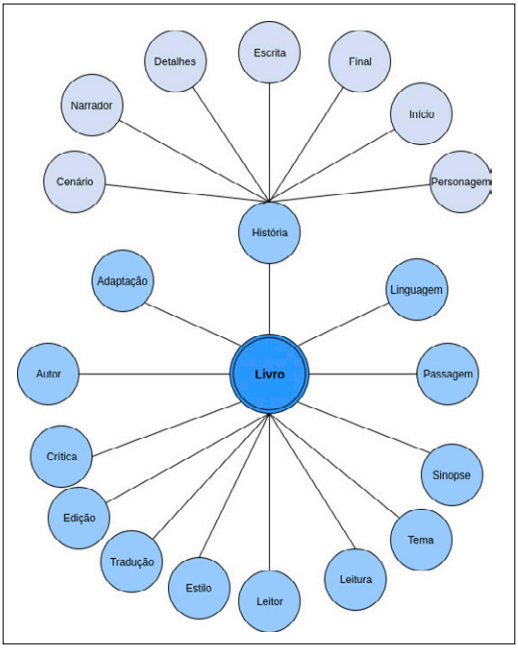


Figure 12 – Hierarchy for the book domain

Appendix 3

We present below the prototypical groups on the smartphone, camera and book domains. We show them in Portuguese because the *corpus* is in this language. Items marked with black color represent prototypical groups.

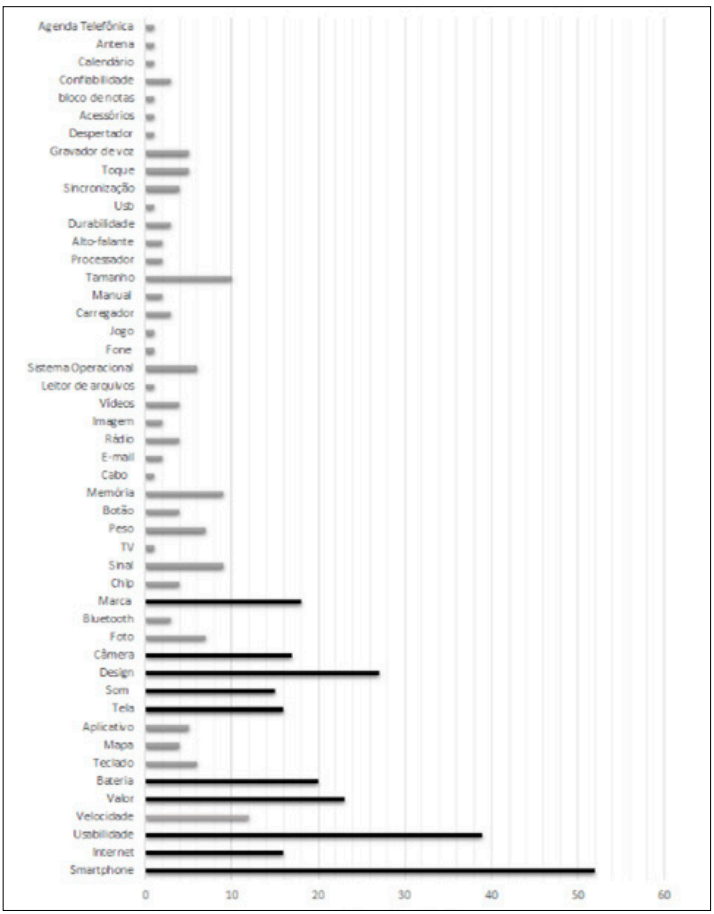


Figure 13 – Prototypical groups for the smartphone domain

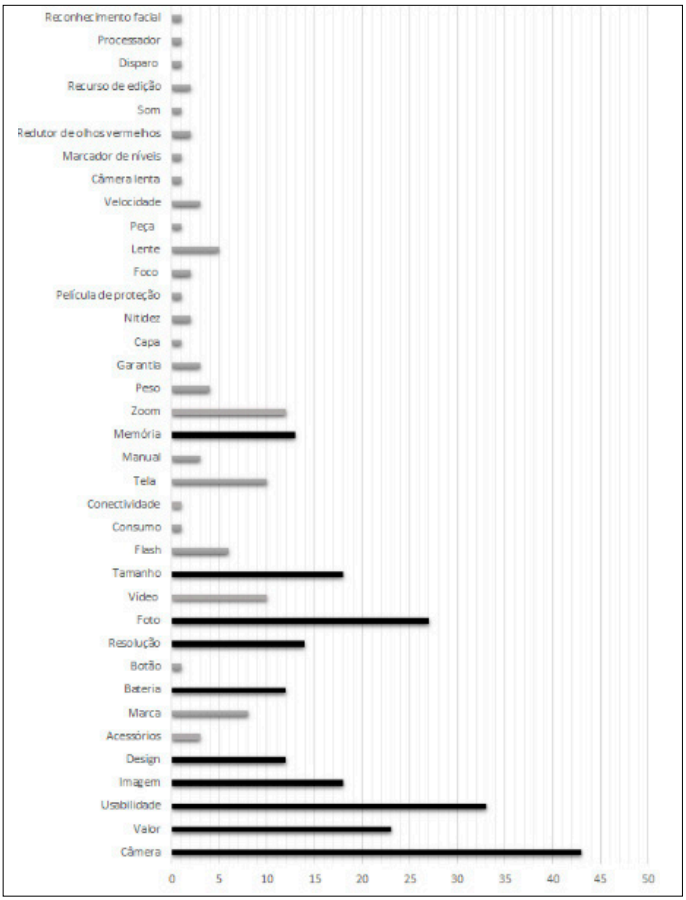


Figure 14 – Prototypical groups for the camera domain

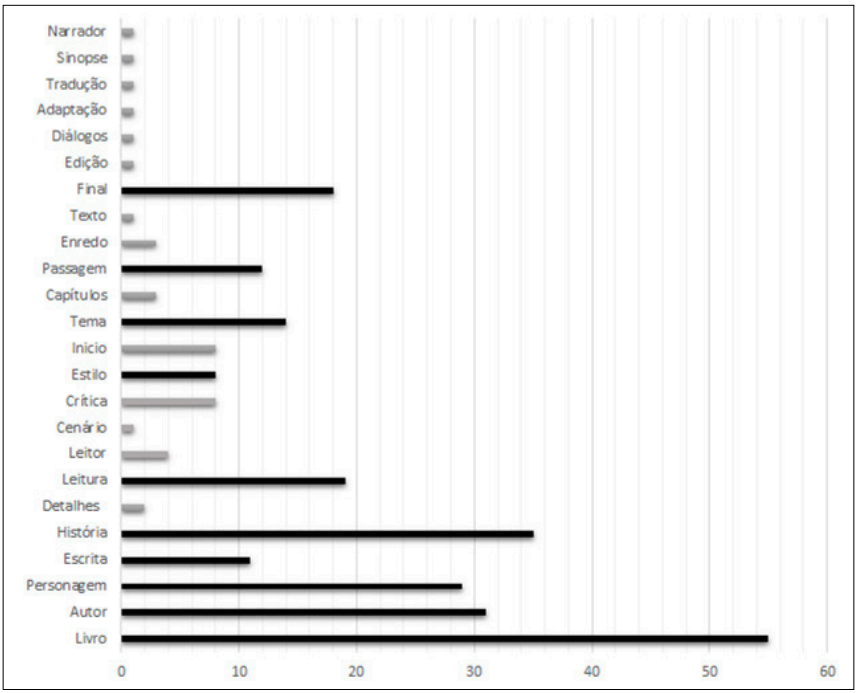


Figure 15 – Prototypical groups for the book domain

Revista Brasileira de Linguística Aplicada: multidimensões temáticas[1]

Brazilian Journal of Applied Linguistics: thematic multi-dimensions

Maria Claudia Nunes Delfino
Rafael Fonseca de Araújo
Tony Berber Sardinha

Resumo: Este capítulo relata um estudo realizado a partir dos artigos publicados na *Revista Brasileira de Linguística Aplicada* (RBLA) com o objetivo de traçar a história da Linguística Aplicada (LA) no Brasil por meio da análise do léxico mais saliente extraído de um *corpus* formado por textos publicados no periódico entre 2001 e 2015. Foi empregado um método multidimensional baseado em *corpus* para extrair os agrupamentos de palavras mais proeminentes que marcam os principais períodos da Linguística aplicada no Brasil.

Palavras-chave: Análise multidimensional. Dimensões temáticas. Linguística de *Corpus*. *Revista Brasileira de Linguística Aplicada*.

Maria Claudia Nunes Delfino – Professora da Faculdade de Tecnologia de Praia Grande, mestre em Linguística Aplicada pela Universidade Católica de São Paulo – claudia@corpuslg.org.
Rafael Fonseca de Araújo – Professor da Escola Técnica Estadual Alberto Santos Dumont e Universidade Metropolitana de Santos, mestre em Linguística Aplicada pela Universidade Católica de São Paulo – rafael@corpuslg.org.
Tony Berber Sardinha – Professor Titular da Pontifícia Universidade Católica de São Paulo, doutor em Inglês pela Universidade de Liverpool – tony@corpuslg.org.

[1] O presente capítulo foi realizado por membros do GELC (Grupo de Pesquisa em Linguística de *Corpus*) no LAEL da PUC-SP com agradecimento especial à Professora Emérita Maria Antonieta Alba Celani pelas valiosas orientações e apoio sem os quais não seria possível sua realização.

Abstract: This chapter reports a study carried out on the articles published the Brazilian Journal of Applied Linguistics (RBLA) with the objective of outlining the history of Applied Linguistics (LA) in Brazil through the analysis of the most salient lexis extracted from a *corpus* formed by the texts published in the journal between 2001 and 2015. We employed a multidimensional *corpus*-based method to extract the most prominent clusters of words marking the major periods in Applied Linguistics history in Brazil.

Keywords: Multidimensional Analysis. Thematic dimensions. *Corpus* Linguistics. Brazilian Journal of Applied Linguistics.

1 Introdução

Em todo o mundo, pesquisadores de diversas áreas acadêmicas publicam artigos em inúmeras revistas científicas que, por sua vez, são os principais instrumentos de divulgação de estudos com o intuito de fomentar pesquisas, apresentar e discutir resultados. Por meio do exame das publicações dessas revistas científicas, é possível observar as tendências que dão suporte teórico-metodológico a diferentes pesquisas, as quais podem variar ao longo do tempo. Também é possível identificar os temas de maior relevância nas comunidades científicas nas quais se circunscrevem.

No Brasil, a *Revista Brasileira de Linguística Aplicada* (RBLA), publicação sem fins lucrativos, é um periódico trimestral, com avaliação por pares, cuja missão é incentivar a pesquisa em Linguística Aplicada (LA). Criada em 2001, a revista recebe artigos originais, de mestres e doutores, que tratam dos muitos fenômenos relacionados a problemas de linguagem da vida real pertinentes à língua em uso em contextos diversos ou à aprendizagem. O periódico também publica resenhas, entrevistas e dois números temáticos por ano. Com apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação de Amparo à Pesquisa do Estado de Minas Gerias (FAPEMIG), a publicação é de responsabilidade do Programa de Pós-Graduação em Estudos Linguísticos, área de concentração em Linguística Aplicada, da Faculdade de Letras da Universidade Federal de Minas Gerais (UFMG), distribuída gratuitamente aos sócios da ALAB. A revista possui a classificação Qualis A1, ou seja, é um periódico de mérito reconhecido, nacional e internacionalmente. É uma publicação que passou por normas e requisitos da *Scientifc Eletronic Library Online* (SciELO), tendo obtido nota máxima.

O presente capítulo relata um estudo que teve por objetivo identificar os temas mais recorrentes de interesse da Linguística Aplicada no Brasil, por meio da investigação do léxico mais saliente nas publicações da RBLA e verificar se há variação temática ao longo do tempo. Para este fim, utilizou-se a Análise Multidimensional (AMD), abordagem para análise de *corpora* que emprega uma série de procedimentos estatísticos multivariados (principalmente análise fatorial) para determinar parâmetros de coocorrência entre características linguísticas (BIBER, 1988). Mais especificamente, foi empregada a variante lexical da

AMD, que utiliza itens lexicais para identificação dos agrupamentos linguísticos de interesse (BERBER SARDINHA, 2014). Os resultados da AMD empregada apontaram dez dimensões temáticas que refletem assuntos mais abordados nas pesquisas em LA no Brasil ao longo de 15 anos de publicações da RBLA.

2 Análise Multidimensional

A AMD é realizada a partir de levantamentos estatísticos de características linguísticas, realizadas por computador e foi desenvolvida por Douglas Biber com o intuito de investigar a variação linguística da língua inglesa na fala e na escrita. Segundo Berber Sardinha (2004, p. 300) a AMD:

> É uma abordagem para a análise de *corpus* que usa procedimentos estatísticos (principalmente análise fatorial), visando o mapeamento das associações entre um conjunto variado de características linguísticas dentro do *corpus* de estudo. Também usa procedimentos automáticos e semiautomáticos para análise do *corpus*, tais como etiquetagem morfossintática (*part of speech tagging*).

Biber (2001) destaca que essa abordagem de análise de *corpus* tem o texto como unidade de análise básica e parte do pressuposto que diferentes registros diferem entre si em termos da frequência dos padrões de coocorrência de seus traços léxico-gramaticais. Tais padrões, por sua vez, definem espaços linguísticos a partir das funções que os subjazem, criando as dimensões de variação (ZUPPARDO, 2014, p. 9). Essa análise admite que parâmetros funcionais múltiplos de variação operem em qualquer domínio do discurso, partindo da hipótese de que os tipos de textos difiram entre si linguística e funcionalmente. E é multidimensional porque admite-se como pressuposto que parâmetros múltiplos de variação operem em qualquer domínio do discurso.

Biber partiu da perspectiva de que os falantes de uma língua têm conhecimento sobre a estrutura e o uso da língua (BIBER, 1988, p. 8), para usá-la de acordo com as exigências funcionais e situacionais. Com essa abordagem, o autor investigou a variação na língua inglesa (LI) confrontando textos dos mais variados registros, para delinear as características linguísticas típicas de cada registro. Ele obteve uma descrição abrangente da LI e esclareceu como os vários registros variam por dimensões que, de acordo com Berber Sardinha (2014), são os parâmetros funcionais de variação formados por padrões de coocorrência de elementos léxico-gramaticais interpretados com base em suas funções comunicativas latentes, ou seja, correspondem a um construto que compreende um conjunto de variedades linguísticas que podem explicitar aspectos funcionais da linguagem em uso e que influenciam a construção de sentidos nos mais variados tipos de textos. Ainda segundo Berber Sardinha (2000a, p. 106):

> Uma dimensão permite visualizar características em comum partilhadas por uma porção significativa dos dados. A interpretação do fator leva em conta tanto as características linguísticas quanto as características partilhadas pelos registros que estão representados no fator. As dimensões permitem redefinir o quadro de registros inicial.

Biber (1988) identificou cinco dimensões de variação que capturam características léxico-gramaticais e discursivas em um *corpus* formado por diversos registros escritos e falados que revelam aspectos funcionais importantes inerentes à língua inglesa, a saber: (1) produção com interação *versus* informação, (2) preocupações narrativas *versus* não narrativas, (3) referência explícita *versus* dependente do contexto, (4) persuasão explícita e (5) informação abstrata *versus* não abstrata.

Na AMD acredita-se que um grupo de características não coocorre frequentemente nos textos de forma aleatória; ao contrário, a coocorrência dessas variáveis se dá de forma sistemática "porque atendem a alguma função comunicativa comum e a interpretação da dimensão funcional subjacente às dimensões é uma hipótese e deve ser confirmada através da análise qualitativa dos textos que compõem um *corpus*" (BIBER, 1988, p. 91).

No Brasil, seguindo os passos de Biber (1988 et. seq.), Berber Sardinha, Kauffmann e Acunzo (2014a) identificaram as dimensões de variação do português brasileiro, por meio da investigação dos padrões de características linguísticas coocorrentes mais salientes no *Corpus* Brasileiro de Variação de Registro (CBVR). As seis dimensões de variação do português brasileiro são: (1) discurso oral *versus* letrado, (2) argumentação, (3) produção com envolvimento *versus* produção com foco informacional, (4) discurso procedural, (5) orientação temporal para o futuro *versus* orientação temporal para o passado e (6) discurso relatado.

Há também trabalhos que recentemente empregam a AMD em registros específicos. Bértoli-Dutra (2014) investigou a variação linguística em letras de música *pop* escritas em língua inglesa. Veirano Pinto (2014), por sua vez, investigou a variação nos principais gêneros cinematográficos norte-americanos no AMC (*American Movies Corpus*). Zuppardo (2014) empregou a AMD para identificar as dimensões de variação no CAM (*Corpus of Aircrafts Manual*), formado por manuais de manutenção de aeronaves. Berber Sardinha e Veirano Pinto (no prelo) realizaram um estudo sobre a linguagem da televisão norte-americana, reunindo no USTV *Corpus* transcrições dos principais tipos de programas televisivos veiculados na televisão nos EUA, entendidos como registros independentes, identificando dimensões de variação entre os programas.

Outros estudos não identificaram novas dimensões de variação, contudo analisaram e comparam novos e diferentes registros em dimensões de variação identificadas anteriormente, ou seja, realizam um mapeamento (adição) de registros ao longo de dimensões preexistentes, o que se entende por AMD aditiva. Berber Sardinha (2014) comparou registros da internet com os registros

tradicionais da língua inglesa ao longo das cinco dimensões de variação supracitadas identificadas por Biber (1988). Delfino (2016) empregou a AMD aditiva em um *corpus* formado por letras de músicas *pop* (CoEL). Fonseca de Araújo (2017) além de mapear um *corpus* formado por textos de *reality TV shows* norte-americanos (CARTS) nas dimensões de variação do inglês, também realizou uma AMD aditiva nas dimensões de variação da televisão norte-americana identificadas por Berber Sardinha e Veirano Pinto (no prelo). Barreto (2016) analisou um *corpus* formado por redações de vestibulares da Universidade Federal do Rio Grande do Norte (Vesti*Corpus*) nas dimensões do português brasileiro identificadas por Berber Sardinha, Kauffmann e Acunzo (2014b).

3 Análise Multidimensional Lexical

Em 2014, Berber Sardinha propôs um novo modelo para a Análise Multidimensional apresentada por Biber em 1988 que se baseia não na interpretação funcional, mas na interpretação semântica dos fatores, entendida como Análise Multidimensional Lexical (AMDL), com o intuito de encontrar dimensões de variação lexical, identificadas por meio da interpretação dos campos semânticos subjacentes à coocorrência do léxico mais saliente. Para tanto, o autor investigou o uso dos adjetivos *American* e *Brazilian*, bem como seus colocados – palavras que ocorrem perto do nódulo – para identificar os parâmetros de representação de identidade nacional e cultural por meio do qual os EUA e o Brasil são representados nas produções textuais em inglês a partir do século XIX disponibilizados pelo Google Books.

Na perspectiva lexical, a Análise Multidimensional, empregada no presente estudo, considera como variáveis (unidade de análise) apenas as palavras de conteúdo ou multipalavras para a identificação das dimensões de variação. A Tabela 1, a seguir, ilustra os aspectos similares e distintos entre a AMD funcional e lexical.

Tabela 1 – Aspectos da AMD funcional e lexical[2]

	Funcional	**Lexical**
Objetivo	Identificar parâmetros subjacentes de variação nos textos de um *corpus*	
Unidade de observação	Textos ou segmentos de texto	Palavras, colocações
Traços linguísticos	Léxico-gramaticais	Lexicais
Base da interpretação	Funcional, comunicativa	Campos semânticos, preferência semântica, "*aboutness*"[2]

Fonte: Berber Sardinha (2017)

[2] O termo "*aboutness*" abrange, além do tópico dos textos, a representação construída através das características lexicais. Nesse sentido, pode-se determinar o posicionamento do(s) autor(es) do(s) texto(s) a partir da sua análise *lexical*.

Como parte do projeto de criação de um dicionário de colocações do português brasileiro, Berber Sardinha, Acunzo e São Bento Ferreira (2014a) empregaram o modelo da AMD lexical para identificar dimensões de colocação como parâmetros comunicativos subjacentes às escolhas das colocações. Investigando o *Corpus* do português brasileiro (CBVR), os autores identificaram oito dimensões, a saber: (1) mitigação e cognição, (2) agropecuária e alimentação, (3) probabilidade, (4) linguagem oral *versus* termos inerentes à internet, (5) ciência e tecnologia, (6) economia *versus* emoções, (7) discurso médico *versus* política, administração pública e governo, e (8) criminalidade e segurança pública. Outro estudo, também realizado por Berber Sardinha, Acunzo e São Bento Ferreira (2014b), investiga as metáforas nas colocações do português brasileiro, encontrando cinco dimensões temáticas: (1) Questões salariais e custo de vida, (2) Questões orçamentárias, (3) Questões tributárias, (4) Questões cambiais e (5) Índices econômicos.

Recentemente, Berber Sardinha realizou dois trabalhos empregando a AMD lexical para identificação da variação lexical em periódicos científicos internacionalmente importantes na LA. O primeiro deles (BERBER SARDINHA, 2016) identificou dimensões de variação lexical em uma análise diacrônica dos principais períodos históricos do TESOL *Quarterly* em um *corpus* formado pelas publicações de 1967 a 2014. O segundo estudo (BERBER SARDINHA, 2017), apresentado no AILA no Rio de Janeiro, além de identificar os temas mais recorrentes da LA mundial, observou as principais tendências e mudanças nos 72 anos de história da área científica por meio da análise de um *corpus* formado pelas publicações dos principais periódicos internacionais: *Applied Linguitics*, ELT *Journa*l, IRAL, *Language Leraning* e TESOL *Quarterly*. Berber Sardinha elaborou essa vertente da Análise Multidimensional de Biber a partir da noção de que os textos podem variar por assunto, propósito, estrutura retórica e estilo, a análise lexical dos textos de um *corpus* pode revelar os assuntos mais recorrentes, refletidos pelos campos semânticos formados por padrões de coocorrência de itens lexicais, interpretados qualitativamente em termos de dimensões de variação lexical (BIBER, 1988, p. 70).

Os estudos de Berber Sardinha descritos acima inspiraram este estudo, que tem por objetivo promover um panorama histórico sobre os temas mais recorrentes de interesse da LA por meio da AMDL dos textos publicados nos primeiros quinze anos da RBLA. Espera-se que a identificação das dimensões lexicais da RBLA revele o estado da arte da LA brasileira por meio das publicações da revista.

4 Metodologia

Conforme mencionado na introdução, a AMD pode ser definida como uma abordagem para análise de *corpus* que utiliza procedimentos estatísticos (mais especificamente a análise fatorial) e que se ocupa do mapeamento das associações

entre conjunto variado de características linguísticas dentro do *corpus* de estudo (BERBER SARDINHA, 2004). Base metodológica deste estudo, a Análise Multidimensional Lexical (AMDL) proposta por Berber Sardinha (2014 et seq.), tem o intuito de observar a variação linguística identificando os campos semânticos subjacentes formados pela coocorrência do léxico mais saliente em *corpus*. É uma adaptação do modelo de AMD desenvolvida inicialmente por Biber (1988 et seq.) para a análise e observação da variação léxico-gramatical em registros falados e escritos de uma língua ou variedade linguística.

O *corpus* utilizado neste estudo, *Corpus* da *Revista Brasileira de Linguística Aplicada* (CRBLA), é composto por toda a coleção da revista desde seu primeiro número até o último número de 2015; 361 textos (305 artigos, 35 resenhas, 17 cartas do editor e 4 entrevistas), totalizando 2,3 milhões de palavras (*tokens)* e 66.814 tipos (*word types)* conforme a Tabela 2.

Tabela 2 – Composição do *corpus* RBLA

Ano	Registro	Textos	*Tokens*
2001	Artigo	4	29.743
	Carta do editor	1	1.939
	Resenha	2	2.343
2002	Artigo	12	48.031
	Carta do editor	1	1.503
	Resenha	2	3.269
2003	Artigo	15	89.586
	Carta do editor	1	779
	Resenha	1	972
2004	Artigo	17	115.625
	Carta do editor	3	2.204
2005	Artigo	18	137.930
	Carta do editor	2	1.153
	Entrevista	1	4.590
2006	Artigo	14	90.927
	Carta do editor	2	1.893
2007	Artigo	17	133.601
	Carta do editor	2	1.507
2008	Artigo	18	145.966
	Carta do editor	1	763
	Resenha	1	5.674
2009	Artigo	24	182.424
	Carta do editor	2	1.630
	Resenha	1	1.613
2010	Artigo	36	256.861
	Carta do editor	4	2.426
	Entrevista	1	5.644
	Resenha	3	7.498

(Continua)

2011	Artigo	23	169.928
	Carta do editor	5	2.703
	Entrevista	1	4.417
	Resenha	1	1.174
2012	Artigo	32	237.092
	Carta do editor	4	3.772
	Entrevista	1	7.046
	Resenha	3	4.240
2013	Artigo	33	248.353
	Carta do editor	3	1.392
2014	Artigo	31	264.208
	Carta do editor	4	2.272
	Resenha	3	12.485
2015	Artigo	11	90.707
TOTAL	Artigo	305	2.240.982
	Carta do editor	35	25.936
	Entrevista	4	21.697
	Resenha	17	39.268
		361	2.327.883

Fonte: Elaborado pelos autores (2017)

O levantamento de dimensões temáticas envolve as seguintes etapas (BERBER SARDINHA, 2004 et seq.): (1) definição de um *corpus* representativo do domínio em questão, em nosso caso todo o acervo da RBLA publicado entre 2001 e 2015 disponível na internet; (2) identificação e contagem do léxico pertinente; (3) normalização das frequências das variáveis (léxico), a fim de nivelar os textos de maior e menor extensão; (4) extração fatorial inicial baseada nas frequências normalizadas; (5) estabelecimento do número de fatores latentes nos dados por meio da análise de um gráfico de sedimentação (*scree plot*); (6) eliminação de variáveis cujas comunalidades estejam abaixo de 0,2; (7) extração fatorial final rotacionada contendo o número de fatores estabelecidos; (8) cômputo da quantidade de variação compartilhada pelos fatores extraídos; (9) retirada do léxico com peso abaixo de 0,3 em qualquer um dos fatores; (10) criação de um padrão fatorial contendo apenas as variáveis que carregaram em cada fator; (11) cômputo das correlações interfatoriais a fim de estabelecer o grau de independência dos fatores; (12) interpretação dos fatores em termos predominantes, incluindo para isso o exame dos textos em que ocorrem para que os fatores assumam o *status* de dimensões; (13) cálculo dos valores padronizados (Z-scores) de cada variável, a fim de nivelar as variáveis de maior e menor ocorrência; (14) cálculo de escores de fator para cada texto a partir dos pesos das variáveis de cada fator; (15) cálculo do escore médio de cada fator em cada período de tempo e (16) cômputo da quantidade de variação explicada pelos fatores por meio de Análise de Variância (ANOVA).

Seguindo as etapas supracitadas, inicialmente o *corpus* foi coletado por meio de um *script* em *Unix Shell*, especialmente desenvolvido para este estudo, que baixou e converteu para formato de texto todo o conteúdo da RBLA existente no *site* scielo.br/rbla. O léxico foi identificado do seguinte modo: para cada ano da revista, foram selecionadas as 100 palavras mais frequentes por meio da ferramenta "lista de palavras" gerada pelo concordanciador AntConc e normalizadas para uma razão de mil ocorrências. Para exemplificar a normalização, tomemos o caso da palavra "abordagem", que teve frequência 91 no ano 2001. Nesse ano, o total de palavras é 34.584. Sendo assim, a frequência normalizada da palavra "palavrasponas dimensabordagem" no ano 2001 é 2,6, pois [91 x 1.000 / 34.584 = 2.631] em relação ao total de palavras de cada ano.

Em seguida, essas listas foram combinadas e foram retiradas as 200 palavras mais frequentes de todos os anos da revista, as quais foram processadas estatisticamente. A lista foi transferida para uma tabela, elaborada em uma planilha de dados do programa Excel (arquivo em formato *.xlx), parte da suíte de programas do Windows Microsoft Office. A extração inicial foi realizada no programa SPSS 23 para *Windows*, com o método de fatoração pelo eixo principal – *Principal Axis Factoring* (*PAF*) (GÓMEZ, 2013). Essa extração gerou um gráfico de sedimentação (*scree plot*) que revelou a existência de dez fatores, conforme mostra a Figura 1. O número de fatores transparece no gráfico por meio da existência de um patamar em que os valores tendem a diminuir seu ritmo de queda. Nessa pesquisa, esse patamar encontra-se entre os fatores 9 e 11. Por isso, optamos pela extração dos dez fatores como melhor solução.

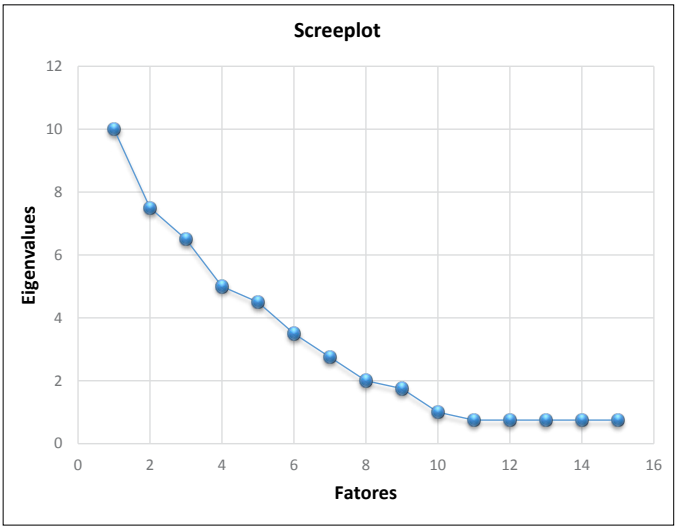


Figura 1 – Gráfico de sedimentação mostrando a existência de dez fatores latentes
Fonte: Elaborado pelos autores (2017)

As variáveis com valor de comunalidade abaixo de 0,2 foram retiradas, por não possuírem representatividade, restando, assim, 213 variáveis no *pool*. Para a rotação dos fatores, foi empregado o método *Promax*, que é um método oblíquo de rotação o qual possibilita correlação entre fatores, que presume a sobreposição de variáveis, característica desejável em um estudo linguístico, haja vista que a linguagem resulta da imbricação de diversos traços léxico-gramaticais (HOEY, 2006, p.5-7). A quantidade de variação compartilhada pelos dez fatores é de 25,4%, conforme mostra a Tabela 3.

Tabela 3 – Percentual de variação

Fator	Variação (%)
1	4,678
2	8,148
3	11,367
4	14,009
5	16,354
6	18,366
7	20,274
8	22,103
9	23,837
10	25,438
Total	25,438

Fonte: Elaborado pelos autores (2017)

Em seguida, foi aplicado o ponto de corte de peso de 0,3 nas variáveis de cada fator (ponto de corte tradicionalmente utilizados em pesquisas com aplicação de análise fatorial). Uma mesma variável somente foi incluída no cálculo dos escores de um fator apenas. As variáveis com peso igual ou superior ao ponto de corte com ocorrência em mais de um fator foram identificadas no padrão fatorial de ambos, porém identificadas por parênteses no fator em que têm peso menor, sendo consideradas para a interpretação das dimensões.

A intercorrelação, que evidencia o grau de interdependência entre os fatores, é ilustrada na Tabela 4, sugerindo que os fatores são independentes, o que é desejado.

Tabela 4 – Matriz de correlações de fator

Fator	1	2	3	4	5	6	7	8	9	10
1	1,000									
2	-0,139	1,000								
3	0,394	-0,139	1,000							
4	0,116	0,009	0,253	1,000						
5	0,099	0,000	-0,104	-0,220	1,000					
6	-,024	-0,112	-0,017	-0,059	0,010	1,000				
7	0,250	0,030	0,248	0,190	-0,064	0,117	1,000			
8	0,115	-0,115	0,143	0,036	-0,067	0,046	0,206	1,000		
9	0,029	0,090	0,016	0,093	0,133	-0,067	0,151	0,040	1,000	
10	0,129	-0,029	0,185	0,041	0,021	0,082	0,108	0,100	-0,004	1,000

Fonte: Elaborado pelos autores (2017)

Os dados mostram que os fatores mais intercorrelacionados são os de números 1 e 3 (0,394), enquanto os mais independentes são os de números 2 e 5 (0,000). Os demais passos da análise são detalhados na seção de resultados, a seguir.

5 Dimensões temáticas

Conforme colocado na seção de metodologia, a análise revelou a existência de dez fatores nos dados. Esses fatores foram interpretados com base em critérios linguísticos e, após a sua interpretação com a inspeção dos dez textos com maior peso em cada dimensão, foram chamados de dimensões que encapsulam os temas principais tratados pelos textos que carregaram no fator, respeitando a presença das variáveis linguísticas mais salientes em cada dimensão e os temas a elas associados nos textos inspecionados e analisados pelos autores. A Tabela 5 ilustra as variáveis linguísticas que compuseram cada um dos fatores.

Tabela 5 – Composição das dimensões

Dimensão 1	+	formação (0,794), profissional (0,701), professor (0,657), professores (0,649), prática (0,633), letras (0,493), docente (0,474), reflexão (0,469), cursos (0,443), universidade (0,411), desenvolvimento (0,353), ensino (0,347), curso (0,325)
Dimensão 2	+	Textos (0,570), gênero (0,543), textuais (0,531), Texto (0,531), textual (0,518), gêneros (0,510), discurso (0,486), análise (0,484), produção (0,476), linguagem (0,384), discursivos (0,363), discursiva (0,344), leitura (0,324), escrita (0,324), perspectiva (0,323), partir (0,316), crítica (0,311), proposta (0,306)
	-	práticas (-0,343)
Dimensão 3	+	Estudo (0,639), aprendizagem (0,558), artigos (0,499), estudos (0,481), resultados (0,440), apresenta (0,437), processos (0,402), línguas (0,337), estrangeiras (0,317), trabalhos (0,301)
Dimensão 4	+	Temático (0,893), número (0,688), grande (0,684), comunidade (0,682), pesquisadores (0,643), *corpus* (0,632), área (0,593), leitores (0,556), poder (0,411), futuro (0,388), (universidade (0,330))
Dimensão 5	+	Aula (0,693), alunos (0,638), sala (0,629), atividades (0,477), aulas (0,436), professora (0,434), escola (0,375), aluno (0,357), escolar (0,342), atividade, (professor (0,334)), (0,323), aprender (0,313), (aprendizagem (0,313)), ações (0,303)
Dimensão 6	+	(texto (0,317)), Informações (0,312)
	-	Sociais (-0,608), social (-0,582), vida (-0,466), identidades (-0,441), pessoas (-0,383), práticas (-0,361), identidade (-0,345), sociedade (-0,314), relações (-0,314), (poder (-0,311))
Dimensão 7	+	Didático (0,880), livro (0,857), didáticos (0,807), livros (0,799), estrangeira (0,445), material (0,389)
Dimensão 8	+	Surdos (0,830), libras (0,829), sinais (0,700), surdo (0,620), língua (0,580), português (0,402), portuguesa (0,355)
Dimensão 9	+	Tecnologias (0,645), digital (0,593), tecnologia (0,472), novas (0,437), letramento (0,391), graduação (0,389), letramentos (0,347), educação (0,339), comunicação (0,329), acesso (0,328), (práticas (0,328)), (cursos (0,316)), (universidade (0,316)), uso (0,305)
Dimensão 10	+	Específicos (0,804), instrumental (0,681), abordagem (0,587), projeto (0,352), (desenvolvimento (0,345)), contexto (0,327)

Fonte: Elaborado pelos autores (2017)

5.1 Dimensão 1: Ensino e Formação de Professores

A Dimensão 1, composta por um polo positivo, tem como principais pesos itens lexicais como *formação* (0,794) e *profissional* (0,701). Ao analisarmos os textos com maiores pesos nessa dimensão, percebemos que os temas principais são ensino e formação de professores. Um exemplo é um artigo de Dutra, de 2006, que aborda o tema ensino e formação de professores (vide Exemplo 1). Com base nisso, sugerimos o rótulo interpretativo *Ensino e Formação de Professores* para a dimensão.

Exemplo 1:
A pesquisa qualitativa centrada na relação aluno-**professor** é baseada na linha de **desenvolvimento** reflexivo do **professor**, destacando como vários tipos de experiências no **desenvolvimento** dão suporte à sua **formação profissional.** (DUTRA, 2006)

Essa dimensão teve escores maiores nos anos 2006 e 2004, conforme mostra a Figura 2. O teste da análise de variância (ANOVA) indica que aproximadamente 80% da variação encontrada nessa dimensão é explicada pela variável temporal e o valor de *p*[3] foi de 0,010 (<0,05), revelando que há diferença estatística significativa, ou seja, pode-se dizer que houve mudança com o passar dos anos nos temas relacionados a ensino e formação de professores. Ainda, de acordo com a Figura 2, percebemos que esse não é mais um tema frequente nas últimas publicações da revista.

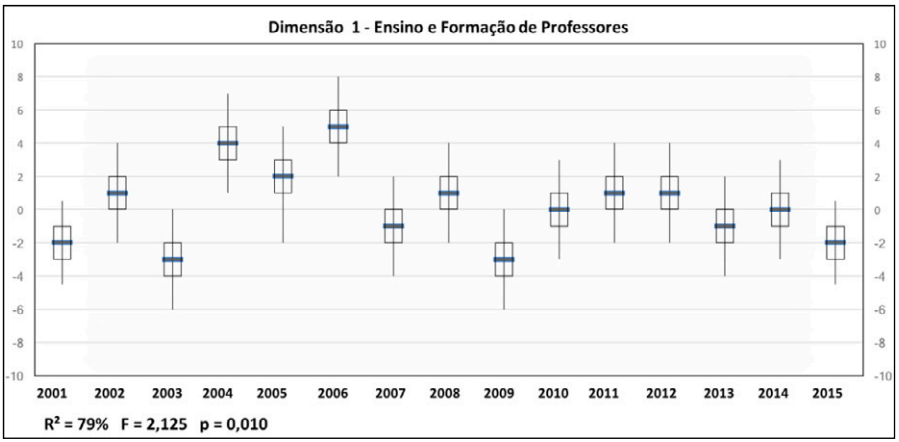


Figura 2 – Escores médios dos anos na Dimensão 1
Fonte: Elaborado pelos autores (2017)

Corroborando os achados, Celani (2008, p. 27) aborda a mudança na visão de como o ensino e o papel do professor como profissional era visto pela sociedade ao retratar que o aluno é o nosso cliente, porém não se poderia dar "o mesmo tipo de relacionamento daquele do profissional da medicina ou do direito ou da

[3] O valor ***p*** (também chamado de nível descritivo ou probabilidade de significância) é a probabilidade de se observar um valor da **estatística** de teste maior ou igual ao encontrado.

administração" promovendo uma discussão que valoriza uma formação contínua do docente.

5.2 Dimensão 2: Texto, Gênero e Discurso

A Dimensão 2, composta por um polo positivo, inclui léxico que indica três temas predominantes: *texto, gênero* e *discurso*. As palavras com maior destaque são *textos* e *gênero*, respectivamente com 0,570 e 0,543 de peso. Ao analisarmos os textos com maiores pesos nessa dimensão, percebemos que os assuntos tratados giram em torno de tipos de textos e discurso. Um exemplo dessa dimensão é o artigo de Pinheiro, *Práticas de produção textual no MSN Messenger: ressignificando a escrita colaborativa,* de 2010:

> Exemplo 2:
> O advento da tecnologia digital, por exemplo, fez com que **gêneros discursivos** sofressem adaptações: encurtamento dos **textos**, uso de *links* eletrônicos, uso da hipermídia, entre outros. (PINHEIRO, 2010)

Essa dimensão teve escores maiores nos anos 2010 e 2012, conforme mostra a Figura 3. A ANOVA indica que apenas 11% da variação encontrada nesta dimensão é explicada pela variável temporal. No entanto, o valor de *p* foi de 0,00 (<0,05), o que indica que a noção de texto, gênero e discurso variou sistematicamente em relação ao tempo, mesmo sendo um tema frequente em todas as publicações da revista.

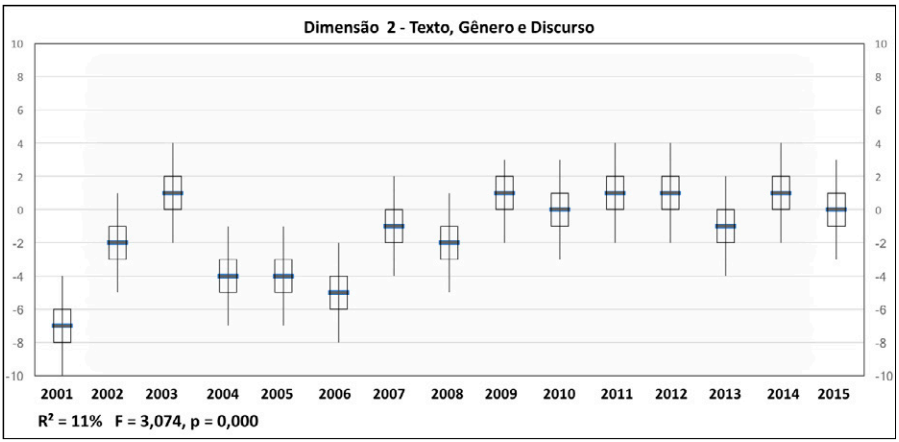


Figura 3 – Escores médios dos anos na Dimensão 2
Fonte: Elaborado pelos autores (2017)

A literatura também apresenta as noções de gênero, texto e discurso como estável, ou seja, presente em publicações ao longo do tempo, como defende Sobral (2008, p. 13), quando afirma que a escolha do gênero, depende da relação entre os interlocutores, e também "das situações sócio-históricas específicas em que se dá essa relação" e, como cada época tem a sua própria situação sócio-histórica, podemos afirmar que a temática *Texto, Gênero e Discurso* necessita ser revisitada constantemente.

5.3 Dimensão 3: Aprendizagem de Língua

A Dimensão 3, composta por um polo positivo, inclui léxico que indica dois temas predominantes: aprendizagem e ensino de línguas. As palavras com maior destaque são *estudo* e *aprendizagem*, respectivamente com 0,639 e 0,558 de peso. Ao analisarmos os textos com maiores pesos nesta dimensão, percebemos que os assuntos tratados giram em torno de aprendizagem de línguas. Um exemplo dessa dimensão é o artigo de Conceição, *Experiências de aprendizagem: reflexões sobre o ensino de língua estrangeira no contexto brasileiro,* de 2006:

Exemplo 3:
O **estudo**, uma proposta de investigação da **aprendizagem** de vocabulário, na qual as relações entre as experiências de **aprendizagem** dos informantes, suas crenças e suas ações na **aprendizagem** são consideradas, aponta uma perpetuação de conceitos de ensino da Abordagem Tradicional ou Gramática e Tradução, ainda presente em nossas salas de aula, através de um ensino focado na leitura e interpretação de textos, uso do dicionário, tradução e listas de palavras para memorização. (CONCEIÇÃO, 2006)

Essa dimensão teve escores maiores nos anos 2006 e 2012, conforme mostra a Figura 4. A ANOVA indica que apenas 3,9% da variação encontrada nesta dimensão é explicada pela variável temporal. E o valor de *p* foi de NS[4] para a variável temporal, mostrando que os temas envolvendo aprendizagem de línguas não mudaram com o passar do tempo.

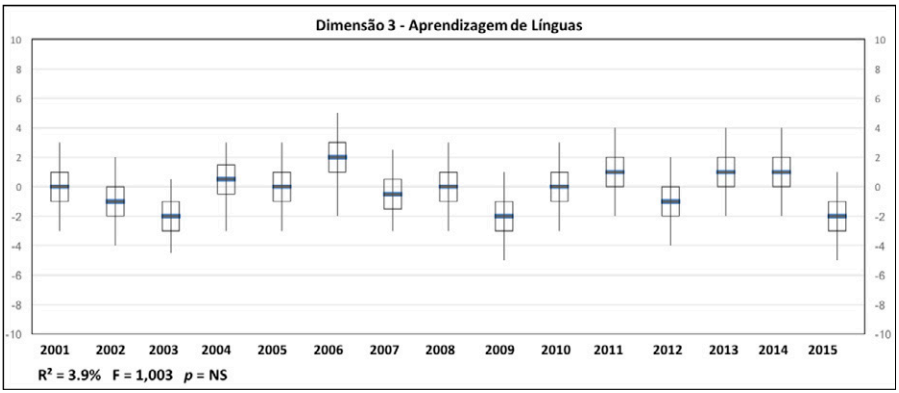


Figura 4 – Escores médios dos anos na Dimensão 3
Fonte: Elaborado pelos autores (2017)

Em concordância, Gardner (1988) destaca que "parece haver um consenso que crenças sobre aprendizagem de línguas, obviamente, são crenças a respeito do que é linguagem, do que é aprendizagem de línguas e sobre aspectos pertinentes à linguagem e à aprendizagem, ou toda tarefa de aprender". Como tais crenças não são estanques e tendem a mudar com o tempo, faz sentido que essa dimensão seja estável e permeie a maior parte das publicações da revista ao longo dos anos.

[4] NS – Não Significativo.

5.4. Dimensão 4: Subáreas da Linguística Aplicada

A Dimensão 4, composta por um polo, inclui léxico que indica um tema predominante: *Subáreas da Linguística Aplicada*. As palavras com maior destaque são *temático* e *comunidade*, respectivamente com 0,893 e 0,682 de peso. Ao analisarmos os textos com maiores pesos nesta dimensão, percebemos que eles tratam, em sua maioria, de assuntos referentes às *Subáreas da Linguística Aplicada*. Um exemplo dessa dimensão é o artigo de Archanjo (2011), que descreve a importância das *Subáreas da Linguística Aplicada* no Congresso Brasileiro de Linguística Aplicada (CBLA), como mostra o exemplo a seguir:

> Exemplo 4:
> A opção pelo CBLA como fonte de empiria – além do fato de ser considerado o congresso máximo da área da LA – se deu pelo entendimento de que esse evento tem sido não somente um fórum privilegiado de divulgação da produção teórico-científica da LA mas, principalmente, por ser um momento em que, a cada três anos, a **comunidade** científica da área se questiona sobre sua prática e sobre os rumos que segue ou pretende seguir. (ARCHANJO, 2011)

Essa dimensão teve escores maiores em 2011, conforme mostra a Figura 5. A ANOVA indica que apenas 4,6% da variação encontrada nesta dimensão poderia ser explicada pela variável temporal, no entanto, o valor de *p* foi de NS para tal variável, indicando que não há diferença estatística significante entre os anos, ou seja, pode-se dizer que as subáreas da LA não variaram com o passar do tempo.

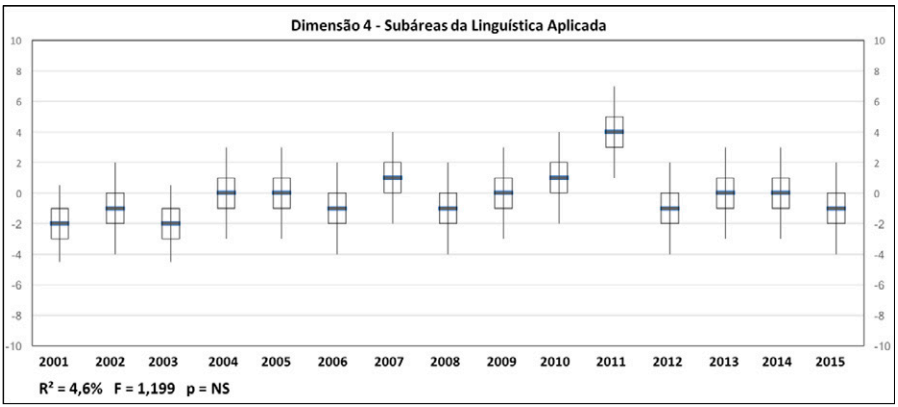


Figura 5 – Escores médios dos anos na Dimensão 4
Fonte: Elaborado pelos autores (2017)

A Figura 5 sugere ainda que a temática *Subáreas da LA* marca a RBLA desde sua fundação de modo praticamente estável. A variação encontrada, na verdade, se deve unicamente ao fluxo normal de temas da revista, de um volume para outro, não sendo forte o suficiente para indicar tendências ou épocas da revista.

Em consonância com Archanjo (2009), podemos dizer que embora as *Subáreas da LA* tenham se ampliado bastante ao longo da história evolutiva desse campo de conhecimento, é ainda visível o fato de que aquelas dedicadas ao ensino-aprendizagem de línguas materna e estrangeira continuam a merecer destaque não apenas pelo número de trabalhos produzidos, mas também pela diversificação dos enfoques que as revestem hoje em dia.

5.5 Dimensão 5: Sala de Aula

A Dimensão 5, composta apenas por um polo positivo, inclui léxico que indica dois temas predominantes, *didática* e *sala de aula*. As palavras com maior destaque são *aula* e *alunos*, respectivamente com pesos 0,693 e 0,638. Ao analisarmos os textos com maiores pesos nesta dimensão, percebemos que os assuntos tratados refletem eventos relacionados à sala de aula. Exemplo dessa dimensão é o artigo de Ribeiro, do 1º número do volume 6 da revista, de 2006:

> Exemplo 5:
> Os saberes adquiridos na disciplina de Didática e Prática de Ensino de língua inglesa são transpostos pela **professora** para as situações de ensino em **sala** de **aula.** (RIBEIRO, 2006)

Essa dimensão teve escores maiores nos anos 2006 e 2008, conforme mostra a Figura 6. A ANOVA indica que apenas 4,5% da variação encontrada nessa dimensão poderia ser explicada pela variável temporal. Entretanto, o valor de *p* não foi estatisticamente significativo, mostrando que os temas envolvendo a sala de aula não variaram sistematicamente ao longo do tempo, ou seja, a sala de aula e os eventos investigados em torno dela continuam os mesmos, apesar de todas as mudanças que ocorreram no mundo. Em outras palavras, a variação encontrada, na verdade, se deve ao fluxo normal de temas da revista de um volume para outro, não sendo forte nem sistemática o suficiente para indicar tendências ou épocas da revista, conforme pode ser observado na Figura 6.

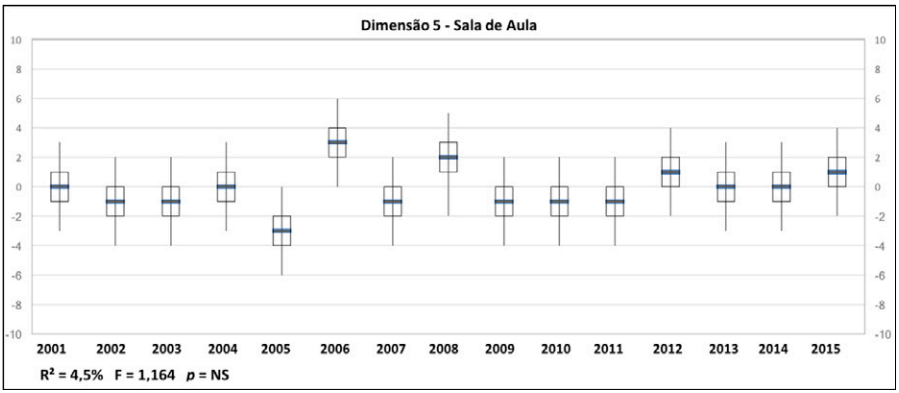


Figura 6 – Escores médios dos anos na Dimensão 5
Fonte: Elaborado pelos autores (2017)

A literatura também mostra que as mudanças dentro da sala de aula não ocorrem com a rapidez necessária para motivar o aluno ao aprendizado. Celani (2016, p. 547-548) levanta alguns pontos a serem questionados e refletidos com relação a esse contexto, tais como: como estimular o aluno e fazer com que ele veja relevância no que é feito em sala de aula?; qual atmosfera o aluno espera encontrar em uma sala de aula?; as concepções do que é aprendizagem e de domínio de classe estão sendo consideradas pelos professores? As respostas a esses questionamentos podem nos levar à reflexão do porquê esse tema não ter variado sistematicamente ao longo dos anos.

5.6 Dimensão 6: Práticas Sociais e Questões Identitárias

A Dimensão 6, composta por um polo negativo, inclui léxico que indica dois temas predominantes: *práticas sociais* e *questões identitárias*. As palavras de maior destaque são *sociais* e *vida*, respectivamente com -0,608 e -0,466 de peso. Ao analisarmos os textos com maiores pesos nessa dimensão, percebemos que os assuntos tratados abordam práticas sociais que enfocam questões identitárias, principalmente dos professores. Exemplo dessa dimensão é o artigo de Ribeiro, *Influências dominantes na construção da prática pedagógica de uma aluna professora de língua inglesa,* presente no volume 6, n. 1, de 2006:

Exemplo 6:
A terceira pergunta pedagógica – Como me tornei assim? – deve guiar o professor em reflexões que o levem a questionar sua ação pedagógica como fruto de forças **sociais** e culturais e o torne consciente das múltiplas forças que o levaram a tornar-se o profissional que é. (RIBEIRO, 2006)

Essa dimensão teve escores maiores nos anos 2001 e 2006, conforme mostra a Figura 7. A ANOVA indica que apenas 11,7% da variação encontrada nessa dimensão é explicada pela variável temporal. No entanto, o valor de *p* foi de 0,00 (<0,05) o que indica que há diferença estatística significativa entre os anos, mostrando que as práticas sociais e questões identitárias são abordadas de maneiras distintas ao longo dos anos, como nos ilustra a Figura 7; esse é um assunto que foi muito comum até 2010, mas não é mais tão frequente nos anos seguintes.

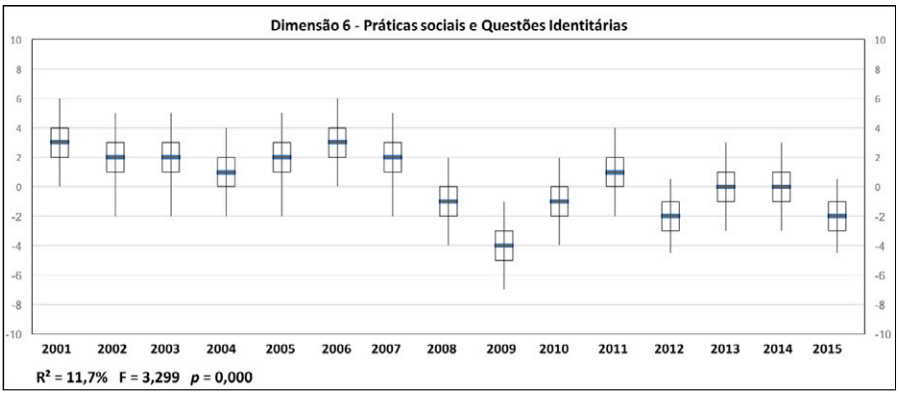


Figura 7 – Escores médios dos anos na Dimensão 6
Fonte: Elaborado pelos autores (2017)

A literatura nos mostra que o tema *questões identitárias* e, dentro deste, a discussão em torno dos conflitos identitários quanto à origem e condição social de alunos e professores esteve muito em voga durante e logo após a introdução da Lei de Diretrizes e Bases da Educação (LDB), Lei 9.394/1996. No entanto, após sua implementação, foi aos poucos perdendo força, o que talvez explique o fato de publicações da RBLA não contemplarem mais esse tema com frequência.

5.7 Dimensão 7: Material de Ensino e Recursos Didáticos

A Dimensão 7, composta por um polo positivo, inclui léxico que indica dois temas predominantes: materiais de ensino e recursos didáticos. As palavras com maior destaque são *didático* e *livro*, respectivamente com pesos 0,880 e 0,857. Ao analisarmos os textos com maiores pesos nesta dimensão, percebemos que os assuntos tratados versam sobre material didático. Um exemplo dessa dimensão é o artigo de Kersch e Guimarães, *A construção de projetos didáticos de leitura e escrita como resultado de uma proposta de formação continuada cooperativa*, do 3º número do volume 12 da revista, de 2012:

> Exemplo 7:
> Assim, a profissionalização de um docente supõe a superação da simples colocação em prática dos materiais e técnicas didáticas disponíveis, passando para um outro patamar, que implica desenvolver capacidades de adaptação e criação de novos dispositivos **didáticos.** (KERSCH; GUIMARÃES, 2012)

Essa dimensão teve escores maiores em 2012, conforme mostra a Figura 8. A ANOVA indica que apenas 4,2% da variação encontrada nesta dimensão poderia ser explicada pela variável temporal, mas o valor de *p* não foi estatisticamente significativo, mostrando que as publicações relativas ao *Material de Ensino e Recursos Didáticos* não se alterou com o passar do tempo, semelhante aos achados que se referem à *Aprendizagem de Línguas* (Dimensão 3).

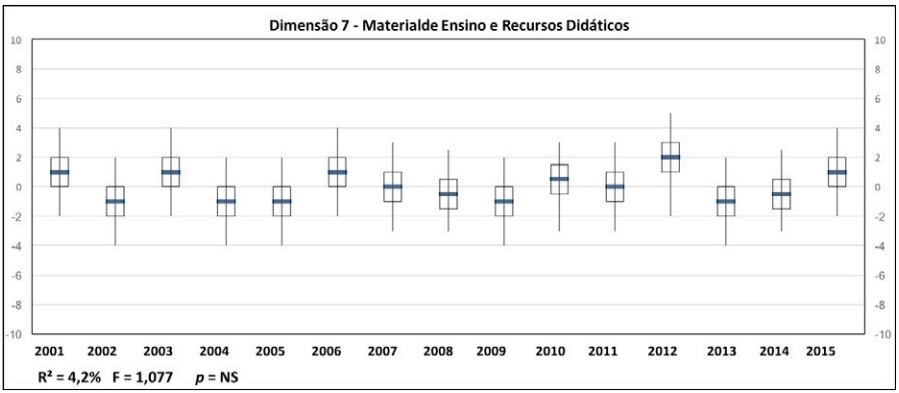


Figura 8 – Escores médios dos anos na Dimensão 7
Fonte: Elaborado pelos autores (2017)

Em outras palavras, a temática *Material de Ensino e Recursos Didáticos* marca a RBLA desde sua fundação de modo praticamente estável. A variação encontrada, na verdade, se deve unicamente ao fluxo normal de temas da revista, de um volume para outro, não sendo forte o suficiente para indicar tendências ou épocas da revista.

Ramos e Freire (2004), ao falarem do ensino por computador, salientam as mudanças que essa e qualquer outra ferramenta traz para o *design* de material didático que precisam ser repensados constantemente. Esse olhar das pesquisadoras vem ao encontro dos nossos achados, indicando que esse tema é comum em todas as publicações da revista.

5.8 Dimensão 8: Libras e Língua Portuguesa

A Dimensão 8, composta por um polo positivo, inclui léxico que indica dois temas predominantes: *libras* e *língua portuguesa*. As palavras com maior destaque são *surdos* e *libras,* respectivamente com 0,830 e 0,829 de peso. Ao analisarmos os textos com maiores pesos nesta dimensão, percebemos que os assuntos tratados giram em torno de tipos de alunos surdos nas aulas de libras e de língua portuguesa. Um exemplo dessa dimensão é o artigo de Finau do 4º número do volume 14 da revista, de 2014:

> Exemplo 8:
> A pesquisadora mostra como ocorrem trocas na ordem das palavras nas sentenças escritas em **língua portuguesa**, as quais se aproximam da organização sintática da **libras**: "Gosta ele a casa"; "Eu milho é gosta". (FINAU, 2014)

Essa dimensão teve escores maiores em 2014, conforme mostra a Figura 9. A ANOVA indica que 24,9% da variação encontrada nessa dimensão é explicada pela variável temporal. O valor de *p* foi 0,000 (<0,05), o que indica que houve variação nas publicações em relação a *Libras e Língua Portuguesa* com o passar dos anos.

Em outras palavras, a temática *Libras e Língua Portuguesa* marca as publicações da RBLA com mudanças em sua abordagem com o passar do tempo, mudanças essas que relacionam as duas línguas cada vez mais, o que anteriormente não acontecia, cada língua era estudada separadamente e hoje em dia as duas estão interrelacionadas. Ao observarmos a Figura 9, percebemos que essa temática é claramente um assunto recente na revista, se fazendo presente principalmente nos anos de 2014 e 2015.

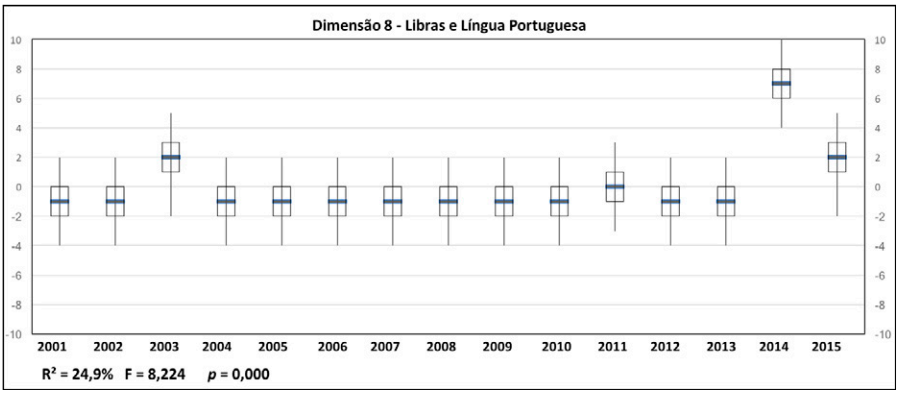


Figura 9 – Escores médios dos anos na Dimensão 8
Fonte: Elaborado pelos autores (2017)

Dias (2012) defende que libras e língua portuguesa são duas línguas que precisam conviver lado a lado e propõe a importância de motivar o aluno surdo no sentido de identificar a língua portuguesa como sua parceira, sem gerar confrontos entre as duas línguas, que era o que acontecia até então. Essa mudança de paradigma juntamente às primeiras políticas que levam em consideração a língua e a comunidade surda no Brasil podem explicar o aumento das publicações e interesse sobre esse tema, o que não era um fato comum até em então.

5.9 Dimensão 9: Novas Tecnologias na Educação

A Dimensão 9, composta por um polo positivo, inclui léxico que indica dois temas predominantes: *tecnologia* e *educação.* As palavras com maior destaque são *tecnologias* e *digital*, respectivamente com 0,645 e 0,593 de peso. Ao analisarmos os textos com maiores pesos nessa dimensão, percebemos que os assuntos tratados abordam os tipos de tecnologia utilizados na educação com base digital. Exemplo dessa dimensão é o artigo de Dias, de 2012: *WebQuests: tecnologias, multiletramentos e a formação do professor de inglês para a era do ciberespaço*:

> Exemplo 9:
> A sociedade mudou sob a influência dessas **tecnologias** e, com isso, novos tipos de **letramento** são necessários, inclusive à comunidade discursiva dos professores de inglês. (DIAS, 2012)

Essa dimensão teve escores maiores nos anos 2012, 2013, 2014, conforme mostra a Figura 10. A ANOVA indica que 16% da variação encontrada nessa dimensão é explicada pela variável temporal. O valor de *p* foi 0,000 (<0,05), o que mostra que as publicações envolvendo *Novas Tecnologias na Educação* têm mudado com o passar do tempo.

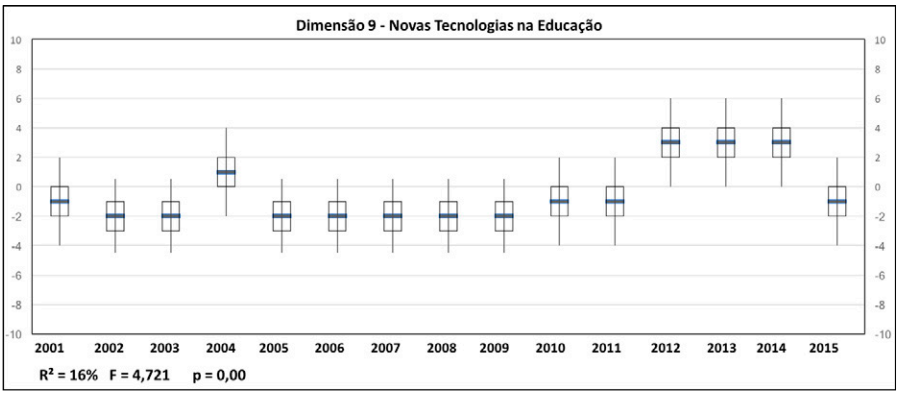


Figura 10 – Escores médios dos anos na Dimensão 9
Fonte: Elaborado pelos autores (2017)

Em outras palavras, a temática *Novas Tecnologias na Educação* marca a RBLA com mudanças em sua abordagem com o passar do tempo. A variação encontrada, na verdade, não se deve unicamente ao fluxo normal de temas da revista, de um volume para outro, sendo forte o suficiente para indicar tendências ou épocas da revista.

A literatura tem abordado a importância das novas tecnologias na educação. Freire (2009, p. 14) relata a necessidade da sintonia entre a educação e novas tecnologias:

> O contexto social em que vivemos é marcado pela rapidez e imediatismo proporcionados por novas modalidades de acesso, armazenamento, recuperação e intercâmbio de informações. Essa caracterização não apenas nos coloca diante de possibilidades únicas de construção e manipulação de conhecimentos, mas, também, origina formas distintas de trabalho, comunicação e interação com o meio, com o outro e com o próprio indivíduo. Parece haver urgência no desenvolvimento de competências e habilidades que respondam mais adequadamente às especificidades desse contexto, à necessidade de um pensar e fazer diferenciados, à carência de instrumentos e metodologias que sejam adequadas a uma percepção inusitada de tempo e espaço e a uma motivação singular para ensinar e aprender.

Moran (2006), por sua vez, relaciona a educação presencial com as tecnologias oriundas da educação à distância, advogando uma mudança no relacionamento cada vez mais intricado tecnologia-educação:

> [...] a educação presencial está incorporando tecnologias, funções, atividades que eram típicas da educação a distância, e a EaD está descobrindo que pode ensinar de forma menos individualista, mantendo um equilíbrio entre a flexibilidade e a interação.

Em virtude do impacto do acesso cada vez mais fácil à internet e aos recursos tecnológicos disponíveis na sociedade contemporânea, essa reflexão sobre o uso de novas tecnologias na educação tem atraído o interesse de estudos recentes publicados na revista.

5.10 Dimensão 10: Abordagem Instrumental

A Dimensão 10, composta por um polo positivo, inclui léxico que indica um tema predominante: *Abordagem Instrumental*. As palavras com maior destaque são *específicos* e *instrumental,* respectivamente com pesos 0,804 e 0,681. Ao analisarmos os textos com maiores pesos nessa dimensão, percebemos que os assuntos tratados debatem tipos de temas específicos, como a abordagem instrumental. Um exemplo dessa dimensão é o artigo de Borges do 4º número do volume 11 da revista, de 2011:

Exemplo 10:
Nesse sentido, tanto a dimensão de propósitos **específicos** (abordagem **instrumental**) quanto à dimensão de propósitos gerais (abordagem comunicativa e abordagem comunicacional) comporiam o MC de ensino de línguas ou o que provavelmente Widdowson pontuou em seu depoimento como abordagem geral ou mesmo desenvolvimento. (BORGES, 2011)

Essa dimensão teve escores maiores em 2011, conforme mostra a Figura 11. A ANOVA indica que apenas 9,4% da variação nessa dimensão pode ser explicada pela variável temporal. O valor de p foi de 0,002 (<0,05), o que indica que os temas relacionados à *Abordagem Instrumental* mudaram com o passar do tempo.

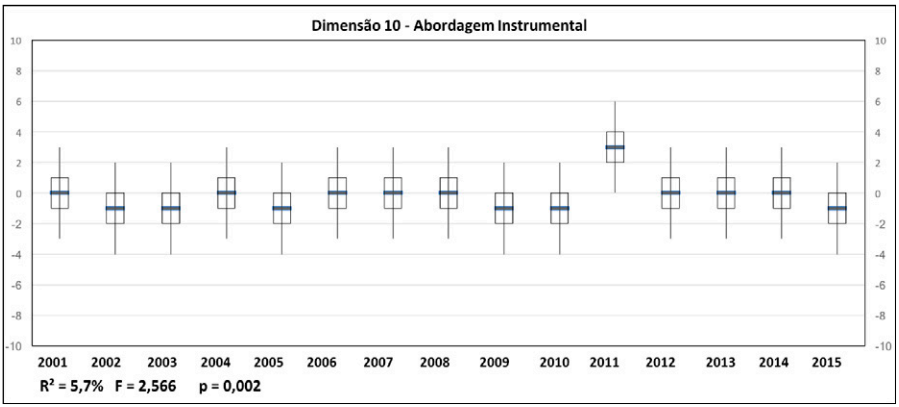


Figura 11 – Escores médios dos anos na Dimensão 10
Fonte: Elaborado pelos autores (2017)

A literatura nos mostra que o Projeto Instrumental no Brasil, que teve início com os professores Maria Antonieta Alba Celani e John Holmes, entre outros, foi e é uma abordagem de rupturas e desbravamentos no ensino-aprendizagem de línguas e, exatamente por esse caráter, necessita se reinventar constantemente, mostrando tendências em determinadas épocas ao longo do tempo, como relata Freire (2009, p. 12-13) no prefácio do livro *A abordagem instrumental no Brasil*:

> Olhando para o passado, para uma história de mais de 25 anos, constatamos que o Projeto se desenvolveu e se mantém vivo, a partir de professores e pesquisadores que se dedicaram à investigação de necessidades, expectativas e lacunas de aprendizagem, e à caracterização de diversos contextos, perfis profissionais, gêneros textuais e atividades de trabalho, a fim de planejar cursos direcionados às especificidades identificadas.
>
> [...] A Abordagem Instrumental, portanto, encontrou artífices dedicados e fiéis nos docentes acima: profissionais que, convencidos do potencial da abordagem, vivenciaram na prática e com antecedência, o que seria, anos depois, reconhecido, teorizado e nomeado.
>
> Olhando para o futuro, sem deixar de lado as experiências do passado, vários são os caminhos que se descortinam à nossa frente, para o planejamento e operacionalização de cursos; definição de conteúdos e procedimentos pedagógicos; preparação, adaptação e avaliação de material impresso e digital; formação de professores em/para ambientes presenciais e a distância. Fica a certeza de que o nosso olhar precisa ser, ao mesmo tempo, prospectivo e retrospectivo, para que não percamos a oportunidade de continuar buscando novos caminhos, sem esquecer da origem, dos fundamentos, das conquistas prévias.

A variação encontrada no que se diz respeito à *Abordagem Instrumental*, na verdade, não se deve unicamente ao fluxo normal de temas da revista, sendo forte o suficiente para indicar tendências ou épocas da revista. Ao observarmos a Figura 11, percebemos que essa temática é saliente no ano de 2011 devido a um número especial do periódico dedicado ao tema.

6 Discussão dos resultados

A análise das publicações da RBLA identifica alguns diferentes temas da área nos contextos profissionais e educacionais, tais como o ensino/aprendizagem de língua estrangeira. Há a presença marcante de temas envolvendo ensino e formação de professores, como observado em nossa análise, em que 5% da variação de todos os componentes dessa revista se deve a esse tema.

Por meio da abordagem multidimensional, foi possível identificar agrupamentos de textos nas publicações da revista que possuem perfil lexical semelhante, isto é, padrão de variação lexical revelando campos semânticos sobre um mesmo tópico, o que aqui definimos como *dimensões lexicais*.

Houve mudanças, com o passar dos anos, nos temas relacionados ao *Ensino e Formação de Professores* (Dimensão 1), *Texto, Gênero e Discurso* (Dimensão 2), *Práticas sociais e Questões Identitárias* (Dimensão 6), *Libras e Língua Portuguesa* (Dimensão 8), *Novas Tecnologias na Educação* (Dimensão 9) e *Abordagem Instrumental* (Dimensão 10). Por outro lado, o tema *Sala de Aula* (Dimensão 5) não mostrou variação significante entre os anos de publicação da revista analisados, mostrando que o que se publica em relação a esse tema não se alterou com o passar do tempo. O tema *Sala de Aula* merece ser mais bem explorado para que possamos observar as mudanças que de fato ocorreram ou não na sala de aula ao longo dos anos.

Aprendizagem de Línguas (Dimensão 3) é um tema frequente nas publicações da RBLA, de modo praticamente estável, ou seja, esta temática parece estar intrinsicamente ligada à Linguística Aplicada, influenciando até mesmo a percepção que o aprendiz tem da língua materna. O mesmo pode ser dito em relação às preocupações pedagógicas e aos alunos, evidenciadas na Dimensão 5 (*Sala de Aula*), também intrinsicamente ligada à LA de modo praticamente estável.

Neste trabalho, das dez dimensões identificadas pela AMDL empregada, apenas duas não abordam explicitamente a questão de ensino-aprendizagem de línguas (Dimensão 2 – *Texto, Gênero e Discurso* – e Dimensão 4 – *Subáreas da Linguística Aplicada*), corroborando os achados de autores, tais como Moita Lopes (1999) e Simpson (2011), que afirmam que, apesar de a Linguística Aplicada ir muito além da sala de aula, esse tema está intimamente relacionado às pesquisas desenvolvidas na área. Vale ressaltar que *Libras e Língua Portuguesa* (Dimensão

8) é um assunto recente nas publicações do periódico (marcadamente presente os anos de 2014 e 2015).

A análise salientou que alguns temas são constantes e presentes em praticamente todos os números da revista. Há temas com maior saliência nas publicações antes de 2010 e com queda nos anos que se seguiram. Em contrapartida, há temas que tiveram maior saliência somente após 2010, o que pode revelar que alguns deles não têm sido abordados com mesma intensidade de interesse que outros nas recentes publicações.

Os resultados apontaram que os temas frequentes nas publicações da RBLA são *Texto, gênero e discurso* (Dimensão 2), *Aprendizagem de Línguas* (Dimensão 3), *Subáreas da Linguística Aplicada* (Dimensão 4), *Sala de Aula* (Dimensão 5), *Material de Ensino e Recursos Didáticos* (Dimensão 7) e *Abordagem Instrumental* (Dimensão 10). Pode-se dizer, portanto, que a Linguística Aplicada no Brasil, representada pela RBLA, uma das publicações mais importantes da área, está intrinsicamente relacionada à aprendizagem de línguas e às questões de gênero e discurso.

Com relação às temáticas mais comumente presentes nas publicações anteriores a 2010, os achados mostram que tanto *Ensino e Formação de Professores* (Dimensão 1) quanto *Práticas Sociais e Questões Identitárias* (Dimensão 6) são assuntos não tão abordados nas publicações mais recentes. Podemos dizer que o primeiro tema não é mais privilegiado por haver, hoje em dia, uma preocupação maior com os alunos e não tanto com a formação docente. O mesmo se dá com as questões identitárias, que enfocavam muito a figura do professor. Na verdade, esses dois temas, aparentemente distantes, estão interligados, e, com a mudança de olhar das pesquisas do professor para o aluno e as novas tecnologias na sala de aula, podemos dizer que as práticas sociais deixaram de ter uma presença marcante nas publicações da revista.

Por outro lado, libras é uma temática presente com maior frequência nas publicações do periódico a partir de 2010, já compreendida como a língua materna do aprendiz e língua portuguesa, como segunda língua. O aumento de enfoque nas pesquisas e publicações mais recentes sobre o tema pode ser justificado pelo reconhecimento da libras como língua oficial no Brasil. Com relação a *Novas Tecnologias na Educação*, um tema também com uma frequência altamente marcada nas temáticas da RBLA após 2010, pode-se dizer que a razão de tal fato poderia estar na inserção da internet e aplicativos informatizados no contexto escolar.

Em suma, partindo-se do pressuposto de que o periódico analisado é representativo da Linguística Aplicada no contexto brasileiro, pode-se dizer que os achados da pesquisa refletem a direção que a LA está tomando no Brasil.

7 Considerações finais

A pesquisa relatada no presente artigo espera ter contribuído para a construção de um panorama temático-histórico da LA entre 2001 e 2015 no Brasil. Por meio da análise do perfil lexical das publicações da RBLA, destacando o léxico mais saliente presente nos textos, revelou-se a existência de dez dimensões que refletem os tópicos de interesse amplamente abordados e também tradicionais nas pesquisas em LA desenvolvidas no Brasil. Esse panorama permite observar a trajetória traçada por teóricos e praticantes da LA brasileira nos primeiros quinze anos de publicação no principal veículo de divulgação de pesquisas da área.

A AMD Lexical empregada neste estudo apontou, por meio das dimensões identificadas, tendências temáticas do periódico que podem fomentar questionamentos em relação aos caminhos e perspectivas da área científica; a dimensão de impacto dos resultados, abordagens e práticas empregadas nos estudos publicados; as expectativas dos pesquisadores-autores atuais e leitores ou aspirantes a pesquisadores da LA no Brasil.

Entretanto, é importante destacar que o índice H (índice que mede a produtividade e impacto do número de artigos com citações maiores ou iguais a esse índice) das publicações do periódico é baixo. Tal fato levanta uma série de questões que precisam ser pensadas e abordadas nas universidades, quais sejam: os artigos não são lidos com muita frequência por pesquisadores e alunos de programas de pós-graduação cujas pesquisas se circunscrevem na LA? Os artigos lidos não são citados com certa regularidade? O baixo índice reflete a qualidade do que é escrito e publicado? Os temas abordados nas publicações são de interesse da comunidade científica da área? Os temas abordados realmente refletem a agenda da LA contemporânea?

A continuidade da pesquisa ora relatada possibilitará a oportunidade de reflexão sobre os questionamentos apontados acima e pelas tendências temáticas que configuram o panorama temático-histórico traçado pelas publicações dos artigos científicos que são comumente publicados no periódico. Pretende-se também redesenhar o *corpus* do estudo, considerando apenas o registro *artigo científico*, bem como a inclusão das publicações referentes aos anos de 2016 e 2017, a fim de observar se os achados inicialmente obtidos se confirmariam.

Referências

ARCHANJO, R. Linguística Aplicada: uma identidade construída nos CBLA. *Revista Brasileira de Linguística Aplicada*, v. 11, n. 3, p. 609-632, 2011. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982011000300002>. Acesso em 15 jul. 2016.

_____. Os caminhos da Pesquisa em Linguística Aplicada nos Congressos Brasileiros de Linguística Aplicada (CBLA): em foco o ensino de línguas. In: ABRALIN, 2009. *Anais*... p. 2366-2374.

ASSOCIAÇÃO BRASILEIRA DE LINGUÍSTICA APLICADA (ALAB). Disponível em: <http://www.alab.org.br>. Acesso em: jun. 2016.

BARRETO, J. P. S. *A redação de vestibular sob uma perspectiva multidimensional*: uma abordagem da Linguística de *Corpus*. Tese (doutorado em Linguística Aplicada e Estudos da Linguagem – LAEL). Pontifícia Universidade Católica de São Paulo, São Paulo, 2016.

BERBER SARDINHA, T. Análise Multidimensional. *D.E.L.T.A.*, n. 16, v. 1, p. 99-127, 2000a.

_____. Linguística de *Corpus*: histórico e problemática. *D.E.L.T.A.*, v. 16, n. 2, 2000b.

_____. *Linguística de* Corpus. São Paulo: Manole, 2004.

_____. 25 years later: Comparing Internet and pre-Internet registers. In: BERBER SARDINHA, T.; VERIANO PINTO, M. (Ed.). *Multi-Dimensional analysis, 25 years on*: a tribute to Douglas Biber. Amsterdam: John Benjamins, 2014, p. 81-105.

_____. *Corpus* Linguistics and History: Lexical dimensions in TESOL Quarterly. In: AMERICAN ASSOCIATION FOR *CORPUS* LINGUISTICS CONFERENCE, Ames, IA, 2016. *Annals...* [s/l]: [s/n], 2016.

_____. A *corpus*-based history of Applied Linguistics. In: WORLD CONGRESS OF APPLIED LINGUISTICS (AILA), 18., 2017, Rio de Janeiro. *Annals...* [s/l]: [s/n], 2017.

BERBER SARDINHA, T.; KAUFMANN, C.; ACUNZO, C. M. A multidimensional analysis of register variation in Brazilian Portuguese. *Corpora*, v. 9, n. 2, p. 239-271, 2014a.

_____. Dimensions of register variation in Brazilian Portuguese. In: BERBER SARDINHA, T.; VERIANO PINTO, M. (Ed.). *Multi-Dimensional analysis, 25 years on*: a tribute to Douglas Biber. Amsterdam: John Benjamins, 2014b, p. 35-80.

BERBER SARDINHA, T.; ACUNZO, C. M.; SÃO BENTO FERREIRA, T. *Dimensions of collocation in Brazilian Portuguese*: Exploring the Brazilian *Corpus* on Sketch Engine. [No prelo].

_____. Metáforas da Economia no Dicionário de Colocações do Português Brasileiro. *Filologia e Linguística Portuguesa*, São Paulo, v. 18, n. 1, 2016.

BERBER SARDINHA, T.; VEIRANO PINTO, M. Dimensions of register variation across American television registers. *International Journal of* Corpus *Linguistics*. [No prelo].

BÉRTOLI-DUTRA, P. Multi-Dimensional Analysis of pop songs. In: BERBER SARDINHA, T.; VERIANO PINTO, M. (Ed.). *Multi-Dimensional Analysis, 25 years on*: A Tribute to Douglas Biber. Amsterdam: John Benjamins, 2014, p. 149-175.

BIBER, D. *Variation across speech and writing*. Cambridge: Cambridge University Press, 1988.

_____. *Variation in English*: Multi-dimensional studies. London: Pearson, 2001.

BORGES, E. F. V. Instrumental e comunicativo no ensino de línguas: mesma abordagem, nomes diferentes? *Revista Brasileira de Linguística Aplicada*, v. 11, n. 4, p. 815-835, 2011 . Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982011000400002&lng=en&nrm=iso>. Acesso em 12 de junho de 2016.

CELANI, M. A. A. Ensino de línguas estrangeiras: ocupação ou profissão? In: LEFFA, V. *O professor de línguas estrangeiras*: construindo a profissão. Pelotas: EDUCAT, 2008, p. 23-44.

_____. Um desafio na Linguística Aplicada contemporânea: a construção de saberes locais. *D.E.L.T.A.*, v. 32 (2), p. 543-555, 2016.

CONCEICÃO, M. P. Experiências de aprendizagem: reflexões sobre o ensino de língua estrangeira no contexto escolar brasileiro. *Revista Brasileira de Linguística Aplicada*, v. 6, n. 2,

p. 185-206, 2006. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982006000200009&lng=en&nrm=iso>. Acesso em 12 de junho de 2016.

DELFINO, M. C. N. *O uso de música para o ensino de inglês como língua estrangeira em um ambiente baseada em* corpus. Dissertação (mestrado em Linguística aplicada e Estudos da Linguagem – LAEL). Pontifícia Universidade Católica de São Paulo, São Paulo, 2016.

DIAS, A. F.A., Língua portuguesa e libras: duas línguas que precisam conviver lado a lado. *Revista Escrita*, n. 15, p. 1-16, 2012.

DUTRA, D. P. Carta da Editora. *Revista Brasileira de Linguística Aplicada*, v. 6, n. 1, p. 07-10, 2006. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982006000100001&lng=en&nrm=iso>. Accesso em 10 de junho de 2018.

FINAU, R. Aquisição de escrita por alunos surdos: a categoria aspectual como um exemplo do processo. *Revista Brasileira de Linguística Aplicada*, v.14, n. 4, p. 935-956, Dec. 2014. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982014000400008&lng=en&nrm=iso>. Acesso em 10 de junho de 2016.

FREIRE, M. M. Prefácio. In: CELANI. M. A. A.; FREIRA, M. M.;, RAMOS, R. C. G. (Org.). *A abordagem instrumental no Brasil*: um projeto, seus percursos e seus desdobramentos. Campinas: Mercado de Letras, 2009.

FONSECA DE ARAÚJO, R. *A linguagem dos reality TV shows norte-americanos*: análise e classificação. Dissertação (mestrado em Linguística aplicada e Estudos da Linguagem – LAEL). Pontifícia Universidade Católica de São Paulo, São Paulo, 2017.

GARDNER, R. C. The socio-educational model of second-language learning: Assumptions, findings, and issues. *Language Learning*, v. 38, n. 1, p. 101-126, 1988.

GÓMEZ, P. C. *Statistical methods in language and linguistic research*. Bristol: Equinox, 2013.

HOEY, M. *Lexical Priming*: a new theory of words and language. London: Routledge, 2006.

KERSCH, D. F.; GUIMARÃES, A. M. M. A construção de projetos didáticos de leitura e escrita como resultado de uma proposta de formação continuada cooperativa. *Revista Brasileira de Linguística Aplicada*, v. 12, n. 3, p. 533-556, Sept. 2012. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982012000300006&lng=en&nrm=iso>. Acesso em 10 de junho de 2016.

MORAN, J. L. *Desafios da internet para o professor*. Campinas, 2006. Disponível em: <http://www.mat.ufrgs.br/~vclotilde/disciplinas/Site%20V%EDdeos/html/textos_pdf/desafios_da_internet_para_o_professor.pdf >. Acesso em: jul. 2016.

MOITA LOPES, L. P. Fotografias da Linguística Aplicada no campo das línguas estrangeiras no Brasil. *D.E.L.T.A.*, São Paulo, v. 15, n. especial, p. 419-435, 1999.

PINHEIRO, P. A. Práticas de produção textual no MSN Messenger: ressignificando a escrita colaborativa. *Revista Brasileira de Linguística Aplicada*, v. 10, n. 1, p. 113-134, 2010. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982010000100007&lng=en&nrm=iso>. Acesso em 12 de junho de 2016.

RAMOS, R. C. G.; FREIRE, M. M. Curso de leitura via rede: da preparação à conscientização. In: COLLINS, H.; FERREIRA, A. O. (Org.). *Relatos de experiência de ensino e aprendizagem de design de material didático online 115 línguas na internet*. Campinas: Mercado de Letras, 2004, p. 27996.

SOBRAL, A. U. As Relações entre texto, discurso e gênero: uma análise ilustrativa. *Revista Intercâmbio*, v. XVII, p. 1-14, 2008.

SIMPSON, J. *The Routledge handbook of Applied Linguistics*. London: Routledge, 2011.

RIBEIRO, S. A. Influências dominantes na construção da prática pedagógica de uma aluna-professora de Língua Inglesa. *Revista Brasileira de Linguística Aplicada*, v. 6, n. 1, p. 81-99, 2006. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982006000100006&lng=en&nrm=iso>. Acesso em 10 de junho de 2016.

VEIRANO PINTO, M. *Dimensions of variation in North American movies*. In: BERBER SARDINHA, T.; VERIANO PINTO, M. (Ed.). *Multi-Dimensional Analysis, 25 years on*: a tribute to Douglas Biber. Amsterdam: John Benjamins, 2014, p. 109-147.

ZUPPARDO, M. C. *Dimensões de variação em manuais aeronáuticos*: um estudo baseado na análise multidimensional. Dissertação (mestrado em Linguística Aplicada e Estudos da Linguagem – LAEL). Pontifícia Universidade Católica de São Paulo, São Paulo, 2014.

Tradução, Lexicografia e Terminologia

Developing a rule-based Brazilian Portuguese-to-Libras machine translation system

Desenvolvendo um sistema de tradução automática de português brasileiro para Libras baseado em regras

Francisco Aulisio dos Santos Paiva
Plínio Almeida Barbosa
Pablo Picasso Feliciano de Faria
José Mario De Martino

Francisco Aulisio dos Santos Paiva – Doutorando na Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas, mestre pela Universidade Estadual de Campinas, bolsista Capes – aulisio.paiva@gmail.com.
Plínio Almeida Barbosa – Professor Associado da Universidade Estadual de Campinas – Instituto de Estudos da Linguagem, doutor pelo Institut de la Communication Parlée e Institut National Polytechnique de Grenoble – pabarbosa.unicampbr@gmail.com.
Pablo Picasso Feliciano de Faria – Professor Doutor da Universidade Estadual de Campinas – Instituto de Estudos da Linguagem, doutor pela Universidade Estadual de Campinas – pablofaria@iel.unicamp.br.
José Mario De Martino – Professor Associado da Universidade Estadual de Campinas – Faculdade de Engenharia Elétrica e de Computação, doutor pela Universidade Estadual de Campinas – martino@fee.unicamp.br.

Abstract: Sign Language Machine Translation is a way that can bring, for deaf people, more education and accessibility to information. This paper presents an ongoing rule-based automatic translation system from Brazilian Portuguese (BP) to Brazilian Sign Language (Libras). Our approach is based on the analysis of a BP-Libras parallel *corpus* composed of the content of a science textbook. We describe a methodology to formalize the morphosyntactic and semantic rules that translate BP text to Libras glosses. Each gloss will be used to control a signing avatar. As a result, we aim to provide a way for deaf children to study by using signed material in Libras, providing more technological resources for students to improve their learning.

Keywords: Machine Translation. Rule-Based. Brazilian Portuguese. Libras.

Resumo: A tradução automática de língua de sinais é uma forma que pode trazer mais educação e acessibilidade à informação para pessoas surdas. Este artigo apresenta um sistema em desenvolvimento de tradução automática de português brasileiro (PB) para a língua de sinais brasileira (Libras) baseado em regras. Nossa abordagem baseia-se na análise de um *corpus* paralelo PB-Libras composto a partir de um livro de ciências. Descrevemos uma metodologia para formalização de regras morfossintáticas e semânticas que convertem um texto em PB para uma representação por glosas. Cada glosa será utilizada para controlar um avatar sinalizador de Libras. O objetivo de nossa abordagem é desenvolver uma tecnologia para auxiliar crianças surdas em seus estudos, ajudando-as a melhorar seu aprendizado.

Palavras-chave: Tradução automática. Baseada em regras. Português brasileiro. Libras.

1 Introduction

Machine Translation (MT) deals with the translation from one natural language to another through computer programs. According to Hutchins (2010), the first ideas for MT started around 1949 with the proposals of cryptography, statistical methods, information theory and the exploration of logic and specific features behind the languages. MT is a field of computational linguistics that can be applied to Sign Language translation.

Sign language is not a universal language, for example, the United States, United Kingdom, Portugal, Spain and Brazil, each one of these countries has its own sign language. It is worth mentioning that while American and British English are most of the time highly intelligible to each other, American Sign Language (ASL) and British Sign Language (BSL) are mutually unintelligible. The same applies to Sign Languages in Portugal and in Brazil.

There are two official languages in Brazil: Brazilian Portuguese (BP) and Brazilian Sign Language (Libras). De Martino et al. (2016a) point out that there are 9.7 million Brazilians with some hearing impairment, from whom approximately 3 million are illiterate in written BP. The authors highlight some of the reasons that lead to illiteracy, such as the lack of skilled teachers working with bilingual education, the lack of specific didactic materials and the sparse

availability of interpreters for translating the written and oral content of school subjects to Libras.

In addition to illiteracy, deaf students experience difficulties such as delay in the development of important competences and exclusion from some school activities, as reported by Lacerda (2006). Gudyanga et al. (2014) highlighted that deafness can affect children's development because they may have trouble communicating with hearing people, resulting in isolation and a feeling of frustration. In this sense, computational approaches to the automatic translation of oral language to Sign Language can help foster inclusive education and access to information.

Like any other Sign Language, Libras is a gestural-visual language that uses hand movements and facial expression to convey meaning. According to Quadros and Karnopp (2004), the Libras signs are composed of five parameters: handshape, location and movement of the hand, palm orientation, and non-manual expressions (facial expressions and body movements). As observed by McCleary and Viotti (2007), one characteristic of Sign Languages is that they do not have a widely accepted writing system, especially because pictographic forms of the systems proposed so far are difficult to understand and use.

An alternative to facilitate the representation of signs is the use of glosses. A gloss transcription is a set of words of a chosen verbal language written in capital letters that represent signs of close meaning. A variety of studies of ASL and Libras have already used this type of system, such as: Klima and Bellugi (1979), Friedman (1979), Ferreira-Brito (1995), Liddell (2003), Finau (2004), Quadros and Karnopp (2004), and Felipe (2007). For instance, to represent the sentence "O menino gosta de futebol" (*The boy likes soccer*), the gloss representation is "MENINO GOSTAR FUTEBOL" (*BOY LIKE SOCCER*). Note that the article "o" and the preposition "de" are excluded, because in Libras they are not signed; verbs are always glossed in their infinitive form, because there are no marks of verbal inflection in Libras. To refer to time, there are signs that indicate either past or future.

Gloss transcription is a representation of Libras. Therefore, it is essential that the sequence of the glosses in the transcription mirrors the order of signing, for an appropriate translation from BP to Libras. The approach adopted in this work uses gloss as an intermediate language in the translation process. In our approach, the glosses will be used as indexes to a database containing motion capture data describing the realization of the signs. The motion captured description indicates the behavior of the pertinent joints of the body during sign production. This description is used to control the movements of a three-dimensional avatar. Figure 1 presents an overview of our approach to the BP-Libras MT problem.

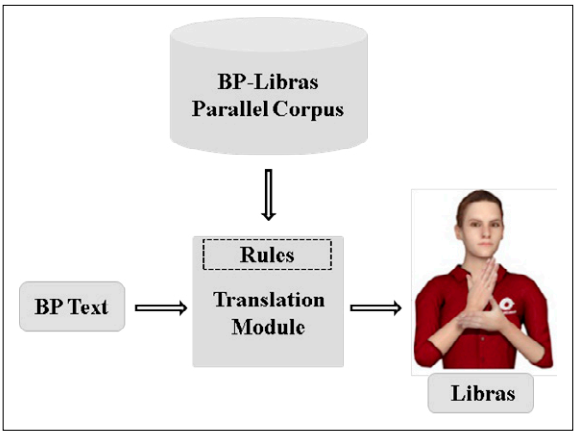


Figure 1 – Overview of our approach to the BP-Libras MT problem

This paper presents the first steps towards a rule-based BP-Libras MT system. The rules are conceived on the basis of the linguistic analysis of a BP-Libras parallel *corpus*. The parallel *corpus* presented in De Martino et al. (2016b) contains the BP-Libras translation of the content of a K-12 science textbook. In the case of oral languages, machine learning approaches have been intensively used for rule inference from *corpora* with huge amounts of sentences. However, this is not yet possible in the case of Sign Languages, because parallel *corpora* are rare, as also highlighted by Porta et al. (2014).

In this context, a feasible alternative for translating Sign Language is to use a knowledge-based approach that identifies and applies linguistic rules to translate sentences in BP to Libras. In this paper, we describe our knowledge-based approach.

This paper is organized as follows. In Section 2, we present works related to Sign Language MT. In Section 3, we discuss basic linguistic concepts and issues of natural language processing. Section 4 describes our approach to tackle the BP-Libras MT problem: the *corpus*, the levels of linguistic analysis, the computational tools and the avatar. Section 5 presents results of the morphosyntactic and semantic analyses, the inferred rules, and the steps of the MT algorithm. In addition, we discuss implications for textbook translation. In Section 6, we present the conclusions and perspectives for our project.

2 Related Works

In this section, we discuss selected works related to Sign Language MT, presenting their approaches, contributions and limitations. In addition, we discuss the importance of using a parallel *corpus*.

Currently, Gupta (2012) indicates that there are two main approaches for MT systems: knowledge-based and data-driven. Knowledge-based MT usually stands for Rule-based Machine Translation (RBMT), whereas data-driven MT embraces Example-based Machine Translation (EBMT) and Statistical Machine Translation (SMT). The RBMT approach denotes a system based on mainstream linguistic levels of description, that is, on rules created from the morphological, syntactic and semantic components of languages. EBMT and SMT, on the other hand, are built upon the analysis of bilingual parallel *corpora* combined with machine learning approaches, as reported by Okpor (2014). According to Koehn (2010), parallel *corpora* are collections of paired texts of two languages. Similarly, in EBMT, the parallel *corpus* contains a source text paired with its translation to the target language. Like EBMT, SMT is also a data-driven approach. However, in the SMT paradigm the translation is modeled as a statistical optimization problem, in which translation patterns are learned from parallel *corpora*.

During the late 20th century, pioneering work on Sign Language MT systems began to emerge. Veale et al. (1994, 1998) is one of the first attempts to implement a rule-based Sign Language translation MT system. The paper presents a methodology to translated English into ASL, Japanese Sign Language (JSL) and Irish Sign Language (ISL). However, their implementation focused only on English to ASL translation. The results of the translation module are presented by a signing avatar. In addition, due to several semantic problems between ASL and English, the implementation was restricted to a small domain, resulting in few animations in ASL at output, as analyzed by Huenerfauth (2003).

Another relevant English-to-ASL MT RBMT system was developed by Zhao et al. (2000). This system translates English text to a gloss representation supplemented with a notation for non-manual signs. The authors consider morphological and syntactic information, and each gloss is associated with a sign in a dictionary. This association supports and facilitates the control of an avatar. As pointed out by Huenerfauth (2003), the system has limitations in representing non-manual expressions, such as head and eye movements.

These two works are rule-based and do not mention the use of bilingual parallel *corpora*. Recognizing that the availability of parallel *corpora* involving SL is still insufficient for data-driven approaches, Porta et al. (2014) also developed a rule-based system from Spanish to Spanish Sign Language glosses, based on dependency-theory syntactic analysis. The approach establishes relations of

dependence between the words of a sentence, such as the relation of the subject to the verb.

Usually, the building of a Sign Language parallel *corpus* begins by choosing the domain of application and then recording videos with a deaf informant. After the recording, the utterances are transcribed into glosses representing signs. The gloss transcription is also supplemented with information about non-manual expressions. Annotation tools, such as Elan software (EUDICO Linguistic Annotator) of Sloetjes and Wittenburg (2008), are used to annotate the video with the glosses and the additional information related to non-manual expressions. Porta et al. (2014) point out that this annotation task is time-consuming.

The works by Othman and Jemni (2011), dealing with English-to-ASL MT system, and Mandita and Anwar (2014), considering Indonesian to German Sign Language, are initiatives involving Sign Language SMT. Currently, there is no pure statistical translation system for BP to Libras especially due to the lack of extensive parallel *corpora*.

BP-Libras MT systems, such as Falibras (BRITO et al., 2012) and Vlibras (LIMA, 2015), use translation rules. The rules of both approaches are quite simple and include rules like: suppression of articles and prepositions, convert of inflected verbs to infinitive; removal of the connecting verb "é" *(is)*[1], and the translation treatment for past sentences. In Lima (2015), 69 sentences were translated by two Libras interpreters and used to test the VLibras set of rules.

To further improve the BP-Libras RBMT solutions, we advocate a more in-depth systematic analysis of parallel *corpora* to derive extensive morphosyntactic and semantic rules.

The parallel *corpus* presented in De Martino et al. (2016b) is currently being analyzed to derive translation rules for our BP-Libras system. This BP-Libras parallel *corpus* contains paired BP-Libras sentences of an elementary textbook with approximately 2,000 sentences pairs composed of 3,000 different glosses/signs.

3 Computational Linguistics Overview

In this section, we present basic concepts for the processing of natural language applicable to the elaboration of rules for automatic translation.

3.1 Grammar in Chomsky's Perspective

The grammar (or the grammatical knowledge of a speaker) is defined as a finite set of rules that can generate an infinite set of well-formed sentences,

[1] This generates an error when it comes to yes/no sentences.

as proposed by Chomsky (1957). Chomsky's perspective has support in formal mathematical concepts, such as the definition of a recursive function. Recursion is a process that defines a function in terms of itself. Baker (2001) illustrates this idea with English sentences, as can be seen in these 3 rules:

I) A sentence S is formed by a Noun Phrase (NP) and a Verb Phrase (VP). In the form, S ⇒ NP + VP. A NP consists of a noun and its potential modifiers, such as articles, adjectives, possessive pronouns, and participles;
II) A VP consists of a verb (V) possibly accompanied by a NP or a clause (CP). In the form, VP ⇒ V, V + NP, V + CP, V + NP + CP;
III) A CP consists of a sentence S, possibly accompanied by a complementizer (C). In the form, CP ⇒ S or C + S.

By taking the words below and applying the rules above, Baker (2001) shows that it is possible to produce an infinite number of sentences conveying different ideas, as follows.

Noun phrases: *John, Mary, Joseph*; Verbs: *thinks, likes*; Complementizer: *that*.

1. Mary likes Joseph.
2. John thinks that Mary likes Joseph.
3. Joseph thinks that John thinks that Mary likes Joseph.
4. Mary thinks that Joseph thinks that John thinks that Mary likes Joseph.
5. John thinks that Mary thinks that Joseph thinks that John thinks that Mary likes Joseph.

Each of these sentences is the result of the application of rules I, II, III. Therefore, based on Chomsky's seminal idea, the rules and patterns must be elaborated precisely and explicitly in order to explain the formation of sentences, as indicated by Baker (2001). As seen in section 2, the rule-based approach is extensively used in machine translation systems for written/oral languages. Thus, the rules of translation must be appropriately elaborated to cover the linguistic levels necessary to generate a good quality translation.

3.2 Language Levels

The various aspects studied by linguists can be divided into six areas: phonetics, phonology, morphology, syntax, semantics and pragmatics. Quadros and Karnopp (2004) summarize the study of each of these areas as follows.

- Phonetics studies and describes the sounds of a language. In addition, it analyzes its articulatory, acoustic and perceptual properties.
- Phonology studies sounds in relation to their function and organization in the language. The smallest contrastive unit of sound is called a phoneme.
- Morphology studies the combination of elements, called morphemes, that form the word. For instance, in Brazilian Portuguese, tense, person and number.
- Syntax studies the structure of the sentence (constituency), its relations of subcategorization, agreement, subordination and order.
- Semantics studies the meaning of the words and sentences.
- Pragmatics studies the relation of meaning to a given context. That is, it studies the language in use in a context and the principles of communication.

When working with written texts, phonetic and phonological levels need not be considered if no pronunciation is required. However, such studies are required in speech synthesis, for example. On the other hand, in our BP-Libras MT system, it is important to take into consideration the morphological, lexical, syntactic and semantic levels.

3.3 Levels of Grammatical Analysis for Machine Translation

To process a word or sentence having translation as an end goal, the following analyses should be carried out, as reported by Rosa (2011).

1. Morphological analysis: the words are decomposed into their canonical form with the information of prefixes or suffixes. For instance: "árvores" *(trees)* ⇒ "árvore" + "s" (plural).

2. Lexical analysis: the canonical forms found are supplemented with their grammatical categories. For instance: "árvores" *(trees)* ⇒ "árvore" (noun)

3. Syntactic analysis: the identification of the sentence structure and the function of its words and constituents. For instance, a syntactic analysis of the sentence "As árvores estão bonitas" *(The trees are beautiful)* could be as shown in Figure 2:

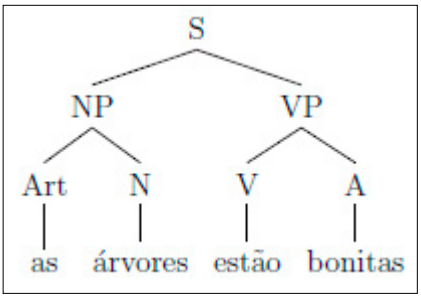


Figure 2 – Syntactic tree

The sentence S, and phrase levels NP and VP can be seen in this syntactic tree. In addition, grammatical categories of words are also shown: article (ART), noun (N), verb (V) and adjective (A). Finally, grammatical functions can be inferred: the NP, as a sister of a VP, stands for the subject of the sentence, while the sister of the V node stands for its complement, in the case above, an adjective (A).

4. Semantic and pragmatic analyses: the meaning of the sentence inferred by taking into account the intention of the statement from the context. For instance, the BP word "saudade" means grateful remembrance of an absent person or something else. Other aspects are evaluated here as well, such as the semantic content of an adjunct phrase, for instance, temporal or locative verbal modifiers.

In the next paragraphs, we show the levels of analysis used so far in the specification of the rules-based MT system.

3.4 Rule-Based Machine Translation

An MT system supporting word-by-word translation hardly achieves satisfactory results for long and complex sentences. For proper translation, it is important to cover all aforementioned levels: morphological, syntactic and semantic. Morphosyntactic and semantic transfers from a source language to any target language are represented by the scheme in Figure 3, known as the Vauquois' triangle.

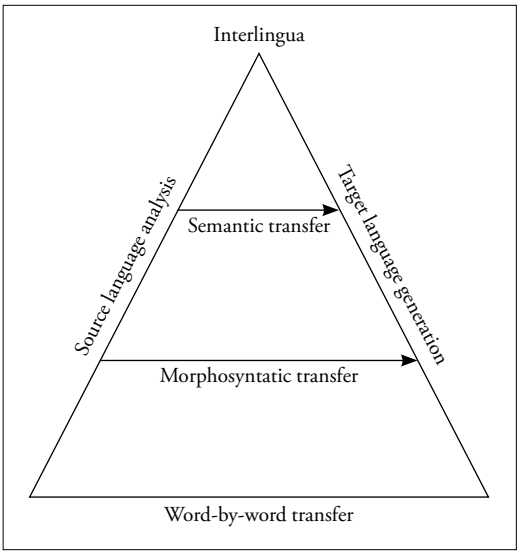


Figure 3 – Vauquois' triangle

The Vauquois' triangle represents different translation approaches considering the depth of the analysis of the source language and the degree of language independence of the representation of meaning in the languages involved. From the bottom to the top, the depth of the analysis and language independence increase. Considering Figure 3, it is possible to classify MT approaches into three different categories: direct, transfer and interlingua.

The direct approach translates the sentences of the source language word-by-word usually using a large bilingual dictionary. The transfer approach applies morphosyntactic and semantic rules to translate one language into another. Finally, the interlingua approach converts the source language into an intermediate representation to generate the target language, as reported by Jurafsky and Martin (2008). According to Dorr et al. (1999), the interlingua can be a symbolic language, which is independent of the source language or target language.

It is important to observe that our intermediate language is a representation of Libras in glosses. Therefore, our intermediate language is dependent of both the source language (BP) and the target language (Libras). Furthermore, since we use a set of rules dependent on the relationship of these two languages, our approach is best characterized as a transfer approach.

Jurafsky and Martin (2008) highlight that transfer architectures can provide good quality translation by applying the idea of contrastive knowledge. This knowledge refers to morphosyntactic and semantic differences between the two languages involved in the translation process. As shown in the Vauquois' triangle,

the transfer process has three phases: analysis of the source language, transfer, and generation of the target language.

The syntactic analysis of the source and target language allows for the identification of syntactic rules for translation. For instance, consider the following example of English to Portuguese translation of a nominal phrase:

beautiful (Adjective) **tree** (Noun) ⇒ **árvore** (Noun) **bonita** (Adjective)

Note that there is an inversion "Adjective Noun" to "Noun Adjective". In this context, the inversion can be as a syntactic rule for the MT process.

Besides syntactic rules, transfer-based systems may also benefit from semantic rules. Jurafsky and Martin (2008) exemplify the need for semantic rules considering the translation of the word "home" from English to German. There are at least 4 possibilities: *nach Hause* (meaning *going home*); *Heim* (meaning *home game*); *Heimat* (meaning *homeland, home country* or *spiritual home*) and *zu Hause* (meaning *being at home*). Ambiguities of this kind can be detected by means of a bilingual dictionary or bilingual parallel *corpora*. They can be used as sources of information for a deeper and systematic analysis that enables the solution of this issue for some cases. This example highlights the importance of using *corpora* in machine translation studies.

4 Methodology

Our written BP-Libras MT system is composed of three main components: a parallel *corpus*, a signing avatar of Libras, and computational resources for analysis and text processing. We describe each of these components in the following sections.

4.1 Bilingual Parallel Corpus

The lack of a bilingual BP-Libras parallel *corpus* encouraged De Martino et al. (2016b) to work on the creation of such resource. In addition, the authors devised the *corpus* to be used as support for a MT system. In order to be useful for education of deaf students, they focused on translating a school textbook. Through the analysis of this *corpus*, we identify morphosyntactic and semantic rules for BP-Libras translation.

The steps to obtain the *corpus* were: selection of the textbook; translation of the textbook from BP to Libras by people fluent in both languages; recording in video and motion capture data of an interpreter signing the translation to Libras;

linguistic annotation of the recorded videos. Below we describe these activities reported by De Martino et al. (2016a, 2016b).

4.2 Selection of the textbook

The choice of a K-12 science textbook is due to its use in classrooms given that deaf children face many difficulties in the process of learning, as mentioned above.

4.2.1 BP-Libras translation by interpreters and motion capture

A team of people fluent in Libras translated each sentence of the textbook. The team is composed of hearing, native BP speakers proficient in Libras and deaf people having Libras as their first language and reading skills in BP. Seeking to ensure a wide-accepted translation, each translation was based on well-known Libras dictionaries, such as the trilingual dictionary for Brazilian Sign Language of Capovilla and Raphael (2008).

After the establishment of the best translation by the team, the translated sentence was recorded in video and transcribed using glosses. During recording, a Motion Capture (Mocap) system was used to record the movements of the body, limb, head and face of interpreter. The Mocap data are used to control the animation of a 3D avatar. See the recording scenario in Figure 4.

Figure 4 – Recording session (Source: DE MARTINO, 2016b, p. 60)

4.2.2 Utterances Annotation

The analysis of the videos was performed in the Elan software of Sloetjes and Wittenburg (2008). This software allows for the annotation of a variety of aspects of the signs used in each sentence including the non-manual expressions, such as the facial expressions. The following items compose the annotation: sentence in Portuguese; its translation into English; gloss transcription; hand configurations; narrative marks, when applicable; other comments on the video.

4.2.3 A Signing Avatar

In our system, we are using the avatar described in De Martino et al. (2016a). Figure 5 shows our avatar.

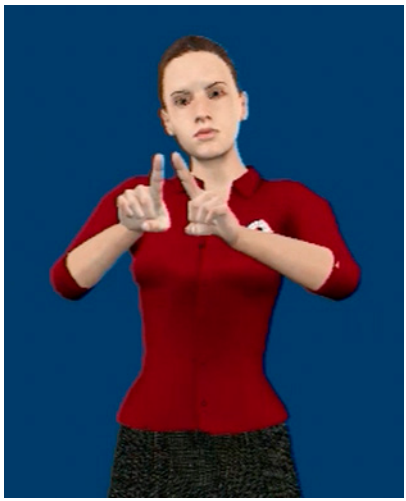


Figure 5 – Signing avatar

The geometric model of the avatar is composed of a textured polygonal mesh that allows for a realistic appearance and a virtual skeleton that represents the bones and joints of the human body. The movements captured in the recording phase are used to control the joints of the virtual skeleton and thus to reproduce the movements recorded during the MoCap session.

The avatar will be ultimately controlled by the intermediate representation of glosses generated by the Translation Module (Figure 1). Each gloss represents a Libras sign, and the automatic translation system converts the BP text into a sequence of glosses following the rules of the Libras grammar.

4.3 Analysis and Text Processing for the MT System

In this section, we present the analyses performed in the *corpus* and the tools used for the task. Morphosyntactic and semantic rules are identified through the analysis of the *corpus* presented.

4.3.1 Portuguese Lexicon

For our morphosyntactic analysis, we used a dictionary of the Unitex system. Unitex is a system for linguistic studies that includes a dictionary and grammars in several languages, such as Spanish, English, French, Greek, and European Portuguese.

From the Unitex system we use the DELAF-BP[2] dictionary developed by Muniz (2004) and members of NILC (Interinstitutional Center for Computational Linguistics USP-São Carlos-SP). The standard followed in the building of this dictionary was designed by the LADL (Laboratoire d'Automatique Documentaire et Linguistique) in France. This standard is known as DELA (Dictionnarie Electronique du LADL) and two types of dictionaries use the standard to identify words in running texts: DELAF (Inflected Words DELA) or DELACF (Inflected Compound Words DELA).

Currently, DELAF-BP has approximately 880,000 BP words. The words of this dictionary are constituted by their lemma, grammatical category and the inflection that corresponds to this form according to the following structure:

word,lemma.part-of-speech+traits:inflection

The lemma is the canonical form of a word. In the case of a verb, the lemma is represented by its infinitive form. For nouns and adjectives, the lemma is represented by the masculine and singular form. Muniz (2004) describes that there are invariable categories such as most adverbs, prepositions, conjunctions, and some determinants. Some examples are given in Table 1:

[2] Available: < http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html >. Last access: September 11, 2017.

Table 1 – Examples of lemmas and grammatical categories

Dictionary examples	Abbreviation
(Eagle) águia,águia.N:fs	Noun (N) Feminine (f) Singular (s)
(Dog) cachorro,cachorro.N:ms	Noun (N) Masculine (m) Singular (s)
(Where) aonde,aonde.PREPXADV	PREPXADV: Contraction of Preposition (PREP) and Adverb (ADV)
da,do.PREPXDET+Art+Def:fs	PREPXDET: Contraction of Preposition (PREP) and Determinant (DET) definite (Def) article (Art) feminine (f) singular (s)
(How) como,como.CONJ	CONJ (Conjunction)
(Of) de,de.PREP	Preposition (PREP)
(Live) vivo,viver.V:P1s	Verb (V); Present of the Indicative (P), First person (1) , singular (s)

We use DELAF-BP together with a lemmatizer to extract the lemmas and their grammatical classes.

4.3.2 Lemmatizer

For our morphosyntactic analysis, we are using the lemmatizer for BP developed by Maziero (2012). The algorithm has the architecture, shown in Figure 6:

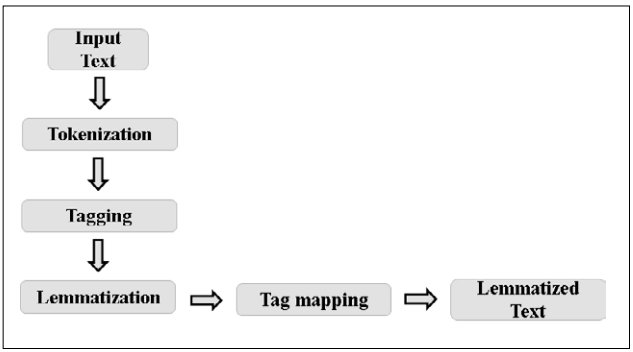


Figure 6 – Steps of the lemmatizer

From a given BP text as input, a tokenization process extracts words and symbols. After that, the morphosyntactic analysis takes place by identifying the grammatical classes of each word. Next, the words are tagged and the mapping of each word to their respective morphological tags is done. Then, the DELAF-BP dictionary is used to complement morphological information for each word. Finally, the algorithm returns the annotated text. Examples of the output are shown in the next section.

By comparing this output with the corresponding Libras translations in the parallel *corpus*, we infer the necessary translation rules. These rules are presented in the next section.

5 Results and Discussion

This section presents the results of the application of the methodology described above. The MT system translating rules are based on a set of 247 sentence pairs of our parallel *corpus*. The sentences are divided into three categories: declarative, Wh-questions and yes/no-questions. They are discussed below in the light of sentence pairs chosen to illustrate the decisions of implementation.

5.1 Steps of the implemented algorithm

Our automatic translator is being implemented in Python 2.7 and it generates the glosses of Libras from a BP written sentence in the following four steps.

GENERAL OUTLINE OF THE TRANSLATION PROCESS
Step 1. Loads the DELAF-BP dictionary;

Step 2. Iterates over user's input sentences (while non-empty);

> **Step 2.1.** Annotate the sentence with lemmas and corresponding grammatical categories in the dictionary;
> **Step 2.2.** Applies pre-processing and translating rules;

Step 3. Delivers a gloss sequence.

These are the details of **Step 2.2.** above.

Step 2.2.1. The pre-processing rules implemented so far are:

PRE-PROCESSING RULES

P1. Identification of the modality of the sentence: declarative; interrogative; imperative.

P2. Exclusion of grammatical categories: articles, preposition + article ("da", "do"), preposition ("de", "para", "a", "entre");

P3. Deletion of verbs: "ser" and "estar" (*to be* – usually not signed in Libras);

P4. Identification of verb to be "é" *(is)*: signed only in yes/no-question;

P5. Plural identification: noun with tag "p" rewritten as "-PL".

Exemplo: "Sombras" *(shadows)* - SOMBRA-PL.

P6. Treatment for NÓS *(we)*: EU-PL (hidden or not) e ELES *(they)*: ELE-PL;

P7. Word Exclusion: "Se" *(if-itself)* - remains only in case of conditional - and "que" *(that)*, "em" *(in)*, "um" *(a-an)* - when they appear in the middle of the sentence.

P8. Rule for past tense: identify the verb tense by adding the gloss JÁ *(already)*

Step 2.2.2. The translating rules implemented so far are:

TRANSLATING RULES

R1. Wh-question:
Interrogative pronouns: "O que" *(what)*; "Por que" *(why)*; "Como" *(how)*; "Qual" *(what)*; "Quanto(s)" *(how many)* are recognized and moved to the end of the sentence.

In some cases they are replaced by a synonym "qual": e.g.: "que" = "qual".

R2. Yes/no-question:
Processing verb to be: recognizes inflection "é" (**P4**) in sentence and moves to the end.

R3. Synonyms:
Synonym lexicon: Some words are replaced by signs of similar meaning, e.g.: "ocorrer" *(occur)* = ACONTECER *(happen);* "e" *(and)* = TAMBÉM *(also)*

R4. Disambiguize function:
Recovers lemma with highest frequency, e.g.: "pode" ⇒ "podar" *(prune)* or "poder" *(can)*

R5. Context analysis:
Adjustments of the glosses to Libras grammar, e.g., exclusion of POR MEIO *(by means of)* – not signed; Join "dia a dia" with hyphens (day-to-day) to DIA-A-DIA.

5.2 Some examples

The rules presented above are illustrated by the following set of ten sentence pairs. For each sentence pair, we present: the text in Portuguese, its word-by-word translation; the English text; and its Libras glosses. The word-by-word translation is presented to help an English-speaking reader understand the order and meaning of each word in Portuguese.

Remember that in the Libras gloss representation the verbs are in the infinitive form. This is already provided by the lemmatizer output, since the lemma of a verb is its infinitive form. In addition, rule P1 works only to identify the modality of the sentence. Below we show the rules applied so far to form the representation by glosses.

Example 1	
BP text	A areia é a maior partícula do solo.
Word-by-word	The-sand-is-the-largest-particle-of.the-soil
English text	Sand is the largest particle of soil.
Libras glosses	AREIA MAIOR PARTÍCULA SOLO

(1) Morphosyntactic analysis: **o** (DET+Art+Def fs) **areia** (N ms|N fs) **ser** (V P3s) **o** (DET+Art+Def fs) **maior** (A fs|A ms) **partícula** (N fs) **do** (PREPXPRO+Dem ms|PREPXDET+Art+Def ms) **solo** (N ms)

After the analysis, the following rules were applied to BP sentences to produce the glosses of Example 1: **P2**, **P3**, which excludes the article and the preposition, besides deleting the verb "to be", because it is not signed in Libras.

Example 2	
BP text	Sabemos que o açúcar se dissolve na água.
Word-by-word	(We)know-that-the-sugar-(itself)-dissolve-in.the-water
English text	We know that sugar dissolves in water.
Libras glosses	EU-PL SABER AÇÚCAR DISSOLVER ÁGUA

(2) Morphosyntactic analysis**: saber** (V P1p) **que** (CONJ) **o** (DET+Art+Def ms) **açúcar** (N ms) **ele** (PRO+Pes R3fs|PRO+Pes R3fp|PRO+Pes R3ms|PRO+Pes R3mp) **dissolver** (V Y2s|V P3s) **no** (PREPXDET+Art+Def fs|PREPXPRO+Dem fs) água (N fs)

In Example 2, rules **P2, P3, P6** and **P7** were applied. Rule **P6** includes the EU-PL gloss to refer to the sign representing the personal pronoun "nós" (*we*). In addition, rule **P7** excludes the words that are not signed, such as "que" (*that*) and "se" (*if*).

Example 3	
BP text	A reportagem refere-se a uma época de verão em nosso país, mas alerta para a chegada do período de inverno.
Word-by-word	he-report-refers-(itself)-to-a-time-of-summer-in-our-country-but-alerts-to-the-arrival-of.the-period-of-winter
English text	The report refers to a time in summer in our country, but alerts to the arrival of winter.
Libras glosses	REPORTAGEM MOSTRAR ÉPOCA VERÃO NOSSO PAÍS MAS ALERTAR CHEGAR PERÍODO INVERNO

(3) Morphosyntactic analysis**: o** (DET+Art+Def fs) **reportagem** (N fs) **referir** (V+PRO P3s) **a** (PREP) **um** (DET+Num Cfs|DET+Art+Ind fs) época (N fs) **de** (PREP) **verão** (N ms) **em** (PREP) **nosso** (PRO+Pos 1ms|N ms) país (N ms) **mas** (CONJ) **alertar** (V Y2s|V P3s) **para** (PREP) **o** (DET+Art+Def fs) **chegada** (N fs) **do** (PREPXPRO+Dem ms|PREPXDET+Art+Def ms) período (N ms) **de** (PREP) inverno (N ms)

In Example 3, rules **P2** and **R3** were applied**.** Rule **R3** was used because the BP word "referir" *(refer)* does not have an exact corresponding sign. Then, the gloss MOSTRAR (show) is used instead. The same happens with the noun "chegada" *(arrival)*, replaced by the infinitive form of the corresponding verb. Note also the importance of analyzing the conjunction MAS *(but)* for the creation of adversative sentences in Libras.

Example 4	
BP text	O que você já sabe sobre o planeta onde vivemos?
Word-by-word	The-what-you-already-know-about-the-planet-where-(we) live
English text	What do you already know about the planet we live on?
Libras glosses	VOCÊ JÁ SABER SOBRE PLANETA ONDE EU-PL VIVER O-QUE?

(4) Morphosyntactic analysis: **o** (PRO+Dem ms|DET+Art+Def ms|PRO+Pes A3ms) **que** (PRO+Int mp|PRO+Int ms|PRO+Int fp|PRO+Int fs) **você** (PRO+Pes

N2fs|PRO+Tra 3fs|PRO+Pes N2ms|PRO+Tra 3ms) já (ADV) **saber** (V Y2s|V P3s) **sobre** (PREP) **o** (DET+Art+Def ms) **planeta** (N ms|N fs) **onde** (PRO+Int mp|PRO+Int ms|PRO+Int fp|PRO+Int fs) **viver** (V Y1p|V J1p|V P1p|V S1p)

In Example 4, rules **P2, P7** and **R1** were applied**.** This example shows the application of rule **R1**, which moves the interrogative pronoun O-QUE (*what*) to the end of the gloss sequence.

It is worth mentioning the use of the JÁ (*already*) gloss, also used to mark the past tense, since verbs are always transcribed in the infinitive. See also Example 8.

Example 5	
BP text	O que está representado nas imagens?
Word-by-word	The-what-is-represented-in.the-pictures?
English text	What is represented in the pictures?
Libras glosses	REPRESENTAR IMAGEM-PL O-QUE

(5) Morphosyntactic analysis 5: **o** (DET+Art+Def ms) **que** (PRO+Int mp|PRO+Int ms|PRO+Int fp|PRO+Int fs) **estar** (V Y2s|V P3s) representado (A ms) **no** (PREPXPRO+Dem fp|PREPXDET+Art+Def fp) **imagem** (N fp)

For this interrogative sentence, we need: **P2**, **P3**, **P5**, **R1**. In this Example 5, the novelty is the application of rule **P5**. This rule recognizes the plural noun by the information (N, fp), then the notation "-PL" is appended to its gloss. Therefore, the word "imagens" (*pictures*) is represented by the gloss IMAGEM-PL. It is very important to recognize the plural form, because plural marking changes signing in Libras. This change usually involves the repetition of the sign or the use of both hands.

Example 6	
BP text	O leite é uma mistura?
Word-by-word	The-milk-is-a-mixture?
English text	Is milk a mixture?
Libras glosses	LEITE MISTURA É?

(6) Morphosyntactic analysis: **o** (DET+Art+Def ms) **leite** (N ms) **ser** (V P3s) **um** (DET+Num Cfs|DET+Art+Ind fs) **mistura** (N fs)

In Example 6, rules **P2**, **P4** and **R2** were applied. Here we have a yes/no-question, therefore, rule **P4** is applied to identify the verbal inflection É (*is*), and rule **R2** moves it to the end to sign its function as an interrogative marker.

Example 7	
BP text	Por que os peixes conseguem respirar embaixo da água?
Word-by-word	Why-the-fish-get-to.breathe-under-of.the-water?
English text	Why fish get to breathe under water?
Libras glosses	PEIXE-PL CONSEGUIR RESPIRAR EMBAIXO ÁGUA PORQUE?

(7) Morphosyntactic analysis: **por** (PREP) **que** (PRO+Int mp|PRO+Int ms|PRO+Int fp|PRO+Int fs) **o** (DET+Art+Def mp) **peixe** (N mp) **conseguir** (V P3p) **respirar** (V U3s|V W1s|V W|V W3s|V U1s) **embaixo** (ADV) **do** (PREPXDET+Art+Def fs|PREPXPRO+Dem fs) água (N fs).

In Example 7, rules **P2, P5**, and **R1** were applied. Here rule **R1** is applied to shift the interrogative pronoun "por que" (*why*) to the end. In addition, pre-processing rules **P2** and **P5** exclude the articles and the prepositions and add plural notation.

Example 8	
BP text	Carlos construiu um cata-vento.
Word-by-word	Carlos-built-a-pinwheel.
English text	Charles built a pinwheel.
Libras glosses	C-A-R-L-O-S JÁ CONSTRUIR CATA-VENTO

(8) Morphosyntactic analysis: **carlos** (N+Pr ms) **construir** (V J3s) **um** (DET+Art+Ind ms|DET+Num Cms) **cata-vento** ()

In Example 8, rules **P2** and **P8** were applied**.** As mentioned in Example 4**,** rule P8 adds the gloss JÁ (*already*) to represent something that has already happened. Note that in the morphosyntactic analysis, the word "cata-vento" (*pinwheel*) is untagged because it was not found in the dictionary, and the lemmatizer returns the same word.

Example 9	
BP text	Por que você acha que isso ocorre?
Word-by-word	Why-you-think-that-this-happens?
English text	Why do you think this happens?
Libras glosses	VOCÊ ACHAR ISSO ACONTECER PORQUE

(9) Morphosyntactic analysis: **por** (PREP) **que** (PRO+Int mp|PRO+Int ms|PRO+Int fp|PRO+Int fs) **você** (PRO+Pes N2fs|PRO+Tra 3fs|PRO+Ind ms|PRO+Ind fs|PRO+Pes N2ms|PRO+Tra 3ms) **achar** (V Y2s|V P3s) **que** (CONJ) **isso** (PRO+Dem ms) **ocorrer** (V Y2s|V P3s)

In Example 9, rules **P2, P7, R1** and **R3** were applied. Rule **R3** retrieves a gloss in a synonym lexicon: the word "ocorrer" (to occur) has been replaced by ACONTECER (*happen*), because it is this gloss that represents the sign with the closest meaning.

Example 10	
BP text	O vento pode soprar em várias direções.
Word-by-word	The-wind-can-blow-in-several-directions.
English text	Wind can blow in several directions.
Libras glosses	VENTO PODER SOPRAR VÁRI@S DIREÇÃO-PL

(10) Morphosyntactic analysis: **o** (DET+Art+Def ms) **vento** (N ms) **poder | podar** (V Y3s|V S3s|V Y2s|V S1s|V P3s) **soprar** (V U3s|V W1s|V W|V W3s|V U1s) **em (**PREP) **vários** (A fp) **direção** (N fp)

Finally, in Example 10, rules **P2**, **P5** and **R4** were applied. Rule **R4** was used to remove the ambiguity of the word "pode", because it may correspond to either the verb "podar" (*to prune*) or the auxiliary "poder" (*can*). This rule uses Google to find out the most frequent word form. Given the clear limitation of this raw strategy, this function needs future improvement to choose the best solution according to the context. For instance, in the current state, the sentence "Quero que você pode essa árvore" *(I want you prune this tree)* is analyzed incorrectly, because "poder" *(can)* is the most frequent form found in the Internet, instead of "podar" *(to prune)*.

All the sentences illustrated here are part of the *corpus* composed by the content of a K-12 science textbook, but the rules are able to cover other sentences with the same structures. Since the meaning of utterances emerges from the interaction between words in a specific sentence context, a deep semantic analysis is required to obtain appropriate translations, which is still lacking in the literature. The analyses shown here contribute to the general study of Libras and to the formalization of BP-Libras automatic translation.

5.3 Integrating the translation rules to the avatar

Our future goal is to integrate our MT system to the avatar according to the ideas presented in De Martino (2016a).

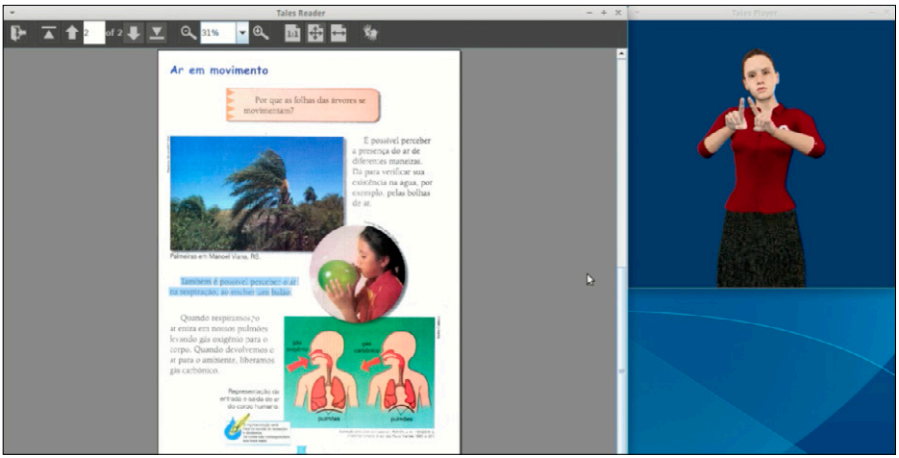


Figure 7 – Example of book translation done by signing avatar

The system will be used to automatically translate the K-12 science textbook from written BP to Libras. When the user selects a particular sentence from the text, a window showing our signing avatar is opened, as showed in Figure 7. Our MT system receives as input a BP text and transfers it to Libras glosses, which is our intermediate representation. In the future, we will develop an algorithm that uses each gloss to concatenate a sequence of signs. The gloss-sign mapping will be interpreted by the avatar animation system and will sign the selected sentence.

6 Conclusion and Perspectives

In this work, we present a rule-based Brazilian Portuguese-to-Libras Machine Translation System. The morphosyntactic and semantic rules so far have produced appropriate translations for 175 out of 245 sentences extracted from our elementary school science textbook, both in declarative and interrogative modalities. The rules can be applied to translate other BP sentences having similar morphosyntatic structures as the already translated sentences. The rules comprise thirteen implemented BP-to-Libras transfer functions ranging from preprocessing of sentences in BP to syntactic inversions and semantic issues.

Future work will integrate the system with an avatar, aiming at the development of a tool that will help deaf people in the learning process. In addition, a system assessment protocol will be developed. This step is of great importance, since translated sentences need to be intelligible to the target user. The objective is to provide more accessibility and inclusive education to deaf people.

Acknowledgments

This work was supported by Capes/ SDH/ MCTI No. 59/2014 – Proc. # 88887.091672 / 2014-01.

References

BHATTACHARYYA, P. *Machine translation*. Boca Raton: CRC Press, Taylor & Francis Group, 2015.

BRITO, P. H. S. et al. FALIBRAS: Uma Ferramenta Flexível para Promover Acessibilidade de Pessoas Surdas. *TISE: Nuevas Ideas em Informatica Educativa*, 8, 2012.

CAPOVILLA, F. C.; RAPHAEL, W. D. *Dicionário enciclopédico ilustrado trilíngue da Língua de Sinais Brasileira*. 3. ed. São Paulo: EDUSP, 2008.

CHOMSKY, N. *Syntactic structures*. The Hague/Paris: Mouton, 1957.

DE MARTINO, J. M. et al. Signing avatars: making education more inclusive. *Universal Access in the Information Society*, Springer-Verlag, Berlin, v. 1, p. 1-16, nov. 2016a.

DE MARTINO, J. M. et al. Building a Brazilian Portuguese – Brazilian Sign Language parallel *Corpus* using motion capture data. In: THE 12TH INTERNATIONAL CONFERENCE ON THE COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 2016, Tomar. *Proceedings workshop on* corpora *and tools for processing* corpora *workshop*, p. 56-63, 2016b.

DORR, B. J.; JORDAN, P. W.; BENOIT, J. W. A survey of current paradigms in machine translation. *Advances in computers*, Burlington, v. 49, p. 1-68, 1999.

FELIPE, T. A. *Libras em contexto*: curso básico. 8. ed. Rio de Janeiro: WalPrint Gráfica e Editora, 2007.

FERREIRA-BRITO, L.; LANGEVIN, R. Sistema Ferreira Brito-Langevin de transcrição de sinais. In: FERREIRA-BRITO, L. *Por uma gramática de Língua de Sinais.* Rio de Janeiro: Tempo Brasileiro, 1995.

FINAU, R. A. Os sinais de tempo e aspecto na Libras. 2004. 238 f. Tese (Doutorado em Letras) – Setor de Ciências Humanas, Letras e Artes, Universidade Federal do Paraná, Curitiba, 2004.

FRIEDMAN, L. *Phonology of a soundless language: phonological structure of the American Sign Language*. 1976. 191f. Dissertation (Doctoral of Philosophy in Linguistics). University of California, Berkeley, 1976.

GUDYANGA, E.; WADESANGO, N.; HOVE, E.; GUDYANGA, A. Challenges faced by students with hearing impairment in Bulawayo urban regular schools. *Mediterranean Journal of Social Sciences*, Rome, v. 5, n. 9, p. 445-451, may 2014.

GUPTA, S. *A survey of data driven machine translation*. 2012. 60f. Dissertation. Department of Computer Science and Engineering, Indian Institute of Techonology, Bombay, 2012.

HUENERFAUTH, M. P. *American Sign Language natural language generation and machine translation systems*. Technical Report, Computer and Information Sciences, University of Pennsylvania, 2003. Available: <http://huenerfauth.ist.rit.edu/pubs/huenerfauth-2003-ms-cis-03-32-asl-nlg-mt-survey.pdf>. Last access: October 05, 2017.

HUTCHINS, J. Machine translation: a concise history. In: WAI, Chan Sin. (Ed.). Special issue: the teaching of computer-aided translation. *Journal of Translation Studies*, Hong Kong, v. 13, p. 29-70, 2010.

JURAFSKY, D.; MARTIN, J. H. *Speech and language processing*: an introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed. New Jersey: Prentice Hall, 2008.

KLIMA, E.; BELLUGI, U. *The signs of language*. Cambridge: Harvard University Press, 1979.

KOEHN, P. *Statistical machine translation*. Cambridge University Press, 2010.

LACERDA, C. B. F. A inclusão escolar de alunos surdos: o que dizem alunos, professores e intérpretes sobre esta experiência. *Cadernos Cedes*, Campinas, v. 26, n. 69, p. 163-184, may/aug. 2006.

LIDDELL, S. K. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press, 2003.

LIMA, M. A. C. B. *Tradução automática com adequação sintático-semântica para LIBRAS*. 2015. 101f. Dissertação (mestrado em Informática). Universidade Federal da Paraíba, João Pessoa, 2015.

MANDITA, F.; ANWAR, T. Statistical machine translation for Indonesian-German sign language. *Journal of Theoretical & Applied Information Technology*, Islamabad, v. 66, n. 1, Aug. 2014.

MAZIERO, E. G. *Lemmatizer for Portuguese*. 2012. Available: <http://conteudo.icmc.usp.br/pessoas/taspardo/LematizadorV2a.rar>. Last access : March 3, 2017.

MCCLEARY, L.; VIOTTI, E. Transcrição de dados de uma língua sinalizada: um estudo piloto da transcrição de narrativas na Língua de Sinais Brasileira (LSB). In: LIMA-SALLES, H. M. M. (Org.). *Bilinguismo dos surdos*: questões linguísticas e educacionais. Goiânia: Canone Editorial, 2007, p. 73-96.

MUNIZ, M. C. M. *A construção de recursos linguístico-computacionais para o português do Brasil*: o projeto Unitex-PB. 2004. 92f. Dissertação (mestrado em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2004.

OKPOR, M. D. Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues*, Mahebourg, v. 11, n. 5, p. 159, 2014.

OTHMAN, A.; JEMNI, M. Statistical sign language machine translation: from English written text to American Sign Language gloss. *International Journal of Computer Science Issues*, Mahebourg, v. 8, n. 3, p. 65-73, 2011.

PORTA, J.; LÓPEZ-COLINO, F.; TEJEDOR, J.; COLÁS, J. A rule-based translation from written Spanish to Spanish Sign Language glosses. *Journal Computer Speech and Language*, v. 28, n. 3, p. 788-811, 2014.

QUADROS, R. M.; KARNOPP, L. B. *Língua de Sinais Brasileira*: estudos lingüísticos. Porto Alegre: Artmed Editora, 2004.

ROSA, J. L. G. *Fundamentos de Inteligência Artificial*. Rio de Janeiro: LTC, 2011.

SLOETJES, H.; WITTENBURG, P. Annotation by category – ELAN and ISO DCR. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 6., 2008, Marrakech. *Proceedings...* Marrakech: ELRA, 2008.

VEALE, T.; CONWAY, A. Cross modal comprehension in ZARDOZ: an English to sign-language translation system. In: INTERNATIONAL WORKSHOP ON NATURAL LANGUAGE GENERATION, 7. *Proceedings...* Kennebunkport: Association for Computational Linguistics, 1994, p. 249-252.

VEALE, T.; CONWAY, A.; BRÓNA, C. The challenges of cross-modal translation: English-to-Sign-Language translation in the ZARDOZ system. *Machine Translation*, Berlin, v. 13, p. 81-106, mar. 1998.

ZHAO, L.; KIPPER, K.; SCHULER, W.; VOGLER, C.; PALMER, M. A machine translation system from English to American Sign Language. In: WHITE, J. S. (Ed.). *Envisioning machine translation in the information future*. *Lecture notes in Computer Science*, v. 1934. Berlin: Springer, 2000.

Proposta de um vocabulário bilíngue de festas populares brasileiras baseada em um estudo de *corpus*

A bilingual dictionary of Brazilian festivals: a model based on a *corpus* study

Giovana Martins de Castro Marqueze

Resumo: O presente trabalho expõe o primeiro resultado de uma pesquisa em andamento que propõe uma obra terminográfica bilíngue na direção português-inglês sobre festas populares brasileiras, tendo como público-alvo profissionais que produzem textos turísticos em inglês sobre o Brasil, sejam eles traduções ou produções nesse idioma. A pesquisa tem a Linguística de *Corpus* como aporte metodológico e faz uso de um *corpus* paralelo composto por textos turísticos sobre o Brasil escritos originalmente em português e suas respectivas traduções para o inglês. Utiliza-se também um *corpus* comparável composto por um *subcorpus* com textos turísticos sobre o Brasil escritos originalmente em português, e outro contendo textos turísticos escritos originalmente em inglês.

Palavras-chave: Textos turísticos. *Corpora*. Festas populares brasileiras. Obra terminográfica.

Giovana Martins de Castro Marqueze – Mestranda do Programa de Pós-Graduação em Estudos da Tradução do Departamento de Letras Modernas da Universidade de São Paulo (USP), licenciada em Letras pela Universidade Estadual de Londrina (UEL) – giovana.marchese@hotmail.com.

Abstract: This paper shows the first result of an ongoing research whose objective is to propose a bilingual reference work (Portuguese / English) of Brazilian festivals, which has as its target audience professionals who produce tourist texts in English about Brazil, be it translations or texts written originally in this language. The research has *Corpus* Linguistics as its methodology and uses a parallel *corpus* compounded by tourist texts about Brazil written originally in Portuguese and their translations into English; and a comparable *corpus* compounded by a *subcorpus* with tourist texts about Brazil written originally in Portuguese and by a *subcorpus* with tourist texts written originally in English.

Keywords: Tourist Texts. *Corpora*. Brazilian Festivals. Reference Work.

1 Introdução

O turismo tem crescido consideravelmente nas últimas décadas como um fenômeno tanto econômico como social, tendo uma importância cada vez maior na economia de diversos países. No Brasil, o setor representa atualmente 3,6% do PIB (Produto Interno Bruto) e emprega direta e indiretamente mais de 10 milhões de pessoas[1]. Impulsionado pelos investimentos realizados no país para o ciclo de megaeventos dos últimos anos, como a Copa das Confederações, a Copa do Mundo e os Jogos Olímpicos, o Brasil ocupa hoje a 27ª colocação no Ranking de Competitividade de Viagens e Turismo do Fórum Econômico Mundial, estudo que compara 136 países analisando 14 dimensões do turismo[2]. Só no ano passado, o Brasil atingiu um número recorde de 6,6 milhões de turistas estrangeiros, registrando um aumento de 4,8% na entrada de turistas internacionais em relação ao ano anterior[3].

Para que todo esse fluxo de turistas tenha acesso a informações sobre os muitos serviços oferecidos pelo setor turístico – hospedagens, restaurantes, passeios, eventos populares e instalações – textos turísticos em forma de panfletos, brochuras, revistas de bordo, pôsteres, *websites*, guias de viagem, dentre outros, são produzidos em grandes quantidades e em diversos idiomas. Segundo Merkaj (2013), esses materiais não devem ser apenas atraentes no que se refere a fotos das atividades e dos lugares divulgados, mas também é importante que sejam apresentados em uma linguagem inteligível ao público que os consumirá. Afinal, é muitas vezes por meio de um texto turístico que o viajante terá suas primeiras impressões sobre o lugar que pretende visitar, podendo, depois, sentir mais ou

[1] EMBRATUR. *Presidente da Embratur participa de debate com empresariado do Lide.* Embratur. 2015. Disponível em: <http://www.embratur.gov.br/>. Acesso em: 24 set. 2017.
[2] OLIVEIRA, Mariana. *Brasil avança no ranking de Competitividade em turismo do Fórum Econômico Mundial.* Ministério do Turismo. 2017. Disponível em: <http://www.turismo.gov.br>. Acesso em: 24 set. 2017.
[3] BRITO, Débora. *Brasil tem recorde de 6,6 milhões de turistas estrangeiros em 2016.* Agência Brasil. 2017. Disponível em: <http://agenciabrasil.ebc.com.br/>. Acesso em: 24 set. 2017.

menos interesse por ele. Não raro, esse primeiro contato com a cultura estrangeira se dá através da tradução de um texto turístico (KELLY, 1997; MERKAJ, 2013), cabendo ao tradutor a tarefa de intermediar a cultura de partida e a de chegada. O farto conteúdo cultural dos textos turísticos representa, segundo Merkaj (2013), um desafio especial para os tradutores, que devem apresentar aos leitores palavras que denotam objetos e conceitos de determinada cultura que não existem da mesma forma em outra, os chamados *realia*.

Considerando a complexidade de se expor aspectos de uma cultura a alguém que não tenha conhecimento sobre ela e levando em conta a já comentada importância da indústria turística para a economia brasileira, estamos conduzindo uma pesquisa que propõe a elaboração de um vocabulário bilíngue[4] (português / inglês) focado nos termos relativos a festas populares brasileiras e que tem como público-alvo profissionais que produzam textos turísticos em inglês sobre o Brasil, sejam eles traduções ou produções escritas originalmente nesse idioma. A proposta tem a Linguística de *Corpus* como aporte metodológico e está sendo desenvolvida a partir do estudo de um *corpus* comparável e de um *corpus* paralelo bilíngues, ambos compostos por textos turísticos sobre o Brasil.

Este trabalho apresenta os detalhes sobre a metodologia empregada na exploração dos *corpora* e na elaboração da obra terminográfica e traz o primeiro exemplo de verbete concebido a partir da análise do termo "Festa Junina". Antes, na próxima seção, ofereceremos uma visão geral dos textos turísticos.

2 Os textos turísticos

A implantação do turismo como atividade profissional favoreceu o surgimento de gêneros textuais como os textos turísticos. Calvi (2010) faz um estudo sistemático desses textos, cobrindo a grande variedade de gêneros textuais produzidos pela indústria do turismo, e propõe uma classificação hierárquica, separando-os em famílias de gêneros, macrogêneros, gêneros e subgêneros.

Como os detalhes do estudo da autora fogem ao escopo deste trabalho, nos concentramos apenas nos três blocos principais nos quais Calvi (2010) agrupa os diferentes gêneros textuais dentro do setor turístico, a saber:

1) Textos que trazem uma reflexão teórica sobre o fenômeno do turismo e suas características. De acordo com a autora, é nesses textos que surgem os neologismos próprios da indústria turística, como o "turismo de massa" e "turismo rural".

[4] Boutin-Quesnel (1985 apud FAULSTICH, 1995) define os vocabulários como sendo obras terminográficas que elencam termos de um domínio e descrevem os conceitos representados por esses termos por meio de definições e ilustrações. Como a obra terminográfica proposta deverá trazer tanto definições quanto ilustrações, achamos por bem chamá-la de "vocabulário".

2) Textos ligados à gestão do turismo, como passagens aéreas, confirmações de reservas em hotéis, recibos, dentre outros.

3) Textos que descrevem e promovem atrações e destinos turísticos, como os guias e revistas de turismo, brochuras de agências de viagens, dentre outros.

Desses três blocos, é o terceiro que nos interessa e do qual tiramos o material que compõe os *corpora* dessa pesquisa. Portanto, para este estudo, consideramos textos turísticos como sendo qualquer texto publicado por uma organização pública ou privada com o intuito de fornecer informação a um possível visitante e anunciar um destino como uma cidade, um hotel, um restaurante etc., encorajando o público a visitá-lo (KELLY, 1997).

2.1 A linguagem do texto turístico

A atividade do turismo envolve contato direto entre culturas e tudo aquilo que faz parte delas, como o folclore, danças, festas, gastronomia, regras, dentre outros (MUÑOZ, 2012). Tendo isso em vista, Argoni (2012) afirma que a linguagem empregada nas produções textuais do setor trata por si só de uma forma de mediação cultural, uma vez que "traduz" valores de uma cultura ao promover identidades de comunidades específicas, constituindo um caso interessante de comunicação intercultural.

Autores como Kelly (1997), Calvi (2000), Muñoz (2012) e Lucas (2012) afirmam que a linguagem do turismo apresenta características lexicais, sintáticas e funções específicas, como detalhadas abaixo:

1) Características lexicais:
- Uso de adjetivos positivos para conferir beleza ao texto;
- Uso de superlativos;
- Presença de *realia*;
- Uso de termos de vários âmbitos de especialidade, como os da arquitetura e da geografia;

2) Características sintáticas:
- Uso de verbos no imperativo para estimular o leitor a visitar o local divulgado;
- Uso do presente do indicativo para passar uma sensação duradoura de viagem;

3) Função
- Função informativa, uma vez que informa sobre as atrações e locais divulgados;

- Função apelativa, já que apela diretamente ao destinatário a fim de convencê-lo a visitar os locais promovidos.

Além das características linguísticas elencadas acima, não podemos deixar de acrescentar aqui os elementos não-verbais desses textos, como mapas, fotos, propagandas e plantas de edifícios. De acordo com Muñoz (2012), eles contribuem para que os textos turísticos cumpram a sua função apelativa, pois funcionam como estímulos que ajudam o turista a decidir aonde ir, o que comprar e onde se hospedar.

2.2 A tradução de textos turísticos

Como dito anteriormente, as traduções de textos turísticos trazem dificuldades especiais, principalmente no que tange aos *realia*. Segundo Lucas (2012), eles representam um verdadeiro desafio para os tradutores, que, além de compreenderem a carga cultural desses termos em seu próprio idioma, deverão eleger, dentre as diferentes soluções de tradução encontradas, a mais adequada e pertinente ao trabalho sendo realizado. Kelly (1997) explica que esse abundante conteúdo cultural dos textos turísticos exige dos tradutores a habilidade de conduzir uma pesquisa minuciosa sobre o termo a ser traduzido para que as informações possam ser passadas da maneira mais clara possível, garantindo a comunicação entre a cultura local e a estrangeira.

Além da questão cultural, Kelly (1997) ainda apresenta uma síntese de outras dificuldades vivenciadas pelos tradutores desses textos, dentre as quais temos a má qualidade do texto-fonte, a qual, segundo a autora, é uma situação comum em textos turísticos, e as limitações impostas pelas imagens e pelas edições bilíngues que trazem o texto em determinado idioma na coluna à esquerda, e em outro idioma, à direita. Essas edições exigem que textos em línguas diferentes tenham o mesmo tamanho, o que restringe as possibilidades de tradução. Ainda segundo Kelly (1997), o fato de as traduções de textos turísticos serem geralmente realizadas para a língua estrangeira do tradutor representa um desafio especial.

Merkaj (2013) aponta a importância de se levar em conta a função desses textos na cultura de chegada e elenca cinco fatores linguísticos e culturais que podem influenciar suas traduções, a saber 1) os sentidos figurados, 2) a diferença em como as culturas entendem certos conceitos, 3) metáforas e expressões, 4) religiões e mitos e 5) valores e estilo de vida.

No entanto, apesar dos desafios de carga cultural e limitações que os tradutores enfrentam quando traduzem textos turísticos, Kelly (1997) e Muñoz (2012) lembram que a indústria turística tem o hábito de contratar tradutores amadores ou despreparados para traduzir os textos que produz, o que pode resultar em

produtos de baixa qualidade, transmitindo, consequentemente, uma imagem negativa do local que está sendo divulgado. Essa má qualidade das traduções do setor já foi alvo de críticas de autores como Duff (1981), Newmark (1993) e Snell-Hornby (1999). A última, ao identificar características de textos publicitários em textos turísticos, afirma que, para países que dependem do turismo, como a Espanha e o Brasil, a publicidade eficaz é essencial, portanto, a tradução desses textos deve ser pensada com cuidado.

3 Metodologia

Conforme já referido, nossa pesquisa vale-se da Linguística de *Corpus* como metodologia e está sendo conduzida a partir de um estudo de um *corpus* paralelo e de um *corpus* comparável. Além disso, foi necessário compilar um *corpus* de referência em português e outro em inglês para a extração de palavras-chave. Esse procedimento visa a obter uma comparação entre um *corpus* de estudo, que é específico, e a linguagem geral, representada pelo *corpus* de referência. Mais adiante detalharemos os materiais que compõem esses *corpora*, o processo de exploração e o propósito do estudo de cada um deles. Antes, discorreremos sobre o AntConc e o AntPconc, ferramentas de análise linguística utilizadas para a realização deste estudo.

3.1 AntConc

Utilizado por muitos cursos de Linguística que visam a apresentar a seus alunos uma ferramenta gratuita e de fácil manipulação para o estudo e observação de padrões de uso da língua (ANTHONY, 2013), o AntConc foi desenvolvido em 2002 por Laurence Anthony, professor na Faculdade de Ciências e Engenharia da Universidade de Waseda, Japão. O *software* é livre e, após seu download, não é necessário que seja instalado. O AntConc possui ferramentas que permitem a geração de lista de palavras, extração de palavras-chave, linhas de concordância e lista de colocados. Na Figura 1, apresentamos a tela inicial do *software*.

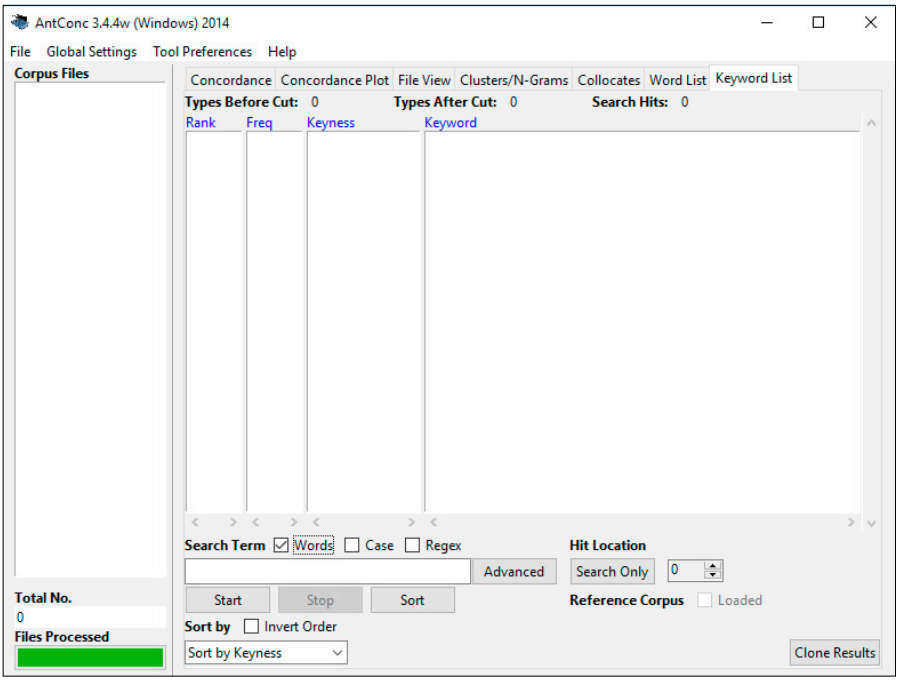


Figura 1 – Tela inicial do AntConc 3.4.4
Fonte: *Print screen* do programa no sistema operacional Window10

A seguir, descrevemos algumas das ferramentas do AntConc que estão sendo utilizadas na nossa pesquisa em andamento:

1) Wordlist (lista de palavras): essa ferramenta gera uma lista de todas as palavras presentes no *corpus* em análise investigado por ordem de frequência e disponibiliza o número total de *tokens* (número total de palavras) e de *types* (formas ou palavras distintas).

2) *Keywords* (palavras-chave): a ferramenta apresenta as palavras mais representativas de um *corpus*, sendo elas o resultado da comparação da lista de palavras do *corpus* de estudo com a de um *corpus* de referência. As palavras que tiverem um número de ocorrências estatisticamente maior no *corpus* de estudo irão compor a lista de palavras-chave. Segundo Berber Sardinha (1999), as palavras-chave geradas normalmente indicam o tema do *corpus*.

3) *Concordance*: o concordanceador gera as linhas de concordância do termo pesquisado, exibindo o contexto no qual o termo está inserido. A palavra em análise aparece no centro da linha, com o restante do texto à sua esquerda e à sua direita, como ilustrado pela Figura 2.

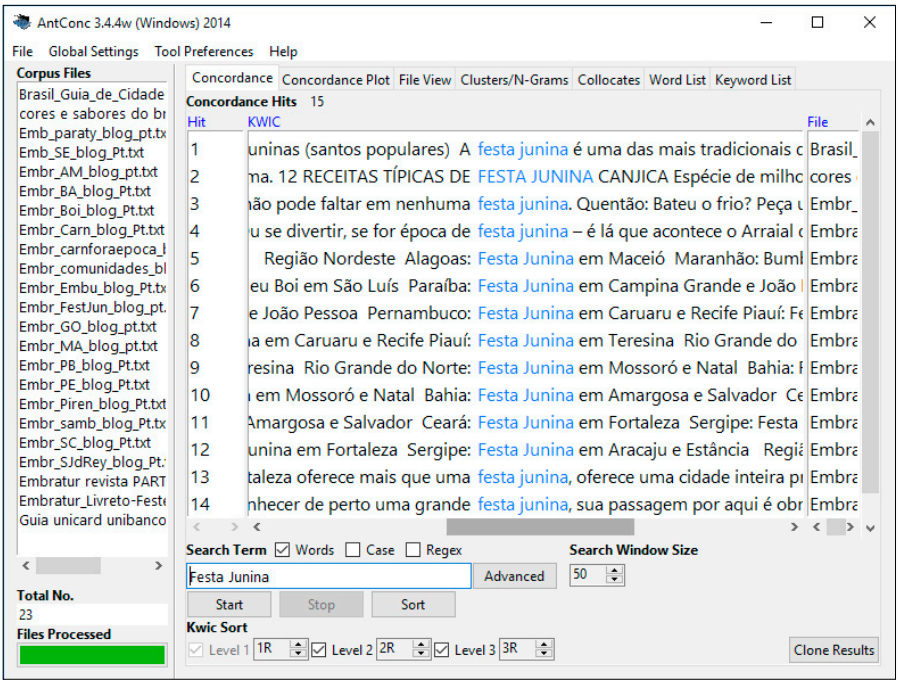


Figura 2 – Tela do concordanceador no AntConc 3.4.4
Fonte: *Print screen* do programa no sistema operacional Window10

4) *File View:* dependendo da pesquisa sendo conduzida, é necessário observar passagens maiores do texto, o que pode ser feito com o *File View*. A ferramenta nos dá acesso ao texto integral de cada fonte (que deve ser explorada individualmente) onde o termo em análise está inserido, permitindo sua visualização em um contexto maior. Isso possibilita ao pesquisador a observação de aspectos que poderiam ficar fora da análise do *corpus* se somente as linhas de concordância fossem levadas em conta. A Figura 3 ilustra a tela dessa ferramenta com o termo "Festa Junina" em destaque.

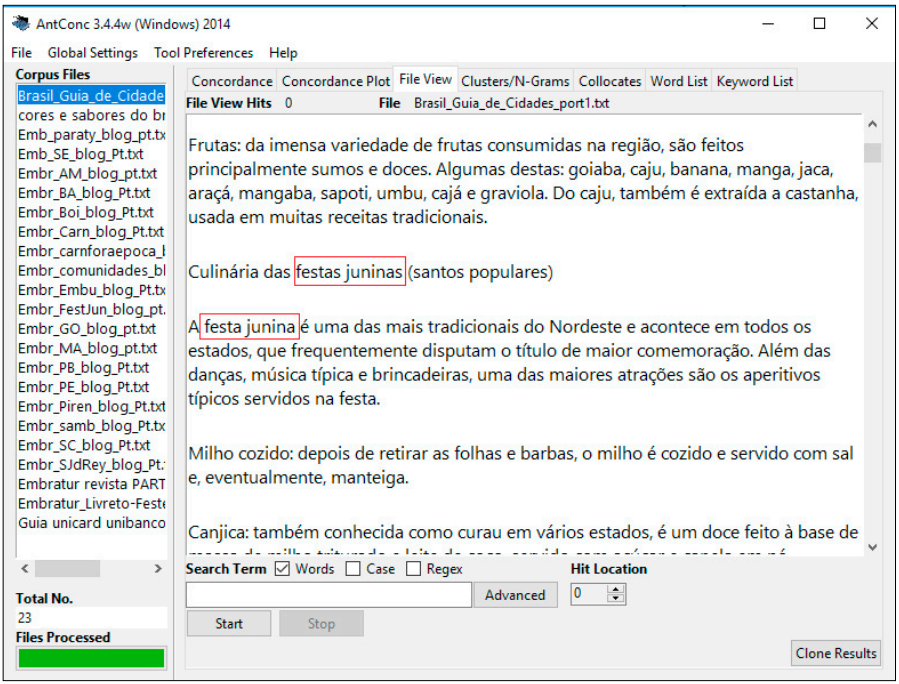


Figura 3 – Tela do *File View* no AntConc 3.4.4
Fonte: *Print screen* do programa no sistema operacional Window10

3.2 AntPconc

Lançado por Laurence Anthony em 2014, esse *software* é um concordanceador de *corpora* paralelos. Para a exploração de um *corpus* nesse programa, cada linha de um texto escrito no idioma original deve coincidir com a linha de seu texto traduzido. Assim, ao buscarmos determinado termo, uma janela é aberta e todas as suas ocorrências no idioma original aparecem na parte superior e cada linha correspondente no texto traduzido aparece na parte inferior, como mostra a Figura 4.

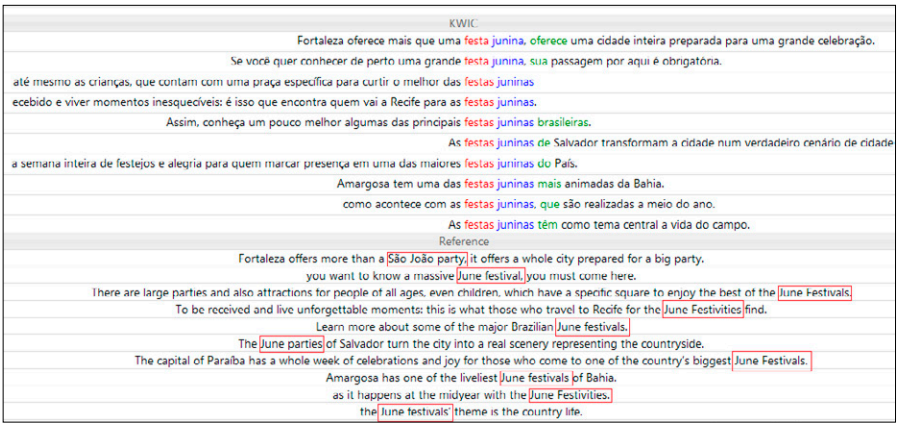


Figura 4 – Linhas de concordância para "Festa Junina" no AntPconc
Fonte: *Print screen* do programa no sistema operacional Window10

Tendo discorrido sobre as ferramentas utilizadas para a análise dos nossos *corpora*, apresentamos abaixo os materiais que os compõem e os detalhes de sua exploração.

3.3 Corpus *comparável*

Nosso *corpus* comparável é formado por textos turísticos escritos originalmente em português e em inglês. O Quadro 1 detalha a composição do *subcorpus* em português.

Quadro 1 – Composição do *subcorpus* em português

Título da obra	Guia Unicard Unibanco Brasil
Ano de publicação:	2005
Autor:	Unicard Unibanco
Editora:	Bei Comunicações
Número de palavras:	135.332
Título da obra:	Guia de Cidades
Ano de publicação:	2012
Autor:	EMBRATUR
Número de palavras:	122.892
Título da obra	Livreto Junino
Ano de publicação:	2013
Autor:	EMBRATUR
Número de palavras:	3.009
Título da obra	Cores e Sabores do Brasil
Ano de publicação:	2014
Autor:	EMBRATUR
Número de palavras:	7.356
Título da obra	Partiu Brasil
Ano de publicação:	2015
Autor:	EMBRATUR
Número de palavras:	16.569
Título da obra:	Blog EMBRATUR
Ano de publicação:	2016
Autor:	Visit Brazil - EMBRATUR
Número de palavras:	5.158
Total geral de palavras:	282.960

Fonte: Elaborado pela autora

Como pode ser visto, das 6 publicações que compõem esse *subcorpus*, 5 são materiais disponíveis no *site* da EMBRATUR (Instituto Brasileiro de Turismo) para divulgação do Brasil no exterior. Somente o *Guia Unicard Unibanco* é uma publicação de empresa privada. Durante a coleta desse material, notamos que há uma escassez de guias turísticos sobre o Brasil escritos originalmente em português. À exceção do *Guia Unicard Unibanco*, todos os outros encontrados eram traduções para o português de publicações estrangeiras escritas originalmente em inglês.

Já os três guias que compõem o *subcorpus* em inglês, como será visto abaixo, são publicações estrangeiras de empresas privadas, sendo o *Lonely Planet* de empresa americana e o *Eyewitness Travel Guide* e o *Rough Guides* de empresas britânicas. Como os três guias juntos totalizavam 838.171 palavras, nos foi necessário extrair as passagens que discorriam especificamente sobre festas populares para que, assim, pudéssemos ter os dois *subcorpora* balanceados. O Quadro 2 detalha a composição desse *subcorpus*.

Quadro 2 – Composição do *subcorpus* comparável em inglês

Título da obra	*The Rough Guide to Brazil*
Ano de publicação:	2014
Autores:	Kiki Deere, Daniel Jacobs, Stephen Keeling e Clemmy Manzo
Editora:	Rough Guides
Número de palavras:	105.352
Título da obra	Eyewitness Travel Brazil
Ano de publicação:	2016
Autor:	Dorling Kindersley
Editora:	Dorling Kindersley
Número de palavras:	35.750
Título da obra	Lonely Planet Brazil
Ano de publicação:	2016
Autor:	Lonely Planet
Editora:	Lonely Planet
Número de palavras:	65.371
Total geral de palavras:	206.473

Fonte: Elaborado pela autora

Para que o AntConc pudesse ler o nosso *corpus*, os textos que estavam no formato *.pdf precisaram ser alterados para *.txt. Essa conversão foi feita com o programa ABBYY Fine Reader[5], que faz o reconhecimento óptico de caracteres (do inglês *optical character recognition,* OCR), convertendo arquivos em *.pdf e documentos digitalizados em dados editáveis.

3.3.1 Exploração do *subcorpus* em português

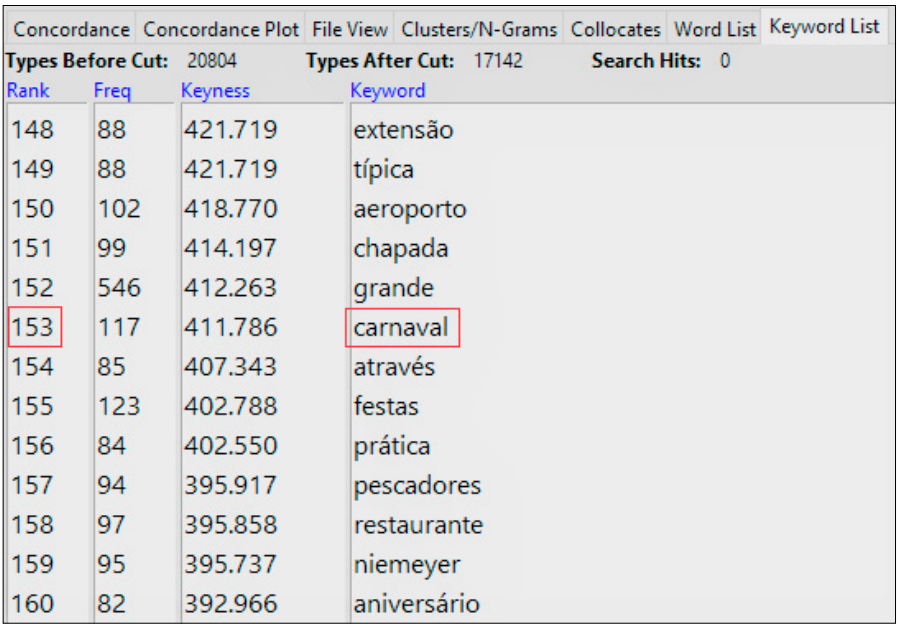
Após corrigirmos alguns erros decorrentes da conversão dos arquivos de *.pdf para *.txt, os textos do *subcorpus* em português foram submetidos à análise semiautomática com a ferramenta *Keywords* do AntConc 3.4.4 para o levantamento da lista de palavras-chave, a partir da qual extraímos os termos relativos a eventos populares brasileiros. Como dito anteriormente, para levantarmos as palavras-chave de um *corpus* de estudo, precisamos contrastá-lo com um *corpus* de referência. Abaixo, detalhamos o processo de compilação do *corpus* de referência em português utilizado nesta pesquisa.

3.3.1.1 O corpus de referência

Visando à otimização da busca nas palavras-chave pelo que era característico da culinária brasileira e constatando que um *corpus* de referência de generalidades

[5] O programa pode ser encontrado em <https://www.abbyy.com/pt-br>.

revelava palavras recorrentes em qualquer receita culinária, Rebechi (2015), que, em sua tese de doutorado, elaborou um dicionário de culinária brasileira, utiliza como referência um *corpus* de culinária geral. A exemplo de Rebechi (2015), compilamos para esta pesquisa um *corpus* de referência para cada idioma do nosso *corpus* de estudo, compostos por textos de turismo geral, excluindo aqueles sobre o Brasil. Dessa forma, os termos comuns aos textos turísticos seriam anulados e termos relacionados à cultura brasileira sobressairiam. A título de ilustração, vejamos, na Figura 5, a posição do termo "carnaval" no nosso *subcorpus* de estudo em português quando comparado ao *subcorpus* de língua geral do *Lácio-Ref*[6] e, na Figura 6, a posição do mesmo termo quando utilizamos o *corpus* de referência com textos de turismo geral.


Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Types Before Cut: 20804 **Types After Cut:** 17142 **Search Hits:** 0

Rank	Freq	Keyness	Keyword
148	88	421.719	extensão
149	88	421.719	típica
150	102	418.770	aeroporto
151	99	414.197	chapada
152	546	412.263	grande
153	117	411.786	carnaval
154	85	407.343	através
155	123	402.788	festas
156	84	402.550	prática
157	94	395.917	pescadores
158	97	395.858	restaurante
159	95	395.737	niemeyer
160	82	392.966	aniversário

Figura 5 – Posição de "carnaval" com o *Lácio-Ref* como *corpus* de referência
Fonte: *Print screen* do programa no sistema operacional Window10

[6] O Lácio-Ref é o *corpus* de referência do projeto Lácio-*Web*. Mais informações sobre o projeto em: <http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm.

Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Types Before Cut: 20804 **Types After Cut:** 15627 **Search Hits:** 0

Rank	Freq	Keyness	Keyword
69	198	221.262	cultural
70	109	219.969	brasília
71	183	219.072	festa
72	108	214.418	infraestrutura
73	70	214.397	pça
74	117	211.368	carnaval
75	106	209.132	ondas
76	287	208.351	visitantes
77	66	202.146	cerrado
78	1819	201.275	das
79	789	200.958	pela
80	78	198.594	amazonas
81	183	191.488	habitantes

Figura 6 – Posição de "carnaval" com um *corpus* de referência de turismo geral
Fonte: *Print screen* do programa no sistema operacional Window10

As listas apresentam as palavras em ordem decrescente de chavicidade. Como pode ser observado, o termo "carnaval" ocupa a 153ª posição quando um *corpus* de referência de língua geral é utilizado (Figura 5). O mesmo termo sobe 79 posições quando usamos o *corpus* de referência com textos de turismo geral (Figura 6), o que contribui para a otimização da busca pelo que é típico do Brasil, justificando a nossa escolha em utilizar um *corpus* de referência de turismo geral.

Para ajudar na compilação do *corpus* de referência em português, utilizamos o *BootCat*, *software* gratuito que cria *corpora* simples rapidamente a partir da *web*, tendo como base palavras-chave inseridas pelos usuários. Para tanto, inserimos alguns termos de turismo geral no programa, que, por sua vez, elaborou um *corpus* para o português de aproximadamente 540 mil palavras. O Quadro 3 exibe as palavras inseridas no programa.

Quadro 3 – Palavras-chave em português inseridas no *BootCat*

Idioma	Palavras-chave
Português	turismo, viagem, Europa, Ásia, Oceania, América do Norte

Fonte: Elaborado pela autora

Para ampliarmos o número de palavras do *corpus* de referência, a fim de que fosse, no mínimo, três vezes maior que nosso *corpus* de estudo[7], adicionamos textos de guias de turismo geral. Dessa forma, nosso *corpus* de referência em português (doravante CorTurRef_Pt) tem pouco mais de um milhão e cem mil palavras, sendo quatro vezes maior que o *subcorpus* de estudo em português.

3.3.1.2 Extração dos termos para análise

Na tentativa de eliminarmos a subjetividade da busca manual pelos termos na *Keywords* do AntConc e nos certificarmos de que os escolhidos são típicos do Brasil, eles foram confrontados em três obras produzidas no Brasil sobre a cultura do país[8]. São elas: *O dicionário do folclore brasileiro* (2012), *Festas populares do Brasil* (2010) e *Maravilhas do Brasil: festas populares* (2006). Para que os termos escolhidos fossem confirmados como próprios do Brasil, eles deveriam constar de pelo menos duas das três obras escolhidas. Sendo assim, trinta e dois termos foram selecionados e divididos em cinco categorias. Uma delas é a categoria de festas populares, as outras quatro são categorias referentes a essas festas, como detalhado abaixo.

- **Festas populares:** Carnaval, Bumba-meu-boi/Boi-bumbá, Festival Folclórico de Parintins, Festa(s) Junina(s) / Festa (s) de São João, Festa do Divino, Círio de Nazaré, Procissão do Fogaréu, Cavalhadas.
- **Músicas e Danças:** samba, samba-enredo, forró, xote, xaxado, frevo, marchinha(s), axé, tambor-de-crioula, quadrilha(s).
- **Símbolos:** (Boi) Caprichoso, (Boi) Garantido, boneco(s) gigante(s) de Olinda, arraial.
- **Organização:** escola(s) de samba, trio(s) elétrico(s), bloco(s) de rua/de carnaval, bumbódromo, sambódromo, abadá.
- **Comidas típicas:** quentão, pé de moleque.

3.3.2 Exploração do *subcorpus* em inglês

Após o levantamento desses termos, iniciamos o processamento do *subcorpus* de textos turísticos em inglês sobre o Brasil a fim de averiguarmos os elementos que caracterizam as festas populares brasileiras nos guias turísticos escritos nesse idioma. O mesmo procedimento adotado para a extração de palavras-chave no

[7] Como o tamanho do *corpus* de referência influencia o número de palavras-chave obtido, é ideal que ele seja de três a cinco vezes maior que o *corpus* de estudo (SARDINHA, 2004).
[8] Esse mesmo procedimento foi realizado por Costa (2006) em sua dissertação de mestrado para a identificação do que era típico do Brasil em materiais de divulgação cultural.

subcorpus em português foi realizado no *subcorpus* em inglês, porém, usando o CorTurRef_Ing, *corpus* de referência em inglês de turismo geral compilado para esta pesquisa nos mesmos moldes do CorTurRef_Pt. No Quadro 4, exibimos as palavras-chave inseridas no *BootCat* para sua compilação.

Quadro 4 – Palavras-chave em inglês inseridas no *BootCat*

Idioma	Palavras-chave
Inglês	tourism, travel, trip, Europe, Asia, Oceania, North America

Fonte: Elaborado pela autora

O *BootCat* gerou um *corpus* de 450 mil palavras, o que, mais uma vez, não seria suficiente, já que nosso *corpus* de estudo em inglês tem pouco mais de 206 mil palavras. Dessa forma, nos foi necessário adicionar textos de guias de turismo geral em inglês para que esse *corpus* de referência também fosse, no mínimo, três vezes maior que o de estudo. Ao final desse processo, o CorTurRef_Ing apresentou 815.213 palavras, sendo quase quatro vezes maior que nosso *subcorpus* em inglês. Ao serem confrontados, um total de 13.502 palavras-chave foram geradas. Nesta etapa, nos foi possível observar que muitos dos elementos típicos das festas brasileiras foram transferidos em português para os guias, facilitando a identificação dos termos que fazem referência às festas populares brasileiras.

A título de ilustração, a Tabela 1 exibe as dez primeiras palavras-chave relacionadas a festas populares brasileiras nos *subcorpora* em português (à esquerda) e em inglês (à direita) de textos turísticos sobre o Brasil.

Tabela 1 – Dez primeiras palavras-chave (KW) referentes a festas populares brasileiras nos *subcorpora* de textos turísticos em português (à esquerda) e em inglês (à direita)

	KW	Posição	Frequência		KW	Posição	Frequência
1	João	13	248	1	Carnaval	10	310
2	Festa	71	183	2	Samba	17	183
3	Carnaval	74	117	3	Forró	36	104
4	Samba	90	69	4	João	37	104
5	Forró	93	58	5	Bloco	46	76
6	Arraial	110	54	6	Boi	60	63
7	Boi	122	78	7	Festa	111	70
8	Parintins	148	40	8	Bumba	117	37
9	Festas	158	123	9	Schools	124	72
10	Juninas	166	36	10	Divino	142	32

Fonte: Elaborado pela autora

Com o estudo desse *subcorpus* estamos extraindo dados que nos auxiliam nas escolhas lexicais para a redação da definição das unidades terminológicas, além de possíveis equivalentes, quando for o caso.

3.4 Corpus *paralelo*

Nosso *corpus* paralelo é composto por textos turísticos sobre o Brasil escritos originalmente em português com suas respectivas traduções para o inglês. Por serem edições bilíngues, o material que compõe o *subcorpus* em português é o mesmo que compõe o *subcorpus* nesse idioma no *corpus* comparável, detalhado anteriormente no Quadro 1. Abaixo, o Quadro 5 traz as versões que compõem o *subcorpus* em inglês.

Quadro 5 – Composição do *subcorpus* em inglês

Título da obra:	Unicard Unibanco Brazil Guide
Texto-fonte:	Guia Unicard Unibanco Brasil
Ano de publicação:	2005
Autor:	Unicard Unibanco
Editora:	Bei Comunicações
Tradução:	Ibycaba Traduções
Número de palavras:	135.202
Título da obra:	Tourist Guide
Texto-fonte:	Guia de Cidades
Ano de publicação:	2012
Autor:	EMBRATUR
Tradução:	Tes Brasil
Número de palavras:	108.621
Título da obra:	June Festivals
Texto fonte:	Livreto Junino
Ano de publicação:	2013
Autor:	EMBRATUR
Tradução:	Não consta
Número de palavras:	2.894
Título da Obra:	Colours and Flavours of Brazil
Texto-fonte:	Cores e Sabores do Brasil
Ano de publicação:	2014
Autor:	EMBRATUR
Tradução:	Não consta
Número de palavras:	7.241
Título da obra:	Visit Brazil
Texto-fonte:	Partiu Brasil
Ano de publicação:	2015
Autor:	EMBRATUR
Tradução	Não consta
Número de palavras:	13.594

Título da obra:	Embratur Blog
Texto-fonte:	Blog Embratur
Ano de publicação:	2016
Autor:	EMBRATUR
Tradução:	Não consta
Número de palavras:	5.293
Total geral de palavras:	**272.845**

Fonte: Elaborado pela autora

Esse *corpus* foi inicialmente explorado no AntPconc 1.1.0 para que pudéssemos identificar e analisar as traduções dos termos selecionados. Além de estudarmos essas traduções por meio das linhas de concordância do programa, também utilizamos a ferramenta *File View* do AntConc, que, como dito anteriormente, nos permite visualizar o texto integral, tornando possível a observação de aspectos não revelados pelas linhas de concordância. Ressaltamos aqui a importância de se manter o arquivo original de onde os textos que compõem o *corpus* foram extraídos. Esses arquivos nos dão acesso à linguagem não verbal dos textos, o que pode evidenciar elementos interessantes de serem analisados.

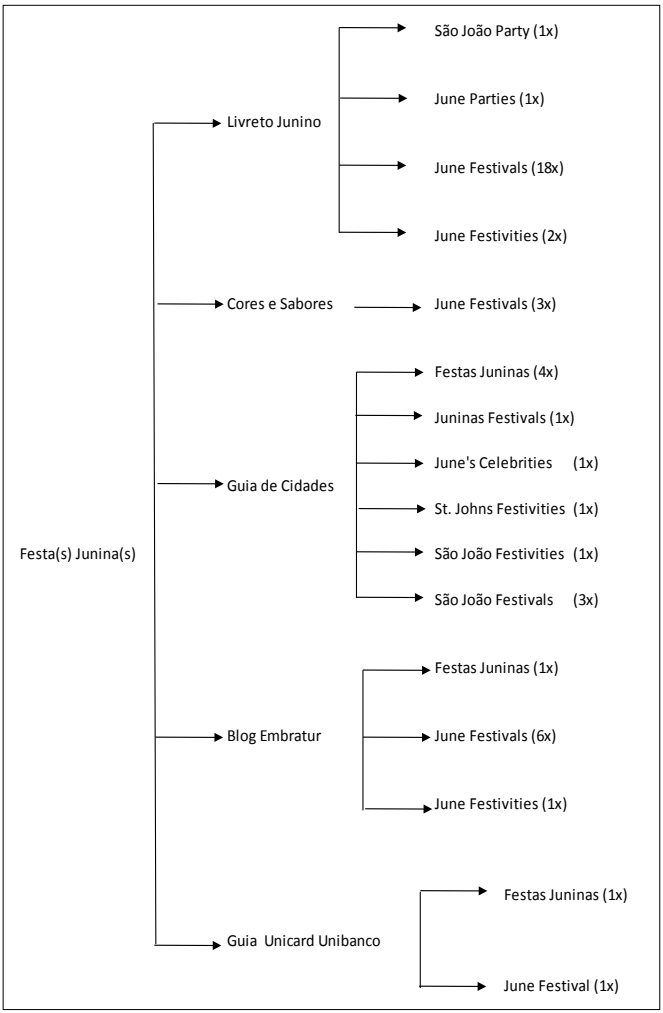
Com o estudo desse *corpus*, estamos identificando se há ou não uma padronização de escolhas tradutórias e detectando os desafios encontrados pelos tradutores quando da tradução dos termos selecionados, o que nos auxilia na elaboração da proposta do vocabulário de festas populares brasileiras, no sentido de que possibilita uma reflexão sobre como uma obra terminográfica poderá vir ao encontro das necessidades dos tradutores de textos turísticos sobre o Brasil.

Na próxima seção, detalharemos a análise do termo "Festa Junina" e exibiremos o nosso primeiro modelo de verbete.

4 Análise do Termo "Festa Junina"

4.1 Análise do corpus paralelo

O Esquema 1 apresenta as soluções encontradas pelos tradutores para a versão do termo "Festa Junina" para o inglês, obtidas com a visualização do nosso *corpus* paralelo no AntPconc. As traduções são exibidas indicando a fonte onde se encontram. O número de ocorrências de cada uma está indicado entre parênteses.



Esquema 1 – Traduções do termo "Festa Junina"
Fonte: Elaborado pela autora

Segundo os dicionários *on line* de língua portuguesa *Caldas Aulete*, *Priberam da Língua Portuguesa* e *Michaelis – Dicionário Brasileiro da Língua Portuguesa*, o qualificador "junino(a)" refere-se ao que acontece no mês de junho e às festas que nele se realizam. Daí, possivelmente, a escolha por "June" para a tradução de "Junina" em 33 das 46 ocorrências do termo em nosso *corpus*.

Em 6 ocasiões, a solução encontrada pelos tradutores foi verter "Junina" por "São João" (5 ocorrências) e "St. Johns" (sic) (1 ocorrência). Ao observarmos trechos maiores do texto, pudemos ver que as passagens onde essas traduções estão

inseridas não esclarecem o leitor sobre o teor religioso dessas festas, portanto, a estratégia de verter "Junina" por "São João" ou "St. Johns" pode ser uma tentativa de explicar que se trata de uma festa religiosa, já que as limitações de espaço possivelmente não permitiam uma adição de informação mais elaborada.

Como pudemos ver acima, das soluções encontradas pelos tradutores do *Guia de Cidades*, uma delas foi a de traduzir "Festas Juninas" por "St. Johns Festivities". Aqui, um apóstrofo, que deveria estar presente antes da letra "s" indicando o genitivo em "John", não foi incluído, resultando em erro. Erro, de acordo com Cruces Colado (2001) é a ruptura das regras de coerência de um texto-alvo, sejam gramaticais, lexicais, semânticas ou culturais. No mesmo guia, uma outra solução encontrada pelos tradutores foi a de verter o termo em análise por June's Celebrities, resultando novamente em erro, uma vez que "celebrities" não significa "festa", "comemoração" ou "celebração", mas sim, "celebridades". Ainda no *Guia de Cidades*, uma das traduções escolhidas para "Festas Juninas" foi "Juninas Festivals". Nesse caso, uma consulta ao texto integral nos mostrou que há uma explicação sobre a origem do qualificador em questão, informando que as festas são assim chamadas por acontecerem no mês de junho. Essa explicação está presente tanto no texto-fonte como no texto-meta.

Um ponto importante a ser comentado é quantidade de soluções diferentes encontradas para a tradução de "Festa(s) Junina(s)" no mesmo material, como acontece principalmente no *Guia de Cidades*. Esse guia apresenta o seu conteúdo dividido por região. Como todas as regiões brasileiras comemoram as Festas Juninas, o termo aparece ao longo de todo o material. Diferentemente, por exemplo, do *Guia Unicard Unibanco* e da revista *Cores e Sabores* que só apresentam o termo em uma passagem sobre festas populares brasileiras. No caso do *Guia de Cidades*, por mais que o cotexto informasse o leitor sobre alguns detalhes da festa, o fato de não haver uma padronização para a tradução do termo dentro da mesma fonte pode, muitas vezes, confundir o leitor. Por exemplo, uma análise mais detalhada desse guia mostrou que em nenhum momento afirma-se que as Festas Juninas também são conhecidas como "Festas de São João", como no caso de Campina Grande que tem o chamado "maior São João do Mundo" e que acontece não só no dia 24/06 (dia de São João) mas em todo o mês de junho. Dessa forma, se no *Guia de Cidades* o leitor depara com o termo "June Festivals" em uma página e com o termo "São João Festivals" três páginas adiante sem que se faça ligação entre as duas festas, ele pode supor que são dois eventos diferentes. Ainda, quando o mesmo guia discorre sobre as festas típicas do Ceará, a sua versão em inglês informa que os turistas vão às Festas de São João atraídos por comidas e diversões típicas *dessa época do ano*: 4a) "The São João festivities in Ceará attract large numbers of tourists in search of [...] food and fun typical of *this time of year* (grifo nosso). Porém, não é mencionado que a época do ano em questão é junho. Já na versão em português, lemos: 4b) "As *Festas Juninas* atraem turistas em busca

de [...] comidas e brincadeiras típicas dessa época do ano" (grifo nosso). Como o brasileiro sabe o mês em que essas festas acontecem, essa informação faz sentido para ele. O mesmo não podemos dizer do estrangeiro que leu a versão em inglês. Nesse caso, a adição dessa informação se fazia necessária para que o texto cumprisse a função de atrair o leitor para a festa em questão.

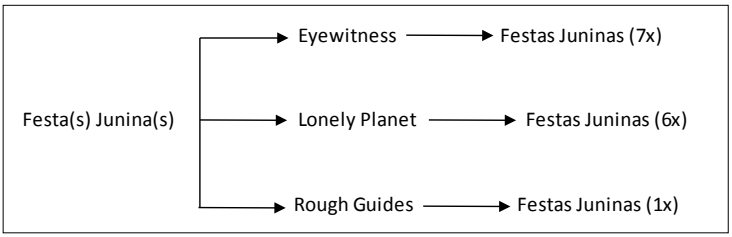
Chamaremos a atenção agora para a escolha dos equivalentes de "festa" no *corpus* de estudo. Em uma pesquisa nos dicionários *Oxford*, *Cambridge*, *Merriam-Webster* e *Macmillan*, vimos que, dos três termos empregados aqui como equivalentes de "festa" ("party", "festival" e "festivities"), somente "festival" tem conotação religiosa, tendo "feast" como sinônimo. Esse último, por sua vez, é definido como o dia em que determinado santo é celebrado e as festas em homenagem a ele. Ao buscarmos esses quatro termos no *Corpus* of Contemporary American English (COCA)[9], confirmamos que tanto "festival" quanto "feast" aparecem ligados a festas religiosas e ao lado de nomes de santos, "Feast of St. Peter", "Festival of St. John the Baptist", o que não acontece com "festivity" – apesar de constar desses dicionários como sinônimo de "festival" – nem com "party". Além disso, uma pesquisa no nosso *subcorpus* de textos turísticos escritos originalmente em inglês mostrou que apenas os termos "festival" e "feast" foram usados para descrever as festas religiosas, enquanto "festivities" foi usado no sentido de "festividades" em geral.

Das 46 ocorrências do termo no *corpus*, em somente 6 momentos a solução foi a de transferi-lo em português para o texto-meta. Nesses casos, há uma explicação – que já estava presente no texto-fonte – sobre quando e onde essas festas são celebradas.

4.2 Corpus *comparável: análise do* subcorpus *em inglês*

O Esquema 2 apresenta o tratamento dado para "Festa(s) Junina(s) pelos autores do material que compõe o nosso *subcorpus* em inglês do *corpus* comparável.

[9] Disponível em: <http://corpus.byu.edu/coca/>.



Esquema 2 – "Festa(s) Junina(s)" no *subcorpus* comparável em inglês
Fonte: Elaborado pela autora

Estudando o esquema acima podemos ver que, sempre que se faz menção às Festas Juninas nas fontes analisadas, o nome da festa é mantido em português. Temos, assim, 14 ocorrências do termo nesse *corpus*: sete no guia *Eyewitness*, seis no *Lonely Planet* e apenas uma no *Rough Guides*.

Como dito anteriormente, a ideia é que esse *subcorpus* nos ajude a escolher as palavras, expressões e construções apropriadas em língua inglesa para a redação da definição dos verbetes. Porém, no caso do termo em análise, o estudo desse *subcorpus* revelou que os autores de dois desses guias pouco sabem sobre as Festas Juninas. O nosso primeiro exemplo vem do *Eyewitness Travel*, no qual todas as ocorrências do termo estão em uma passagem sobre Campina Grande, onde é explicado que as Festas Juninas fazem parte do folclore nordestino. Informação imprecisa, já que são comemoradas em todo o Brasil. O *Rough Guides*, por sua vez, informa que essas festas acontecem exclusivamente na terceira semana de junho e apenas no sudeste brasileiro, sendo voltadas principalmente ao público infantil. Dos três materiais, o *Lonely Planet* é o único que não traz informações incorretas sobre as Festas Juninas. O guia discorre brevemente sobre elas em uma seção sobre eventos populares e ao dissertar sobre alguns estados brasileiros, afirmando que são comemoradas em todo o Brasil, sendo um dos principais festivais folclóricos do país.

É fundamental ressaltarmos que essas informações erradas trazidas por dois guias de renome no setor turístico descrevem a segunda maior festa popular brasileira, o que pode não apenas servir como obstáculo para o desenvolvimento do turismo no país, já que essas festas poderiam atrair milhares de turistas estrangeiros pelo grau de diversidade cultural que apresentam, como podem ir contra a reputação de guias premiados como os mencionados, caso o leitor se dê conta do equívoco. Essa descoberta confirma a nossa finalidade de propor um vocabulário bilíngue que tenha como público-alvo não apenas profissionais que traduzam textos turísticos sobre o Brasil para o inglês, mas também os autores desses textos que não tenham conhecimento sobre a cultura brasileira.

Como as passagens sobre as Festas Juninas nos guias eram muito curtas e não traziam muitos termos que pudéssemos usar na redação da definição do verbete,

nos foi necessário recorrer a outras fontes escritas originalmente em inglês sobre as Festas Juninas para extrairmos o léxico necessário para essa tarefa, a saber: o *site* do jornal *The Telegraph*, que dedicou uma de suas páginas às festas populares brasileiras, o *site The Culture Trip*, que publica artigos sobre as culturas de diversos países ao redor do mundo, e o *site The Fare Compare*, que faz comparações entre preços de passagens áreas e oferece dicas sobre eventos ao redor do mundo. Como os artigos dessas fontes são pequenos, a extração do léxico foi feita manualmente. O Quadro 6 mostra as palavras encontradas usadas na nossa definição.

Quadro 6 – campo lexical de "Festas Juninas" em inglês

Nomes dos santos celebrados	Saint Anthony Saint John the Baptist Saint Peter
Verbos	Feature Attend Enjoy Play Dance Win Gather Occur Celebrate
Colocações adjetivas	Rural life Rural clothing Small prizes
Substantivos	Feast Celebration Festivity Festival Honour Bonfire

Fonte: Elaborado pela autora

Na próxima subseção apresentaremos o nosso primeiro modelo de verbete baseado no estudo do termo "Festa Junina" no nosso *corpus* paralelo e no comparável.

4.3 Exemplo de verbete

Tomando por base a análise acima detalhada, concluímos que o verbete "Festa Junina" deve trazer os seguintes campos:

1) Unidade terminológica na língua de partida
2) Classe gramatical
3) Forma pluralizada da unidade terminológica na língua de partida, o que pode auxiliar um escritor estrangeiro que não está familiarizado com as regras de formação de plural da língua portuguesa.
4) Definição na língua de chegada. Esse campo poderá auxiliar os tradutores a buscarem por palavras em inglês que fazem parte do campo lexical em que o termo está inserido, por exemplo, no caso de "Festas Juninas", a definição trará construções como "feast / festival of St. John", dentre outros. Com o intuito de assegurarmos uma definição confiável do termo, outras fontes, como livros sobre o folclore brasileiro e dicionários, que complementassem e/ou corrigissem as informações contidas nos *corpora* foram consultadas.
5) Remissivas, que estarão no corpo da definição do termo, destacadas em azul ou no campo "see also", o que deverá ser explicitado na introdução da obra terminográfica.
7) When, campo que trará quando as Festas Juninas acontecem no país.
8) Where, campo que trará o local em que essas festas são celebradas.
9) Important dates, que trará uma lista das datas importantes relacionadas ao evento. Por exemplo, os dias de Santo Antônio, São João e São Pedro.
10) Fotos com legenda para melhor ilustração do termo.

Como pode ser visto, a ideia é que o vocabulário proposto inclua não só aspectos linguísticos, mas também extralinguísticos, como imagens e datas importantes.

O modelo do verbete "Festa Junina" pode ser visto na Figura 7.

Festa Junina

Adjectival collocation Pl: Festas Juninas *Festas Juninas* are the traditional feasts that occur at the beginning of the Brazilian winter in June in honour of Saint Anthony (Santo Antônio), Saint John the Baptist (São João) and Saint Peter (São Pedro). The theme of these festivities revolves around rural life. Those attending dress up in rural clothing, dance the *quadrilha*, are treated to typical Brazilian food such as *canjica* and *pamonha*, enjoy bonfires and play games to win small prizes. Every region in Brazil has its own way of celebrating the Festas Juninas. The celebration in Campina Grande, in the state of Paraíba (Northeast of Brazil), is the biggest one, gathering millions of visitors every year.

See also: São João

When? In June, extending until July in some cities.

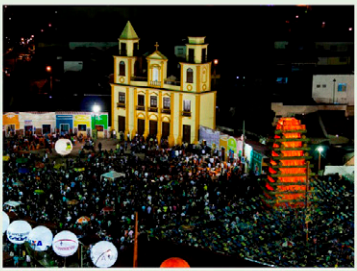
Where? All over Brazil.

Important Dates

Feast of St. Anthony: Jun/13

Feast of St. John: Jun/24

Feast of St. Peter: Jun/29



The *Festa Junina* in Campina Grande, state of Paraíba, is called "São João de Campina Grande".

Girl wearing vintage rural clothing.

Figura 7 – Modelo do verbete "Festa Junina"
Fonte: Elaborado pela autora, com fotos de Kyller Costa Gorgônio (Campina Grande) e Eduardo Coutinho (menina)

5 Considerações Finais

É importante lembrarmos que este trabalho trata de uma pesquisa em andamento, portanto, ainda serão necessárias as análises dos demais termos para que possamos fazer afirmações mais contundentes. No entanto, o estudo do termo "Festa Junina" nos nossos *corpora* constituiu uma amostra interessante de como podemos descobrir quais informações são úteis ao público-alvo do vocabulário proposto e que devam constar da definição dos verbetes. Foi o caso da análise do nosso *subcorpus* de textos originais em inglês. Com ele, conforme visto acima, tínhamos inicialmente o objetivo de investigar o campo lexical de "Festas Juninas" em inglês para auxílio da redação da definição do termo. O estudo desse *subcorpus* se tornou ainda mais pertinente ao descobrirmos que algumas das publicações estrangeiras que o compõem trazem informações erradas sobre essas festas, o que nos guiou nas informações que o modelo do verbete deveria trazer, como os campos que explicam quando e onde as Festas Juninas são comemoradas.

Já o estudo do *corpus* paralelo, como previsto, nos ajudou a definir o léxico que deveria estar presente na definição do verbete para guiar os tradutores na tarefa de produzir textos em inglês com mais naturalidade. Foi o caso do uso de "party" pelos tradutores como equivalente de "festa". Um estudo no *subcorpus* de textos escritos originalmente em inglês e em fontes externas nos revelou que "party" não costuma ser usado para designar festas religiosas, como é o caso das Festas Juninas, não entrando, portanto, na definição do verbete. Exemplos tirados desse *corpus*, como "Juninas Festivals", "June's Celebrities" e "São João Party", corroboram a afirmação de Kelly (1997) sobre o desafio especial que os tradutores enfrentam ao verter textos para um idioma que não seja a sua língua materna. Essa questão fica ainda mais complexa com a presença de *realia*. Uma obra terminográfica como a que tencionamos propor ao final desta pesquisa pode deixar esse desafio menos difícil. Porém, é importante esclarecer que não pretendemos que ela traga imposições sobre como um termo deve ou não ser traduzido, mas que sirva como um guia que auxilie tradutores e redatores de textos turísticos em inglês a produzirem textos com informações precisas e que fluam com mais naturalidade, o que só tem a contribuir positivamente para o contínuo desenvolvimento do turismo no Brasil.

Referências

ALIZADEH, A. Bridging cultures: Tourism and the art of translation. In: 2011 INTERNATIONAL CONFERENCE ON SOCIAL SCIENCE AND HUMANITY, Singapore. *Anais...*Vol. 5. Singapore: Lacsit Press, 2011, p. 261-264.

ANTHONY, L. *AntConc.* Laurence Anthony's *Website*. 2013. Disponível em: <http://www.laurenceanthony.net/software.html>. Acesso em: 09 out. 2017.

ARGONI, M. Tourism communication: the translator's responsibility in the translation of cultural difference. *Pasos: revista de turismo y patrimonio cultural*, Santa Cruz de Tenerife, v. 10, n. 4, p. 5-11, jun. 2012.

BOIERAS, G. *Festas populares*: maravilhas do Brasil. São Paulo: Escrituras, 2006.

BOUTIN-QUESNEL, R. *Vocabulaire systématique de la Terminologie.* Québec: Publications Du Québec, 1985. Tradução para o português de E. Faulstich.

CALVI, M. V. *Il linguaggio spagnolo del turismo*. Viaregio-lucca: Baroni, 2000.

_____. Los géneros discursivos en la lengua del turismo: una propuesta de clasificación. *Ibérica,* Castellón, v. 19, p. 9-31, 2010.

CAMBRIDGE UNIVERSITY PRESS. *Cambridge Dictionary*. Disponível em: <http://www.cambridge.org>. Acesso em: 08 out. 2017.

CASCUDO, L. da C. *Dicionário do Folclore Brasileiro*. 12. ed. São Paulo: Global Editora, 2012.

COSTA, A. T. P. *Brasil mostrando a sua cara: estratégias de tradução no material de divulgação cultural: um estudo baseado em* corpus*.* 2006. 233 f. Dissertação (mestrado). Curso de Linguística Aplicada, Departamento de Línguas Estrangeiras e Tradução, Universidade de Brasília, Brasília, 2006.

CRUCES COLADO, S. El origen de los errores en traducción. In: DOMINGO, P.; GONZÁLEZ, E. R. R.; DOLORES, J. Plaza (Coord.). *Écrire, traduire et représenter la fête.* Valencia: Universitat de València, 2001, p. 813-822.

DUFF, A. *The third language*: recurrent problems of translation into English. Oxford: Pergamon, 1981.

EDITORA MELHORAMENTOS. *Michaelis*: Dicionário Brasileiro da Língua Portuguesa. 2017. Disponível em: <http://michaelis.uol.com.br>. Acesso em: 08 out. 2017.

FAULSTICH, E. Socioterminologia: mais que um método de pesquisa, uma disciplina. *Ciência da Informação*, Brasília, v. 24, n. 3, 1995.

GEIGER, P. *Dicionário Online Caldas Aulete*. Disponível em: <http://www.aulete.com.br>. Acesso em: 08 out. 2017

KELLY, D. The translation of texts from the tourist sector: textual conventions, cultural distance and other constraints. *Trans: Revista de Traductología*, Málaga, n. 2, p. 33-42, 1997.

LUCAS, L. C. S. *Traducción para el turismo y el ocio.* Murcia: Universidad de Murcia, 2012.

MACMILLAN PUBLISHERS. *Macmillan Dictionary*. 2017. Disponível em: <http://www.macmillandictionary.com/>. Acesso em: 08 out. 2017

MARTINS, J. P. S. *Festas Populares do Brasil.* São Paulo: Komedi, 2010.

MERKAJ, L. Tourist communication: a specialized discourse with difficulties in translation. *European Scientific Journal*, Açores, v. 2, p. 321-325, dez. 2013.

MERRIAM-WEBSTER. *Merriam-Webster Dictionary*. 2017. Disponível em: <www.merriam-webster.com>. Acesso em: 08 out. 2017.

MOSS, C. *Brazil events*. The Telegraph. Atualizado em 14 out. 2015. Disponível em: <http://www.telegraph.co.uk/travel/destinations/south-america/brazil/articles/brazil-events/>. Acesso em: 06 out. 2017.

MUÑOZ, I. D. Analysing common mistakes in translations of tourist texts (Spanish, English and German). *Onomázein*, Málaga, v. 2, n. 23, p. 335-349, 2012.

NEWMARK, P. *Paragraphs on Translation.* Clevdon/Philadelphia/Adelaide: Multilingual Matters Ltd., 1993.

OXFORD UNIVERSITY PRESS. 2017. *Oxford Dictionaries*. Disponível em: <http://www.oxforddictionaries.com/>. Acesso em: 08 out. 2017

PRIBERAM. *Dicionário Priberam da Língua Portuguesa*. 2013. Disponível em: <https://www.priberam.pt/>. Acesso em: 08 out. 2017.

REBECHI, R. R. *A tradução da culinária típica brasileira para o inglês*: um estudo sob o enfoque da Linguística de *Corpus.* 2015. 393 f. Tese (doutorado). Curso de Estudos Linguísticos e Literários em Inglês, Departamento de Letras Modernas, Universidade de São Paulo, São Paulo, 2015.

SAMUELS, A. J. *Celebrating the Brazilian harvest with festa junina.* The Culture Trip. Atualizado em 25, out. 2016. Disponível em: <https://theculturetrip.com/south-america/brazil/articles/festa-junina-celebrating-the-brazilian-harvest/>. Acesso em: 06 out. 2017.

SARDINHA, T. B. A influência do tamanho do *corpus* de referência na obtenção de palavras-chave. *Direct Papers*, São Paulo, n. 38, p. 1-18, 1999. Disponível em: <http://www2.lael.pucsp.br/direct/direct_papers.htm>. Acesso em: 09 out. 2017.

_____. *Linguística de* Corpus. Barueri: Manole, 2004. 410 p.

SCHIESSER, R. *Festa junina*: what to expect. The fare compare. Atualizado em 10 out. 2015. Disponível em: <https://www.farecompare.com/eventurist/festa-junina/#/>. Acesso em: 06 out. 2017.

SNELL-HORNBY, M. The ultimate comfort: word, text and the translation of tourist brochures. In: ANDERMAN, G.; ROGERS, M. (Ed.). *Word, text, translation.* Clevedon: Multilingual Matters, 1999, p. 95-103.

Créditos das imagens no exemplo de verbete

Foto de Campina Grande: Kyller Costa Gorgônio/Flickr/WikiCommons – Liberada para reutilização.

Foto da menina com vestido caipira: Eduardo Coutinho/Flickr/Wiki Commons – Liberada para reutilização.

Frames de compreensão e *corpora*: estudo de caso com uso do Sketch Engine

Frames of understanding and *corpora*: a case study using Sketch Engine

Aline Nardes dos Santos
Rove Chishman

Resumo: O objetivo deste trabalho é relatar as vantagens do uso da ferramenta Sketch Engine em um estudo de caso que descreveu, a partir de frames semânticos, as conceptualizações de feto anencéfalo no contexto do processo da ADPF 54, por meio da qual foi autorizada a interrupção de gravidez de fetos anencefálicos no Brasil. O aporte teórico para a descrição dessas conceptualizações foi a abordagem dos frames de compreensão.O trabalho evidencia como o Sketch Engine facilitou a identificação de evocadores potenciais do frame analisado, principalmente por meio da ferramenta Word Sketch. Além disso, essa exploração inicial dos dados revelou algumas das facetas de conhecimento predominantes em cada *subcorpus*, bem como a sua relação com determinadas escolhas lexicais.

Palavras-chave: Sketch Engine. *Frames* de compreensão. ADPF 54.

Aline Nardes dos Santos – Doutoranda e mestra em Linguística Aplicada pela Universidade do Vale do Rio dos Sinos (Unisinos), bolsista CAPES/PROSUC – aline.nardes@gmail.com.
Rove Chishman – Professora titular da Universidade do Vale do Rio dos Sinos (Unisinos), doutora em Letras pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Bolsista Produtividade do CNPq – rove@unisinos.br.

Abstract: This paper aims at reporting the advantages of the use of Sketch Engine in a case study that described, through semantic frames, the conceptualizations of anencephalic fetus in the context of the ADPF 54 process, by means of which the interruption of pregnancy in case of anencephaly was authorizedin Brazil. The theoretical framework for description of conceptualizations was the frames of understanding approach. This case study shows how Sketch Engine facilitated the identification of potential evokers for the analyzed frame, especially through the use of the Word Sketch tool. In addition, this initial data exploration showed some of the knowledge facets that were predominant in each *subcorpus*, as well as its relations with some lexical choices.

Keywords: Sketch Engine. Frames of understanding. ADPF 54.

1 Introdução

Este trabalho consiste em um aprofundamento de nossa pesquisa realizada em nível de mestrado (SANTOS; CHISHMAN, 2017), com foco nas implicações do uso da ferramenta Sketch Engine para as reformulações que propusemos em nossa metodologia de análise. Nosso contexto de pesquisa foi o processo da Arguição de Descumprimento de Preceito Fundamental nº 54 (ADPF 54), que instituiu o direito à antecipação terapêutica de parto de anencéfalos às mulheres brasileiras, dado que o enquadramento desse procedimento como crime, no caso de anencefalia, feria direitos inconstitucionais da gestante, tais como liberdade, dignidade da pessoa humana e direito à saúde. A causa, defendida oficialmente pela Confederação Nacional dos Trabalhadores da Saúde (CNTS), foi julgada como procedente pelo Superior Tribunal Federal (STF), em virtude do sofrimento que a gravidez de anencéfalo causa à gestante: o distúrbio da anencefalia é irreversível, pois o tubo neural do feto não se fecha e causa a dissolução de sua massa encefálica; assim, "Em mais da metade dos casos, os fetos não resistem à gestação, e os poucos que alcançam o momento do parto sobrevivem minutos ou horas fora do útero" (DINIZ; VÉLEZ, 2008, p. 648).

A ADPF não foi apenas debatida em sessões do STF, mas também envolveu audiências públicas em que diversas entidades foram convidadas a expressar seu posicionamento acerca do processo – principalmente representantes de instituições médicas e religiosas. Diante de tal heterogeneidade de posicionamentos, tínhamos interesse em verificar como as perspectivas acerca da entidade *feto anencéfalo* se diferenciavam. Para isso, tomamos como ponto de partida uma noção de significado como *conceptualização* (LANGACKER, 1987; 2008), ou seja, como processo dinâmico que envolve operações sociocognitivas. Dessa forma, "conceptualizações podem constituir experiências sutilmente diferentes, dependendo das escolhas linguísticas feitas pelos falantes" (SANTOS, 2016). Mais especificamente, a categoria de análise utilizada foi o *frame* de compreensão (ZIEM, 2014; FILLMORE, 1985), estrutura sociocognitiva ativada interacionalmente por facetas de conhecimento, as quais são descritas a partir do esquema *slot-filler*.

De modo a identificar os *frames* relativos a *feto anencéfalo* que constavam no processo, a ferramenta Sketch Engine (SE) foi utilizada. Em princípio, tal uso se restringiria à coleta de concordâncias para posterior análise qualitativa. No entanto, em virtude da pertinência de recursos como a Word Sketch para a exploração do *corpus*, essa etapa permitiu outras explorações que resultaram em uma coleta mais sistemática de evocadores de *frame*. Além disso, o uso do Sketch Engine antecipou alguns resultados que foram evidenciados na etapa posterior de descrição qualitativa das estruturas conceptuais evocadas, incluindo algumas das facetas de conhecimento predominantes em cada *subcorpus*, bem como a sua relação com determinadas escolhas lexicais, conforme será ilustrado neste trabalho.

Assim, o objetivo principal deste artigo é abordar as vantagens do uso da ferramenta Sketch Engine (doravante também referida por meio da sigla SE) em nosso estudo de caso, visto que o recurso não só atendeu aos nossos propósitos de pesquisa, mas também contribuiu para que expandíssemos o procedimento inicial de identificação de evocadores, permitindo que fizéssemos uma análise prévia das facetas de conhecimento que constituíam o *frame feto anencéfalo* em cada *subcorpus*.

Em consonância com esse propósito, o texto se organiza da seguinte forma: na segunda seção, contextualizamos a noção de *frame* semântico (FILLMORE, 1982; 1985), especificando a categoria do *frame* de compreensão, retomada por Ziem (2014) a partir dos estudos fillmorianos. Na seção 3, contextualizamos o *corpus* e justificamos a escolha do SE para a extração das informações linguísticas. Na seção 4, trazemos exemplos das vantagens do uso do Sketch Engine na identificação de *frames* de compreensão a partir de *corpora*. Por fim, na seção 5, tecemos algumas considerações, indicando as limitações dessa aplicação do SE e consolidando nossa proposta metodológica a partir dos resultados alcançados.

2 Frames semânticos

A noção de *frame* não pertence propriamente à Linguística, visto que surge a partir de uma série de pressupostos provenientes de áreas distintas. Dentre os conceitos explicitados por Fillmore (1982), como teorias subjacentes à estrutura do *frame* preconizada em seus trabalhos, estão o *esquema* (BARTLETT, 1932), o *script* (SCHANK; ABELSON, 1977), o *frame* computacional, de Minsky (1974) e o *frame* sociológico-interacional, proposto por Goffman (1975). A atividade comunicacional de ativação cognitiva dessas estruturas pode ser definida como "a recorrência, na percepção, no pensamento e na comunicação, a maneiras estruturadas de se interpretar experiências" (FILLMORE, 1976, p. 20).

De modo geral, essa recorrência pode ser ilustrada por meio do exemplo favorito de Lakoff (2004) ao explicar o conceito de *frame* aos seus alunos: ao adentrar a sala de aula, o pesquisador inicia sua fala com a seguinte exclamação:

não pense em um elefante! Naturalmente, seus alunos se veem impossibilitados de evitar pensar justamente em um elefante e em todos os conhecimentos relacionados a essa entidade. Como explica o autor, isso acontece pelo seguinte motivo:

> Qualquer palavra, como *elefante*, evoca um *frame*, que pode ser uma imagem ou outros tipos de conhecimento: elefantes são grandes, têm orelhas frouxas e uma tromba, estão associados a circos, e assim por diante. A palavra é definida em relação a esse *frame.* (LAKOFF, 2004, p. 3)

Dessa forma, o termo *elefante* acarreta a ativação de um ou mais *frames*, que podem ser considerados como estruturas de expectativa relativamente estáveis, partilhadas por uma comunidade e sócio-historicamente situadas (TANNEN; WALLAT, 1987; MIRANDA, 1999).

A perspectiva linguística acerca do *frame* foi estabelecida em trabalhos seminais de Fillmore (1975; 1976), anteriormente ao surgimento da Linguística Cognitiva como área de conhecimento – empreendimento que preconiza uma concepção de linguagem totalmente inter-relacionada às nossas capacidades cognitivas. No entanto, "Ao defender os princípios de não distinção entre conhecimento linguístico e conhecimento enciclopédico e de contextualização cultural da língua, a Semântica de Frames se torna parte integrante do empreendimento cognitivista" (BERTOLDI, 2011, p. 16).

A apropriação do *frame* por Fillmore se entrelaça com sua trajetória de pesquisador – tal qual relatado pelo próprio autor em sua "história privada" a respeito da noção de *frame* semântico (FILLMORE, 1982). Primeiramente, esse conceito é tomado como esquema estritamente linguístico, em uma perspectiva estruturalista, sendo utilizado em operações de comutação, de modo a evidenciar o funcionamento lexical e gramatical de sentenças (FILLMORE, 1975; 1987). Tal concepção se modifica a partir de sua filiação epistemológica ao gerativismo, com a proposição da Gramática de Casos, objetivando construir "[...] uma fórmula para indicar a valência ou os requisitos contextuais de um dado predicador" (FILLMORE, 1975, p. 130), a qual permitiria a verificação de papéis semânticos universais. A partir desse estudo, surge o conceito de *case frame*, definido pelo autor (1982, p. 115) como "[...] uma pequena cena ou 'situação' 'abstrata', de forma que para entender a estrutura semântica do verbo seria necessário entender as propriedades de tais cenas esquematizadas".

O refinamento de sua teoria ocorre quando o pesquisador resolve privilegiar os *frames* em suas análises, transformando os papéis semânticos em categorias subordinadas a essas cenas abstratas (FILLMORE, 2012). Nesse contexto, foi crucial o momento em que o autor passou a se apropriar do conceito de protótipo, advindo dos estudos em psicologia cognitiva acerca do processo de categorização (ROSCH, 1973): os resultados obtidos nessa área de conhecimento, por meio de estudos experimentais com diferentes comunidades, mostraram que o fenômeno

da categorização é mais bem elucidado quando se consideram os efeitos prototípicos que nos levam a agrupar objetos. Por exemplo, o termo *pássaro*, em nosso contexto sociocultural, leva-nos a pensar em exemplares mais prototípicos da categoria – como sabiás ou pardais –, em detrimento de aves menos prototípicas, como o avestruz. No contexto da teoria de Fillmore, a noção de protótipo "explica como o conhecimento [ou o *frame*] ativado por determinada expressão está diretamente ligado à categorização prototípica que fazemos do mundo" (SANTOS, 2016, p. 46).

A partir de sua aproximação com estudos computacionais ligados à recuperação da informação (JURAFSKY, 2014), Fillmore passa a aprimorar seu conceito de *frame* semântico, considerando o modo como o computador deveria associar palavras para que se estabelecesse um processamento similar ao da mente humana – desse modo, computacionalmente, seria necessário associar agrupamentos de palavras relacionadas ao mesmo *frame*, de maneira que fosse possível recuperar automaticamente essas inter-relações. Tais premissas subjazem ao conhecido conceito de *frame* postulado no texto que consolida a Semântica de Frames como programa de pesquisa. Nas palavras do autor,

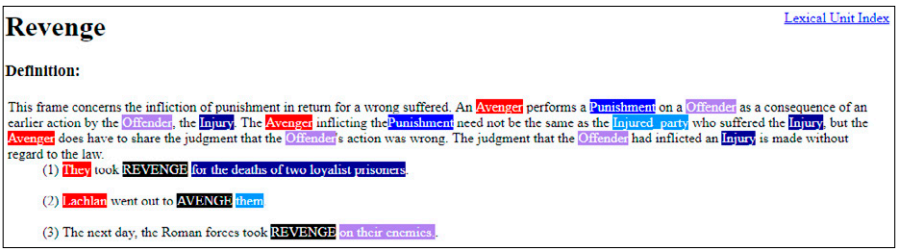
> A Semântica de Frames oferece um modo particular de se olhar para o significado das palavras, e também um modo de caracterizar princípios para criar novas palavras e frases, para adicionar novos sentidos às palavras, e para juntar os sentidos de elementos textuais ao sentido total do texto. Pelo termo *frame* tenho em mente qualquer sistema de conceitos relacionados de tal maneira que para entender qualquer um deles é preciso entender a estrutura que os comporta como um todo; quando um dos itens de tal estrutura é introduzido em um texto ou em uma conversa, todos os outros se tornam automaticamente disponíveis. (FILLMORE, 1982, p. 11)

É importante observar que as bases epistemológicas da Semântica de Frames consolidam uma visão mais abrangente de *frame*, em contraposição às primeiras formulações do autor. A partir de então, essa estrutura de conhecimento é explicitamente associada às categorias de experiência que ancoram os usos que fazemos da língua, revelando o caráter altamente motivado e socialmente ancorado de nossa comunicação verbal.

A aproximação de Fillmore com a lexicografia, juntamente com sua experiência adquirida na área da computação, resulta na concepção da *FrameNet Berkeley*, um recurso lexicográfico computacional baseado em *frames*. A partir dessa concretização de uma base de dados voltada às áreas de Processamento da Linguagem Natural (PLN) e da Inteligência Artificial, bem como à Linguística Computacional, estabelece-se uma metodologia para descrição de *frames*. Nesse contexto, como explicam Fillmore, Johnson e Petruck (2003), as palavras que evocam *frames* passam a ser denominadas *unidades lexicais*, indicando-se um

pareamento entre palavra e significado – ou seja, no contexto da teoria fillmoriana, entre uma palavra e um *frame*.

Apesar de essa metodologia se iniciar por uma caracterização introspectiva do *frame*, segundo o conhecimento dos falantes, essa estrutura conceptual é restrita aos aspectos "[...] para os quais a língua disponibiliza meios expressivos específicos" (FILLMORE; BAKER, 2010, p. 321). Ou seja, o *frame* parte de um conhecimento enciclopédico mais geral e extralinguístico, mas sua descrição, no âmbito da Semântica de Frames, fica reduzida aos fatores linguísticos que descrevem o respectivo cenário. Mais especificamente, como ressalta Ziem (2014), a metodologia da *FrameNet* é orientada pelas valências, as quais descrevem o *frame* a partir das combinatórias sintático-semânticas que envolvem seus evocadores. A título de exemplo, trazemos a seguir o *frame Revenge*, em que os elementos de *frame* são realçados em cores diferentes na definição. Logo abaixo, enumerados, constam três exemplos com os respectivos elementos de *frame* coloridos – portanto, expressos linguisticamente – sendo que a unidade lexical evocadora é realçada na cor preta:



Revenge

Lexical Unit Index

Definition:

This frame concerns the infliction of punishment in return for a wrong suffered. An Avenger performs a Punishment on a Offender as a consequence of an earlier action by the Offender, the Injury. The Avenger inflicting thePunishment need not be the same as the Injured_party who suffered the Injury, but the Avenger does have to share the judgment that the Offender's action was wrong. The judgment that the Offender had inflicted an Injury is made without regard to the law.

(1) They took REVENGE for the deaths of two loyalist prisoners.

(2) Lachlan went out to AVENGE them.

(3) The next day, the Roman forces took REVENGE on their enemies.

Figura 1 – *Frame Revenge* descrito segundo as convenções da *FrameNet Berkeley*
Fonte: *FrameNet Berkeley*. Disponível em: <framenet.icsi.berkeley.edu>

Visto que nosso objetivo de análise não estava centrado na verificação de valências relativas ao *frame feto anencéfalo*, optamos por uma abordagem mais abrangente que nos permitisse descrever as diferentes perspectivas acerca dessa entidade, de modo a sistematizar os diferentes atributos direcionados a esse *frame* a partir de uma análise de *corpus*. Para isso, optamos por uma proposta mais abrangente, cujo foco é a identificação de *frames* de compreensão. Esse desdobramento é abordado a seguir.

2.1 Frames de compreensão

Os estudos semânticos de Ziem (2014) pautam-se na proposta fillmoriana de semântica da compreensão (FILLMORE, 1985), a qual "tem por objetivo explicar todas as facetas de conhecimento necessárias para se compreender inteiramente o significado de uma expressão linguística" (ZIEM, 2014, p. 2). Nesse trabalho,

Fillmore busca contrapor a sua proposta a abordagens vericondicionais e intuitivas do significado, preconizando um programa empírico de pesquisa, voltado a contextos comunicativos reais. Importa ressaltar que tal perspectiva, baseada em dados autênticos, também é incorporada pelo projeto da *FrameNet*; no entanto, conforme indicado na seção anterior, em virtude de sua faceta computacional, a descrição valencial ganha destaque na plataforma, em detrimento da perspectiva voltada às facetas de conhecimento postulada no texto de 1985.

Assim, nesse enquadramento teórico retomado por Ziem, os *frames* de compreensão podem ser entendidos tanto como estruturas cognitivas que organizam nossa experiência quanto como ferramenta analítica para se identificarem esses *frames* a partir de evidências empíricas (FILLMORE, 1985). Visto que a Semântica da Compreensão evidencia não haver separação entre conhecimento linguístico e conhecimento de mundo, em virtude da visão enciclopédica de significado, Ziem considera que expressões linguísticas estão sempre inseridas em um "espaço de compreensão" estruturado por meio desses *frames*. Nessa apropriação dos conceitos fillmorianos, é importante observar que o autor também se vale de estudos desenvolvidos no contexto de pesquisa alemão, nomeadamente os trabalhos de Konerding (1993), Lönneker (2003) e Fraas (1996).

Para propósitos analíticos, com base em estudos como os de Minsky (1974) e Coulson (2001), o autor leva em conta a estrutura dos *slots* e *fillers* como constituintes que embasam a organização esquemática dos *frames*. Os *slots* são realizados linguisticamente por determinadas expressões (*fillers*); por exemplo, em uma frase como *a piscina mede 2 metros*, temos um *slot* de *medida* relativo ao termo *piscina*, cujo *filler* é *mede dois metros*. Assim, cada *filler* revela uma faceta de conhecimento acerca do *frame* ativado pelos falantes.

É importante observar que o conceito de *slot* também se relaciona, nos trabalhos de Fillmore (1976), à ideia de *frame* como estrutura que requer um "preenchimento de detalhes" por parte do falante ao longo da interação. Nas palavras do autor, compreender o significado de uma palavra "requer conhecer o cenário; entender uma sentença contendo esse mundo requer conhecer o cenário e usar o conteúdo lexical e a estrutura gramatical do resto da sentença para preencher alguns detalhes [...]" (FILLMORE, 1976, p. 28). Assim, a estrutura *slot-filler* é apropriada para indicar que *frames* não trazem todas as facetas de conhecimento possíveis, mas sim aquelas que são adequadas aos propósitos dos interlocutores.

Considerando que, em seus textos seminais, Fillmore não traz muitas pistas quanto à verificação empírica de *slots*, Ziem vale-se dos estudos de Lönneker (2003), os quais propõem que *frames* podem ser identificados por meio de proposições, que se referem, nos termos de Searle (1969), à "dimensão do conteúdo da frase que pertence ao conteúdo proposicional da sentença, independentemente de seu modo de expressão" (ZIEM, 2014, p. 245). Desse modo, uma proposição diz algo sobre uma entidade em particular – o objeto de referência, sobre o qual

se atribui uma predicação. Assim, predicação significa "alocação de predicados a um objeto de referência" (ZIEM, 2014, p. 246). Nesse caso, não se trata de uma referência a um mundo objetivo, tal qual defendido por abordagens do significado baseadas em condições de verdade, mas sim de um mundo reconhecido e constantemente reconstruído pelos falantes por meio de processos sociocognitivos – esse é o caso das construções que perspectivizam *feto anencéfalo* presentes em nosso *corpus*, visto que não está em questão a busca de uma verdade absoluta relativa ao estatuto do feto, mas sim o jogo de diferentes matizes que constroem, de forma heterogênea, esse *frame* no contexto em estudo.

De forma a se verificar empiricamente as predicações que indicam a estrutura de um *frame*, é necessário partir de *corpora* autênticos. A subseção a seguir aborda as etapas para identificação dos *frames* de compreensão a partir de tais predicações.

2.1.1 Como identificar frames de compreensão em *corpora*?

Segundo essa perspectiva, o trabalho de identificação de *frames* leva em conta o pressuposto de que um *corpus* autêntico, selecionado de acordo com os propósitos de investigação, traz uma série de predicações quase explícitas sobre o *frame* a ser observado, as quais devem ser explicitadas pelo analista (ZIEM, 2014). O ponto de partida para a busca dessas predicações são os evocadores do *frame* em análise, a partir dos quais é possível verificar como a entidade em questão – no caso deste trabalho, o *feto anencefálico* – é predicada, ou seja, quais são os atributos recorrentes no *corpus*.

Assim, os procedimentos metodológicos para identificação de *frames* de compreensão podem ser elencados da seguinte forma:

a) Identificação dos evocadores do *frame*;

b) Coleta de todas as concordâncias nas quais esses evocadores ocorrem, recuperando todo o período correspondente;

c) Explicitação das predicações (quase) explícitas (posicionamento evocador como sujeito e explicitação das predicações);

d) Agrupamento das predicações: organização das predicações semelhantes e nomeação do *slot* que as abrange;

e) Verificação da frequência de cada predicação, de modo a determinarmos os *slots* mais e menos emergentes em cada *subcorpus*.

Para esclarecer o funcionamento desse processo descritivo, trazemos um exemplo adaptado de uma investigação relatada por Ziem (2014), relativa à descrição do *frame* investidor financeiro em um *corpus* midiático.

Primeiramente, os possíveis evocadores de *frame* foram elencados introspectivamente pelo autor – no caso do *frame* investidor financeiro, o único evocador considerado foi a própria expressão *investidor financeiro*, acrescida de seu plural,

"bem como [de] outras formas inflexionais ou derivativas" (ZIEM, 2014, p. 350). A manchete extraída do *corpus* pelo autor, a título de ilustração de sua metodologia, trazia o seguinte conteúdo:

Ações são compradas por investidores financeiros após quedas recentes em Wall Street.

Na segunda etapa, foram coletadas todas as demais concordâncias que possuíam esse evocador de *frame*. Na terceira etapa, visto que o objetivo da análise era descrever o *frame* "investidor financeiro", o termo foi deslocado para a posição de sujeito e a frase foi reformulada, de modo a explicitar a predicação atribuída a ele – processo que resulta em predicações explícitas, as quais constituem *fillers* do *frame* em estudo. A predicação explícita encontrada foi a seguinte:

Investidores financeiros compram ações.

Além dessa ocorrência, outros atributos similares foram elencados, os quais incluíam atividades como *financiar empresas*, *gerenciar carreiras* e *administrar carteiras*. Trata-se de *fillers* a serem qualitativamente agrupados em *slots* – procedimento que consiste na quarta etapa analítica. Como explica Ziem (2014, p. 361), "[...] predicações explícitas devem ser manualmente e interpretativamente classificadas. Isso deve ser definido caso a caso de forma a cumprir com os critérios específicos e produzir resultados replicáveis e confiáveis". Nesse caso, Ziem (2014) agrupou as predicações citadas a partir do *slot* [função profissional].

A seguir, esquematizamos esse exemplo adaptado, a partir da ilustração elaborada em Santos (2016), de modo a explicar melhor o modo de agrupamento de *fillers* em *slots* por meio da explicitação de predicações:

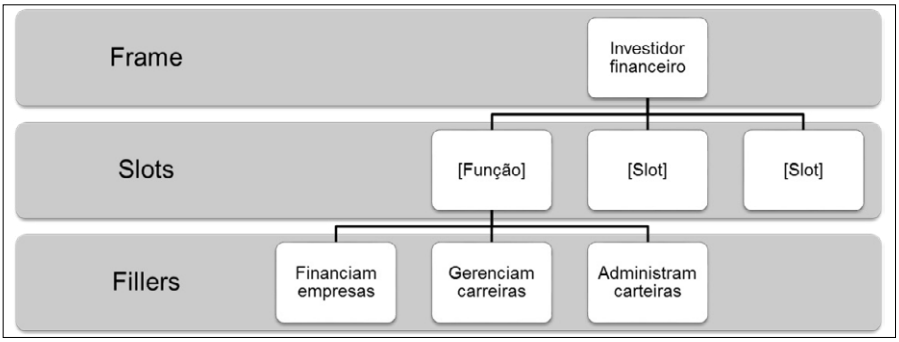


Figura 2 – Frame investidor financeiro
Fonte: Santos (2016)

A quinta e última etapa consiste na verificação da frequência de cada predicação, visto que o *corpus* pode reiterar alguns *fillers*, o que reforça a proeminência dada a uma faceta de conhecimento em detrimento de outra. Tal aspecto evidencia a relação direta existente entre o conceito de *frame* e a operação de perfilamento (LANGACKER, 1987; 2008): trata-se de uma operação cognitiva de saliência que permite a ênfase de determinado aspecto de uma cena (o *perfil*) em relação ao cenário como um todo (a *base*). Um exemplo didático fornecido por Langacker (2008), ao explicar esse conceito, remete à dinâmica de nossa divisão do tempo em anos e meses: o ciclo temporal de doze meses corresponde à base, sendo que cada mês corresponde a uma porção específica desse período. O mesmo ocorre no caso da forma geométrica círculo: tendo-a como base, é possível salientar ou perfilar suas diferentes partes, ou seja, é possível atentar para seu diâmetro, seu raio ou sua circunferência.

Dessa forma, visto que os perfilamentos resultam na expressão das diferentes facetas potencialmente evocadas por uma palavra (CROFT; CRUSE, 2004), é importante considerar que cada *filler* correspondente a um *frame* resulta em uma faceta de conhecimento específica, a qual enfatiza algum aspecto, em detrimento de outros, da entidade que está sendo conceptualizada. Por esse motivo, Ziem (2014, p. 285) enfatiza:

> Com a ajuda de predicações explícitas, usuários da língua perfilam certas facetas do conhecimento sobre um objeto de referência enquanto outras se alternam no *background*. Toda predicação perspectiviza o objeto de referência de uma forma particular. É impossível usar signos linguísticos sem adotar uma perspectiva refratada dos objetos de referência.

Ao encontro disso, Croft e Cruse (2004, p. 18) consideram que *frames* semânticos são uma abordagem pertinente para "descrever diferenças que parecem ser definidas em bases sociais em vez de conceptuais. Mas há uma ponte entre elas. Comunidades são definidas pelas atividades sociais que mantêm os membros unidos".

Realizado este panorama acerca dos pressupostos teóricos que direcionaram nossa investigação, reiteramos que este trabalho tem como foco a primeira etapa analítica de identificação dos evocadores, na qual o uso da ferramenta Word Sketch mostrou-se crucial. Além disso, a partir dessa exploração inicial, indicamos como o uso desse recurso trouxe alguns resultados preliminares relativos às etapas de análise posteriores. Esses aspectos são detalhados na seção a seguir.

3 O *corpus* de estudo e a ferramenta Sketch Engine

Nesta seção, contextualizamos nosso *corpus* de estudo e detalhamos sua divisão em *subcorpora*, de modo a estabelecer comparações entre as conceptualizações encontradas nos discursos religioso, médico e jurídico, os quais permearam as discussões acerca do estatuto do feto anencéfalo ao longo do processo da ADPF 54. Em seguida, abordamos os recursos do Sketch Engine utilizados nas etapas de análise.

3.1 Características do corpus de estudo

A análise empreendida neste trabalho consiste em um *corpus* proveniente do processo da Arguição de Preceito Fundamental 54-8 (ADPF 54). Tal material é composto pelo acórdão da ADPF 54 e pelas notas taquigráficas que registram os depoimentos das quatro audiências públicas realizadas. Conforme referido anteriormente, o foco da análise foi a construção do *frame feto anencéfalo*, a partir das diferentes vozes que debateram sua condição biológica e social ao longo do processo.

Em relação ao documento acórdão, trata-se de um "julgamento proferido pelos tribunais" (BRASIL, 1973) que objetiva proferir uma sentença cujos emissores são o Poder Judiciário. No caso do acórdão da ADPF 54, o documento constitui-se de uma decisão final referente à petição em favor da antecipação terapêutica de parto, indicando os votos individuais dos ministros acerca da matéria em debate. É importante observar que essa decisão consiste em um acórdão inteiro teor, dado que todas as fases do processo são descritas pelo respectivo relator, dispensando-se a consulta a acórdãos anteriores.

Quanto à transcrição das notas taquigráficas, trata-se do registro de quatro audiências públicas que ocorreram para registrar os posicionamentos de autoridades e cidadãos interessados no processo. Em três audiências, representantes de instituições médicas foram ouvidos; e, em uma audiência pública, representantes religiosos foram convidados a se manifestar.

Assim, para fins de comparação entre diferentes facetas de conhecimento, o *corpus* foi dividido em três *subcorpora*: (i) as transcrições das notas taquigráficas da primeira audiência pública, na qual predominam os posicionamentos de representantes de instituições religiosas (*Corpus* NT1); (ii) as transcrições das notas taquigráficas das três audiências públicas seguintes, em que predominam os posicionamentos de representantes de entidades médicas (*Corpus* NT2); e (iii) o acórdão de inteiro teor, em que consta a votação dos ministros e o proferimento da decisão final (*Corpus* Acórdão).

Esses documentos, disponíveis em *.pdf, foram convertidos para formato *.doc e posteriormente para *.txt, de modo a serem processados pela ferramenta Sketch Engine. O *corpus* compilado possui pouco mais de 158.000 *tokens*. É importante observar que, como se trata de um estudo de caso que reúne toda a documentação atinente ao processo da ADPF 54 e como os *subcorpora* foram divididos conforme a identidade dos participantes, não há balanceamento entre esses três segmentos, visto que os representantes religiosos foram ouvidos apenas em uma audiência pública e que o acórdão é muito mais extenso que as transcrições das notas taquigráficas. Tais fatores levaram-nos a considerar as concordâncias independentemente de sua frequência, de modo a explorar exaustivamente as conceptualizações presentes nesses dados.

3.2 O Sketch Engine

Realizada a etapa de compilação do *corpus*, escolhemos o Sketch Engine como ferramenta para extração dos possíveis evocadores do *frame feto anencéfalo*, bem como das concordâncias que indicavam as predicações relativas a essa entidade. A ferramenta, que já vem sendo utilizada pelo grupo SemanTec em seus empreendimentos cognitivo-lexicográficos (CHISHMAN et al., 2014; 2015), dispensa quaisquer instalações, pois trata-se de um recurso *on-line* que exige apenas a inserção de *login* e senha correspondentes.

É importante pontuar que se trata de um *software* pago e que tínhamos acesso às licenças utilizadas por nosso grupo de pesquisa; no entanto, o Sketch Engine permite que se faça um registro gratuito, válido por um mês, a partir do qual as suas principais ferramentas são disponibilizadas.

Dentre esses recursos, está a Word Sketch, uma tabela disponibilizada pelo Sketch Engine que elenca todas as combinatórias sintáticas da palavra pesquisada, sistematizando seu comportamento gramatical e colocacional. Cada informação concernente a essas relações gramaticais vem acompanhada da frequência de concordâncias correspondentes, permitindo acesso imediato aos contextos nos quais essas combinatórias ocorrem. A seguir, esse funcionamento será explorado e ilustrado a partir de exemplos extraídos de nosso *corpus* de estudo.

4 O uso do Sketch Engine na identificação dos *frames* de compreensão

Nesta seção, objetivamos ilustrar o uso do Sketch Engine para verificação dos seguintes aspectos: a) possíveis evocadores para o frame *feto anencéfalo*; b) facetas de conhecimento enfatizadas pelos falantes a partir do estudo das concordâncias; c) existência de uma relação direta entre algumas escolhas lexicais e determinadas

facetas de conhecimento; e d) relação entre usos linguísticos e identidades predominantes em cada *subcorpus*.

4.1 Seleção de evocadores

Conforme suprarreferido, o primeiro passo da análise de *frames* de compreensão consiste na verificação de possíveis unidades lexicais que possam evocar o *frame* em estudo, seguindo a proposta de Ziem (2014). Nessa etapa, o autor propõe um levantamento introspectivo desses evocadores, a partir do conhecimento de mundo do analista. Assim, a questão que guia esse procedimento é a seguinte: que unidades linguísticas servem como ponto de partida para explorarmos o *frame* em análise?

Ao fazer as primeiras explorações do *corpus*, verificamos que o *frame feto anencéfalo* não era apenas evocado por essa expressão – por exemplo, participantes contrários à ADPF também se referiam ao feto como *criança*. Levando em conta essa necessidade de coletar os diversos evocadores que constavam no *corpus*, o uso da ferramenta Word Sketchfoi crucial para o levantamento de tais informações.

Assim, nesse contexto de pesquisa, ao efetuarmos a primeira busca por meio da palavra *feto*, encontramos combinatórias como *feto anencéfalo* e *feto anencefálico*, indicando que deveríamos buscar pelas concordâncias a partir de ambos os termos, dentre outras ocorrências. A imagem a seguir mostra as *word sketches* para *feto* nos três *subcorpora*, mostrando que, além de ocorrências com os evocadores *feto anencéfalo* e *feto anencefálico*, também deveriam ser coletadas as concordâncias que contivessem o termo *feto portador de anencefalia*:

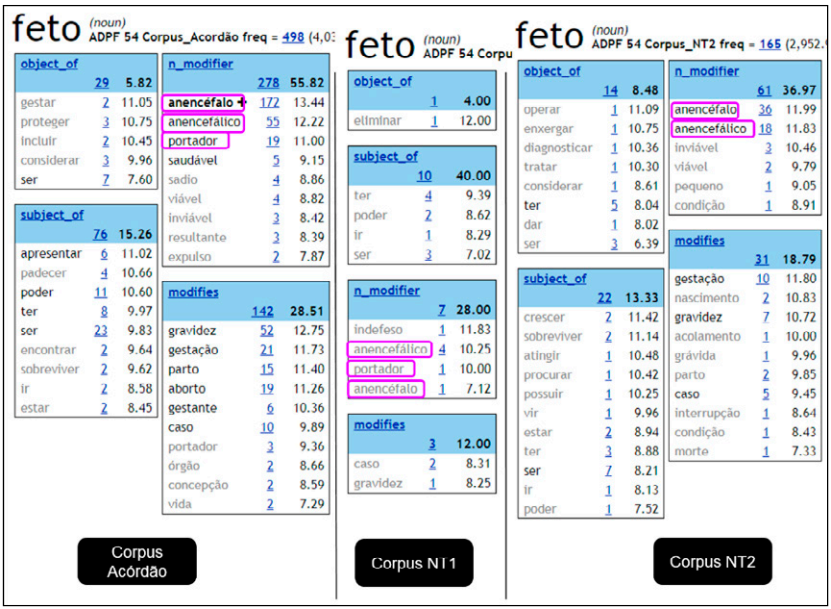


Figura 3 – *Word sketches* para *feto* (*Corpora* Acórdão, NT1 e NT2)
Fonte: Registrada pelas autoras a partir do Sketch Engine

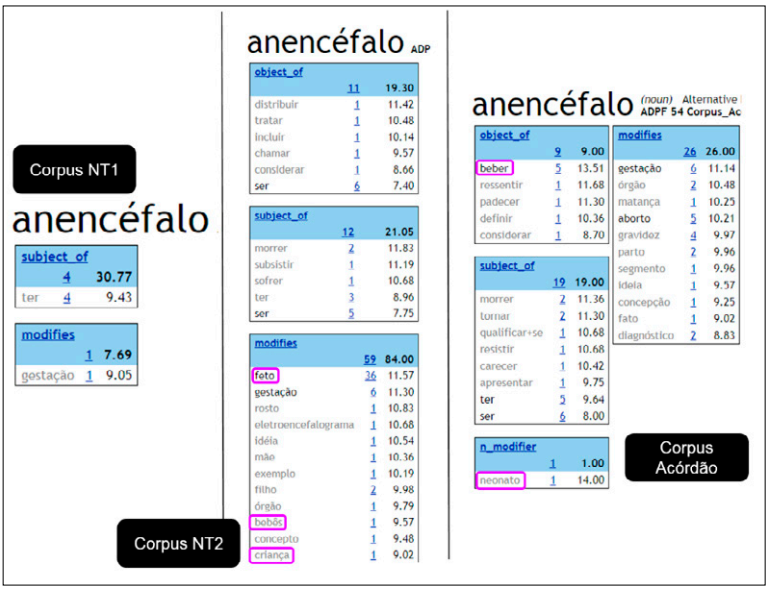


Figura 4 – *Word sketches* para *anencéfalo* (*Corpora* Acórdão, NT1 e NT2)
Fonte: Registrada pelas autoras a partir do Sketch Engine

Nas *word sketches* da palavra *anencéfalo*, conforme mostra a Figura 3, foi possível identificar ocorrências menos frequentes mas igualmente relevantes, como *criança anencéfala* e *neonato anencéfalo*.

Além disso, nessa mesma imagem, observamos a ocorrência de *bebê anencéfalo*, que aparece na coluna de verbos como *beber*, em virtude de um erro proveniente do etiquetador morfossintático disponibilizado pelo Sketch Engine, o Freeling. Tal aspecto evidencia a pertinência de se verificar todas as entradas da Word Sketch de forma qualitativa, examinando as concordâncias correspondentes a cada função sintática.

Finalizada essa primeira etapa, quanto aos evocadores para o *frame feto anencéfalo* acessados a partir das *word sketches* de *feto* e de *anencéfalo*, identificamos os seguintes termos:

Anencéfalo
Feto anencefálico
Feto anencéfalo
Feto portador de anencefalia
Neonato anencéfalo
Criança anencéfala
Bebê anencéfalo

A partir da seleção de evocadores, iniciamos os procedimentos de coleta, que implicaram uma análise de cada concordância para verificar se o trecho trazia predicações referentes ao *frame*. Por exemplo, muitas ocorrências de *feto anencéfalo* estão ligadas a evocadores como *antecipação terapêutica de parto de feto anencéfalo*. Nesses casos, as predicações voltavam-se ao *frame* de interrupção de gravidez. Além disso, considerando a estrutura textual, por vezes foi necessário recuperar anáforas para identificá-las como evocadores – por exemplo, ao procurar pelo evocador *criança com anencefalia* no concordanciador, encontramos "essas crianças" e, ao verificar que se tratava de uma menção a uma criança que seria anencéfala, incluímos o trecho em nossa lista, de modo a cumprir com as etapas analíticas posteriores.

Vale também ressaltar que os resultados da Word Sketch implicaram descarte de concordâncias em alguns casos. Por exemplo, ao consultarmos o termo *pessoa anencéfala*, vimos que pertencia ao trecho "não há pessoas anencéfalas no mundo". Desse modo, não incluímos o termo em nossa lista de evocadores, pois não se tratava propriamente de uma descrição relativa a fetos anencefálicos.

4.2 Verificação de facetas de conhecimento a partir das concordâncias

Segundo a metodologia proposta por Ziem (2014), a etapa de verificação de facetas de conhecimento, ou de *fillers* que preenchem os *slots* de *frames* de compreensão, seria feita apenas após a coleta de todas as concordâncias e a descrição qualitativa de predicações. No entanto, o uso da Word Sketch na primeira etapa indicou que já era possível verificar a predominância de algumas facetas de conhecimento em cada *subcorpora*, as quais poderiam estar vinculadas a certas escolhas lexicais. Dessa forma, após fazer a busca por evocadores, clicamos nos *links* das *word sketches* para efetuarmos uma pré-análise das concordâncias correspondentes, visando a responder à pergunta: em que medida o uso de alguns termos já evidenciaria, apenas por meio dessa consulta prévia, as facetas de conhecimento enfatizadas pelos falantes?

Tal exploração nos mostrou que, quanto ao uso do termo feto anencéfalo/anencefálico no *corpus*, essas expressões estavam mais ligadas às facetas de anomalia e/ou morte do feto nos *subcorpora* NT2 e Acórdão:

Quadro 1 – Exemplos de concordâncias com o evocador *feto anencéfalo* (*Corpora* Acórdão e NT2)

O ***feto anencéfalo*** é, hoje, tido pela medicina como inviável
Conforme demonstrado, o ***feto anencéfalo*** não tem potencialidade de vida.
o ***feto anencéfalo***, se chega a nascer, tem mínima sobrevida, e sequer apresenta capacidades além das fisiológicas
o ***feto anencéfalo*** é, até o estágio atual da medicina, irremediavelmente inviável para a vida extrauterina
O ***feto anencéfalo*** não passa de um organismo prometido à inscrição do seu nome não no registro civil, mas numa lápide mortuária.
O ***feto anencéfalo*** mostra-se gravemente deficiente no plano neurológico.
O ***feto anencéfalo*** é um natimorto cerebral.

Fonte: Elaborado pelas autoras a partir do Sketch Engine

No entanto, essa faceta não consta no *corpus* NT1, das entidades religiosas: há pouquíssimas ocorrências, nas quais se evoca uma faceta de potencialidade de vida do anencéfalo, conforme os exemplos a seguir:

Quadro 2 – Exemplos de concordâncias com o evocador *feto anencéfalo* (*Corpus* NT1)

Não é porque o ***feto anencéfalo*** não tem essas áreas nobres do córtex cerebral - geralmente ele não as tem – que ele não tem consciência.
Assim, o ***feto anencefálico*** é um ser humano vivente
Em suma, nesse tronco encefálico alto, que todo ***feto anencefálico*** tem, nós contemplamos as bases dos mecanismos neurais da respiração

Fonte: Elaborado pelas autoras a partir do Sketch Engine

4.3 Relação entre escolhas lexicais e facetas de conhecimento

A pertinência de tal exploração com o uso do Sketch Engine levou-nos a realizar mais uma investigação prévia à descrição sistemática de *frames*, no intuito de responder a esta questão: o uso por si só dos termos *criança* e *bebê* já evidenciava facetas relativas à vida do anencéfalo? Tal questão foi motivada principalmente por trabalhos como os de Lakoff (2004) e Kövecses (2006), os quais afirmam que o uso dessas expressões sempre ancora o posicionamento dos falantes em *frames* como *ciclo de vida humana*, projetando a vida do feto para um estágio avançado.

No entanto, nos *subcorpora* da ADPF 54, tal premissa não foi totalmente confirmada: o uso de *bebê* nem sempre indicava uma faceta relativa à expectativa de vida do feto, conforme exemplos a seguir. Ressaltamos que a frequência dessa faceta de conhecimento é baixa e ocorre apenas nos *subcorpora* Acórdão e NT2. Ainda assim, tal resultado se mostra interessante para evidenciar como a verificação de *frames* não pode depender apenas de uma análise lexical, mas deve levar em conta os contextos de ocorrência desses termos, de modo a se identificar qual perspectiva acerca da entidade é ativada pelos falantes.

Quadro 3 – Exemplos de concordâncias com o evocador *bebê* – faceta de conhecimento de anomalia e morte (*Corpora* Acórdão e NT2)

Um ***bebê*** anencéfalo é geralmente cego, surdo, inconsciente e incapaz de sentir dor.
na realidade, o ***bebê*** era inviável, que ele era um natimorto, por assim dizer

Fonte: Elaborado pelas autoras a partir do Sketch Engine

O mesmo não se pode dizer quanto às ocorrências de *criança*: esse uso, que ocorre apenas nos *subcorpora* NT1 e NT2, sempre se compromete com facetas relativas à vida do anencéfalo, conforme exemplos a seguir:

Quadro 4 – Exemplos de concordâncias com o evocador *criança* (*Corpus* NT1)

Aqui está a ***criança*** portadora de anencefalia com seus pais. Seus pais referem interação com essa criança
Estudos devem ser feitos para determinar o real estado de consciência dessas ***crianças*** portadoras de anencefalia.

Fonte: Elaborado pelas autoras a partir do Sketch Engine

Quadro 5 – Exemplos de concordâncias com o evocador *criança* (*Corpus* NT2)

A ***criança anencefálica*** – vejam bem - não causa perigo à vida da sua mãe mais do que uma gestação gemelar.
Portanto, não se justifica como aborto terapêutico o que se pretende fazer com a ***criança anencefálica***.

Fonte: Elaborado pelas autoras a partir do Sketch Engine

4.4 Relação entre ocorrências lexicais e identidades dos participantes

Ao longo das explorações realizadas por meio da Word Sketch, também foi possível verificar como o léxico de cada *subcorpus* revelava seus aspectos identitários – ou seja, como indicava algumas peculiaridades do discurso religioso e contrário à ADPF; do discurso médico e do discurso jurídico – no caso desses últimos, é interessante notar as marcas linguísticas provenientes da terminologia dessas áreas, as quais consistem em domínios altamente especializados.

Nesse ínterim, o *corpus* NT1 foi o único que apresentou construções como *eliminar* (o feto com anencefalia), bem como designações de anencefálicos como *indefesos* e *inocentes* – fator que permitiu percebermos, previamente à etapa de construção de *frames*, como as perspectivas acerca do feto, nesse *subcorpus*, eram construídas em prol de um discurso totalmente antagônico em relação à proposta da ADPF.

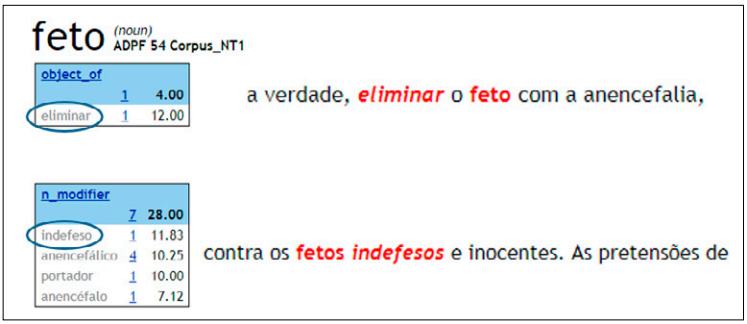


Figura 5 – Exemplos do léxico do *Corpus* NT1 (Word Sketch e respectivas concordâncias)
Fonte: Elaborado pelas autoras a partir do Sketch Engine

Já o *Corpus* NT2 se particulariza pela incidência de termos típicos da área médica, incluindo construções como *operar*, *tratar* e *diagnosticar*, referindo-se às condições neurológicas do feto anencéfalo ou ao estatuto do distúrbio da anencefalia. Tal aspecto indica o ponto de vista a partir do qual a comunidade médica, segundo os participantes que se manifestaram nas audiências públicas, manifesta seus posicionamentos:

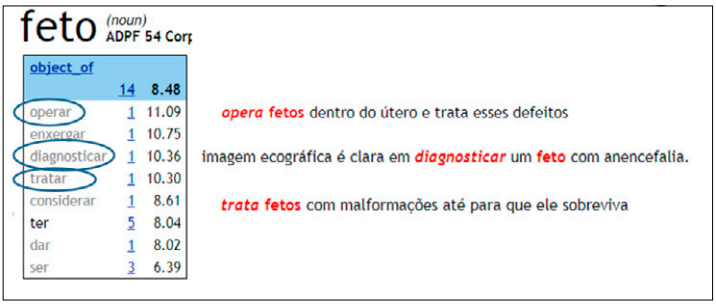


Figura 6 – Exemplos do léxico do *Corpus* NT2 (Word Sketch e respectivas concordâncias)
Fonte: Elaborado pelas autoras a partir do Sketch Engine

Por sua vez, o *Corpus* Acórdão apresentou elementos típicos da terminologia jurídica, indicando que os votos dos ministros pautaram-se consideravelmente no debate acerca do estatuto jurídico do anencéfalo em relação a fetos sem anencefalia, utilizando-se termos como *proteger*, *incluir* e *considerar*:

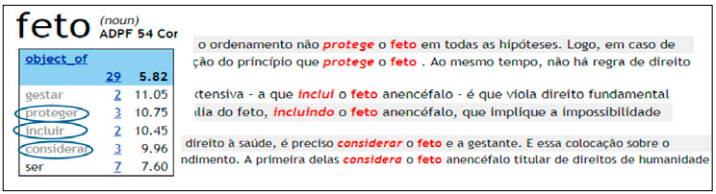


Figura 7 – Exemplos do léxico do *Corpus* Acórdão (Word Sketch e respectivas concordâncias)
Fonte: Elaborado pelas autoras a partir do Sketch Engine

Observamos que, nesse caso, foi necessário acessar as concordâncias e consultar fontes jurídicas para compreender suficientemente o mote desses debates, dado que se trata de termos característicos da área, cujo significado difere de usos em língua geral. Por exemplo, o verbo *proteger* está relacionado especificamente à proteção jurídica do feto por meio de princípios constitucionais. Da mesma forma, *incluir* e *considerar* são usados em relação à possibilidade de enquadramento do feto – e, por vezes, da gestante – no grupo de cidadãos cujos direitos constitucionais, como direito à saúde ou direitos de humanidade, devem ser assegurados.

5 Considerações finais

Este trabalho visou a relatar as vantagens no uso do Sketch Engine na identificação de evocadores de *frames* e de suas respectivas facetas de conhecimento, a partir de um estudo de caso do processo da ADPF 54, tendo como foco o *frame* feto anencéfalo. Para isso, contextualizamos o conceito de *frame* no âmbito da semântica cognitiva, abordamos as etapas de identificação de *frames* de compreensão postuladas por Ziem (2014) e mostramos como o uso do Sketch Engine permitiu uma exploração sistemática não apenas dos evocadores de *frame*, mas também das principais facetas de conhecimento e dos recursos lexicais predominantes em cada *subcorpus*.

Além disso, é importante ressaltar a pertinência de uma análise qualitativa das concordâncias, visto que, principalmente na etapa de identificação dos evocadores de *frames*, foi importante recuperar algumas anáforas que também eram relevantes à descrição de *frames*. Da mesma forma, a observação das concordâncias classificadas pelas *word sketches* permitiu o descarte de ocorrências que não estavam diretamente relacionadas à entidade *feto anencéfalo*.

De modo geral, o uso do Sketch Engine resultou em procedimentos metodológicos mais sistemáticos e baseados nos dados, se comparados à proposta de Ziem (2014). Retomando as indicações do autor, explicitadas na seção 2.1, o uso do *corpus* seria posterior ao levantamento de evocadores, o qual partiria da introspecção do analista. A partir desse levantamento introspectivo, o *corpus* seria

então explorado por meio do concordanciador, de modo a se partir para a etapa de classificação qualitativa de facetas de conhecimento:

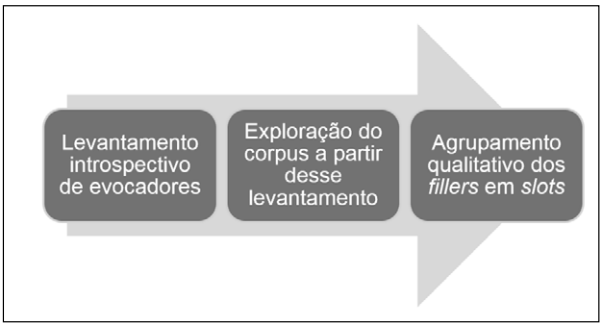


Figura 8 – Proposta inicial de identificação de *frames* de compreensão
Fonte: Elaborado pelas autoras a partir de Ziem (2014)

No contexto deste estudo de caso e considerando a pertinência do recurso Word Sketch e do concordanciador, propusemos uma complementação das etapas iniciais de identificação dos *frames*, incluindo os seguintes procedimentos: (a) levantamento de possíveis evocadores a partir do *corpus*; e (b) análise prévia de algumas facetas de conhecimento, explorando-se os evocadores por meio das *word sketches*. A imagem a seguir visa a sistematizar essas complementações:

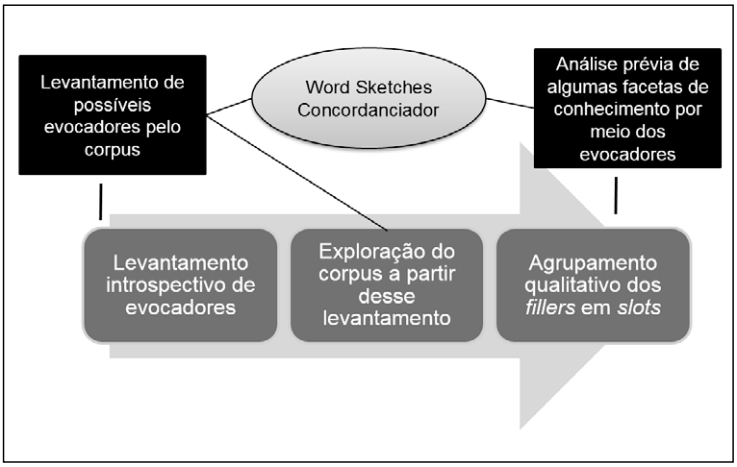


Figura 9 – Complementação da proposta inicial de identificação de *frames* de compreensão
Fonte: Elaborado pelas autoras

Por fim, consideramos que ainda é necessário maior aprofundamento, em estudos futuros, da sistematização metodológica relativa à identificação de evocadores: em nosso trabalho, a ferramenta Word Sketch foi fundamental para que pudéssemos coletar o maior número possível de evocadores; contudo, é possível que termos menos óbvios tenham sido ignorados. Em vista disso, consideramos que o processamento de uma lista de palavras do *corpus* possa ser pertinente a estudos posteriores, de modo que se possa examinar, de forma mais exaustiva, os evocadores potenciais do *frame* analisado.

Referências

BARTLETT, F. *Remembering*: a study in Experimental and Social Psychology. Cambridge: Cambridge University Press, 1932.

BERTOLDI, A. *Semântica de framese recursos lexicais jurídicos:* um estudo contrastivo. 2011. 136 f. Tese (doutorado em Linguística Aplicada). Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos, São Leopoldo, 2011. Disponível em: <http://www.repositorio.jesuita.org.br/bitstream/handle/UNISINOS/4718/AndersonBertoldiLinguistica.pdf>. Acesso em: 15 out. 2017.

BRASIL. *Lei n. 5.869*, de 11 de janeiro de 1973. Institui o Código de Processo Civil. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l5869.htm>. Acesso em: 10 jun. 2014.

CHISHMAN, R. L. de O. et al. Field – Dicionário de expressões do futebol: um recurso lexicográfico baseado no aporte teórico-metodológico da Semântica de *Frames*. *Signo*, Santa Cruz do Sul, v. 39, n. 67, p. 25-35, 2014.

CHISHMAN, R. L. de O. et al. The relevance of the Sketch Engine software to build Field – Football Expressions Dictionary. *RELIN*, Belo Horizonte, v. 23, Edição Especial, p. 769-796, 2015. Disponível em: <http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/8918/8796>. Acesso em: 09 out. 2017.

COULSON, S. *Semantic leaps*. Frame-shifting and conceptual blending in meaning construction. New York: Cambridge University Press, 2001.

CROFT, W.; CRUSE, D. A. *Cognitive Linguistics*. Cambridge: Cambridge University Press, 2004.

DINIZ, D.; VÉLEZ, A. C. G. Aborto na Suprema Corte: o caso da anencefalia no Brasil. *Revista Estudos Feministas*, Florianópolis, v. 16, n. 2, p. 647-652, 2008. Disponível em: <https://periodicos.ufsc.br/index.php/ref/article/view/S0104-026X2008000200019/8797>. Acesso em: 04 nov. 2015.

FILLMORE, C. J. An alternative to checklist theories of meaning. In: COGEN, C. et al. (Ed.). *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society, 1975, p. 123-31.

_____. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, v. 280, p. 20-32, 1976.

_____. Frame Semantics. In: _____. *Linguistics in the Morning Calm*. Seoul: Hansinh Publishing Co., 1982.

_____. Frames and the semantics of understanding. *Quaderni di Semantica*, v. 6, n. 2, 1985, p. 222-254.

_____. A private history of the concept 'Frame'. In: DIRVEN, R.; RADDEN, G. (Ed.). *Concepts of Case.* Tübingen: Narr, 1987.

_____. Encounters with language. *Computational Linguistics*, v. 38, n. 4, p. 701-718, 2012. Disponível em: <http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00129>. Acesso em: 15 set. 2017.

FILLMORE, C. J.; BAKER, C. A frames approach to semantic analysis. In: HEINE, B.; NARROG, H. (Ed.). *The Oxford Handbook of Linguistic Analysis*. New York: Oxford University Press, 2010, p. 313-339.

FILLMORE, C. J.; JOHNSON, C.; PETRUCK, M. R. L. Background to *FrameNet*. *International Journal of Lexicography*, Oxford, v. 16, n. 3, p. 235-250, 2003. Disponível em: <ijl.oxfordjournals.org/content/16/3/235.full.pdf>. Acesso em: 20 out. 2015.

FRAAS, C. *Gebrauchswandel und Bedeutungsvarianz in Textnetzen*. Die Konzepte "Identität" und "Deutsche" im Diskurszurdeutschen Einheit. Tübingen: Narr, 1996.

JURAFSKY, D. Charles Fillmore. *Computational Linguistics*, Cambridge, v. 40, n. 3, p. 725-731, 2014. Disponível em: <http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00201>. Acesso em: 20 set. 2017.

LANGACKER, R. W. *Foundations of cognitive grammar*: theoretical prerequisites. Stanford: Stanford University Press, 1987.

LÖNNEKER, B. *Konzeptframes und Relationen*. Extraktion, Annotation und Analyse französischer *Corpora* ausdem *World Wide Web*. Berlin: Aka, 2003.

GOFFMAN, E. *Frame Analysis*: an essay on the organization of experience. Cambridge: Harvard University Press, 1975.

KONERDING, K.-P. *Frames und lexikalischesBedeutungswissen*. Untersuchungenzurlinguistischen Grundlegungeiner Frametheorie und zuihrer Anwendung in der Lexikographie. Tübingen: Niemeyer, 1993.

KÖVECSES, Z. *Language, mind and culture*. A practical introduction. New York: Oxford University Press, 2006.

LAKOFF, G. *Don't think of an elephant!* Know your values and frame the debate. Vermont: Chelsea Green Publishing, 2004.

LANGACKER, R. W. *Cognitive Grammar:* a basic introduction. New York: Oxford University Press, 2008.

MINSKY, M. *A framework for representing knowledge*. Artificial Intelligence Memo, n. 306. Cambridge: Massachusetts Institute of Technology, 1974.

MIRANDA, N. S. Domínios conceptuais e projeções entre domínios: uma introdução ao Modelo dos Espaços Mentais. *Veredas*: revista de estudos linguísticos, Juiz de Fora, v. 3, n. 1, p. 81-95, 1999.

ROSCH, E. Natural categories. *Cognitive Psychology*, v. 4, n. 3, p. 328-350, 1973.

SANTOS, A. N. *Direito, aborto e anencefalia no Brasil:* uma análise semântico-cognitiva do processo da ADPF-54. 2016. 2016. 161 f. Dissertação (mestrado em Linguística Aplicada). Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, 2016. Disponível em: <http://www.repositorio.jesuita.org.br/handle/UNISINOS/5203>. Acesso em: 09 out. 2017.

SANTOS, A. N.; CHISHMAN, R. L. de O. O papel da Semântica de *Frames* na construção de um recurso dicionarístico: a organização lexicográfica do Field – Dicionário de Expressões do Futebol. *Revista da ABRALIN*, v. 14, n. 3, p. 433-468, 2015.

_____. Direito e anencefalia no Brasil: uma abordagem semântico-cognitiva da ADPF 54. *Revista da Anpoll*, Florianópolis, v. 1, n. 42, p. 52,70, 2017. Disponível em: <https://revistadaanpoll.emnuvens.com.br/revista/article/view/925>. Acesso em: 20 nov. 2017.

SCHANK, R. C.; ABELSON, R. P. *Scripts, plans, goals, and understanding*: An inquiry into human knowledge structures. Hillsdale: Lawrence Erlbaum, 1977.

SEARLE, J. R. *Speech acts*: an essay in the philosophy of language. Cambridge: Cambridge University Press, 1969.

TANNEN, D.; WALLAT, C. Interactive Frames and Knowledge Schemas in Interaction: Examples from a Medical Examination/Interview. In: TANNEN, D. *Framing in Discourse*. New York: Oxford University Press, 1987.

ZIEM, A. *Frames of understanding in text and discourse*. Amsterdam: John Benjamins, 2014.

O estudo do estilo na legendagem: uma pesquisa baseada em *corpus*

The study of style in subtitling: a *corpus*-based research

Janailton Mick Vitor da Silva
Alessandra Ramos de Oliveira Harden

Resumo: Este capítulo faz um recorte teórico e metodológico da pesquisa de mestrado, ainda em andamento, sobre a temática de estilo da tradução no âmbito da legendagem, realizada no Programa de Pós-Graduação em Estudos de Tradução, na Universidade de Brasília. Filiando-se aos Estudos da Tradução Baseados em *Corpus*, a pesquisa objetiva investigar estilo na tradução de legendas de documentários, a partir da comparação de traduções feitas por duas legendistas para a Netflix. Para os fins de análise estilística, serão utilizados os recursos Wordlist e Concord do programa Word Smith Tools® (versão 7.0).

Palavras-chave: Estilo da tradução. Estilo do tradutor. Legendagem profissional. ETBC.

Janailton Mick Vitor da Silva – Aluno da Universidade de Brasília, licenciado pela Universidade Federal de Campina Grande, bolsista CAPES – janailtonm@gmail.com.
Alessandra Ramos de Oliveira Harden – Professora da Universidade de Brasília, doutora pela *University College Dublin* – oliveira.ales@gmail.com.

Abstract: This chapter makes a theoretical and methodological extract from an ongoing Master's research on the theme of translation style in the field of subtitling, carried out at the Graduate Program in Translation Studies at the University of Brasília (POSTRAD/UnB). Following *Corpus*-Based Translation Studies, the research aims to investigate style in the translation of documentary subtitles by comparing the works of two subtitlers for Netflix. For the purposes of stylistic analysis, Wordlist and Concord of Word Smith Tools© (version 7.0) will be used.

Keywords: Translation style. Translator's style. Professional subtitling. CTS.

1 Introdução

Reconhecidos como área acadêmica independente há algum tempo, os Estudos de Tradução vêm se caracterizando como uma disciplina que dialoga com diversos campos do saber. De fato, enquanto atividade e território de investigações teóricas, a tradução abraça inúmeros aspectos das relações humanas. Entre essas muitas áreas de contato que hoje se estabelecem no âmbito dos Estudos da Tradução, estão os estudos em Tradução Audiovisual (TAV), os quais envolvem reflexões acerca de linguagens diferentes e conhecimentos técnicos diversos que se conjugam para permitir maior acesso a produtos culturais de natureza específica. Nesse âmbito, as formas de comunicação são compostas por elementos sonoros e imagéticos, que, combinados e sincronizados, cooperam para a geração de significados.

Neste texto, a ênfase é dada à legendagem, uma das principais formas pelas quais a TAV se concretiza. Sem desconsiderar os aspectos técnicos, por certo inseparáveis de qualquer análise ou avaliação dessa prática, a ênfase aqui é dada ao lado humano das legendas, aquele que se refere ao uso da língua feito na legendagem, ou seja, às escolhas linguísticas e tradutórias realizadas pelo legendista.

Para começar a discussão, é preciso estabelecer alguns pontos de partida. O mais importante deles é apresentar o que entendemos por legendagem. Assim, para os fins da investigação aqui anunciada, a legendagem consiste na tradução para um texto escrito, visível geralmente ao fim da tela, de diálogo original de falantes. Ela inclui elementos discursivos dispostos na imagem e informações provenientes da trilha sonora (DÍAZ CINTAS; REMAEL, 2007). As legendas devem obedecer a alguns critérios básicos, tais como: i) aparecer em sincronia com imagem e diálogo; ii) fornecer uma tradução semanticamente adequada do diálogo da língua-fonte; iii) permanecer na tela tempo suficiente para sua completa visualização; e iv) interagir com signos visuais, acústicos e demais códigos da obra audiovisual em foco (CHAUME, 2004; DÍAZ CINTAS; REMAEL, 2007; GEORGAKOPOULOU, 2009; GOTTLIEB, 1994; 1998; 2005a; 2005b; IVARSSON; CARROLL, 1998).

Essas características das legendas parecem condicionar o trabalho dos tradutores e influenciar suas decisões linguísticas e tradutórias, levando a legendagem

a ser considerada uma prática um tanto "restritiva". Dessa forma, toda escolha do tradutor é definida, antes de tudo, por essas limitações. Assim, muitos pesquisadores criticam o *status* da legendagem, e até o da TAV, considerando-a como adaptação, e não como tradução propriamente dita (GAMBIER, 2003). Essa é uma visão dicotômica cuja origem pode ser identificada na tradicional discussão linguística sobre fidelidade e equivalência formal, uma preocupação mais encontrada entre os estudiosos de vertentes estruturalistas da tradução (OLIVEIRA, 2007).

Neste artigo, porém, a concepção que temos de tradução é mais abrangente, capaz de englobar atividades de variadas naturezas e agentes, e aberta a influências de códigos diversos, como é a legendagem, que se faz entre o linguístico, o visual e o sonoro. Assim, seguindo os preceitos de Díaz Cintas e Remael (2007), compreendemos a tradução sob "uma perspectiva mais flexível e heterogênea e menos estática, que abarque um amplo conjunto de realidades empíricas e reconheça a natureza em constante mudança dessa prática"[1] (DÍAZ CINTAS; REMAEL, 2007, p. 10, tradução nossa)[2].

Nesse sentido, as considerações tecidas neste texto são feitas com base no reconhecimento da natureza híbrida e heterogênea da tradução, que ocorre também com textos multimodais e polissemióticos, frutos de uma era globalizada em que a comunicação e a produção de informações não mais se materializam apenas a partir do canal verbal escrito e falado.

Outro pressuposto importante desta pesquisa é a compreensão de que a atividade tradutória é realizada por homens e mulheres que, como todos os outros profissionais do planeta, trabalham de forma contextualizada, ou seja, sob condições sociais, econômicas e culturais que condicionam o resultado de seus esforços de uma forma ou de outra (PYM, 2012; 1998). De fato, os tradutores não trabalham em isolamento, e o aparato social que os rodeia reflete, em maior ou menor grau, o próprio mercado de tradução (DÍAZ CINTAS, 2004; MUNDAY, 2008). No caso da legendagem, além das restrições técnicas já mencionadas, os tradutores também precisam observar normas institucionais (ligadas à empresa que os contrata) e outras regras que lhes são impostas tanto pelo gênero da obra e audiência pretendida, quanto por questões estéticas ou artísticas definidas por diretores e produtores (DÍAZ CINTAS; REMAEL, 2007).

Dadas essas condições, é provável que, num primeiro momento, seja desafiador ver o tradutor como um produtor de discursos. São tantas as limitações que há dúvidas sobre a possibilidade de os tradutores deixarem suas marcas no texto. É nesse ponto que reside o interesse desta pesquisa: diante das condições sociais, técnicas e institucionais às quais os tradutores estão submetidos, eles podem ocupar

[1] "*Translation must be understood from a more flexible, heterogeneous and less static perspective, one that encompasses a broad set of empirical realities and acknowledges the ever-changing nature of practice*" (DÍAZ CINTAS; REMAEL, 2007, p. 10).

[2] Todas as traduções são de nossa autoria, excetuando-se aquelas já publicadas.

algum espaço nas legendas? Há como deixar um rastro identificável a partir das escolhas linguísticas e/ou tradutórias que fazem?

Para tentar responder a essas perguntas, recorremos aos estudos de estilo iniciados por Baker (2000), que define estilo do tradutor como um tipo de impressão digital. O conceito de estilo do tradutor foi retomado e aprofundado em outras pesquisas, tais como nas de Saldanha (2011a; 2011b), para quem o estilo é uma "forma de traduzir" do tradutor, que: i) é reconhecível em mais de uma tradução sua; ii) é distinguível do trabalho de demais tradutores; iii) constitui um padrão coerente de escolha; iv) apresenta uma ou mais funções discerníveis; e v) independe do estilo do autor e do(s) texto(s) fonte(s) (TF/s) e de limitações linguísticas (SALDANHA, 2011b).

Essa noção de estilo é entendida como um "atributo pessoal", caracterizando o trabalho do tradutor enquanto indivíduo que também apresenta um estilo próprio. Combinada a essa proposta de estudo do estilo, Saldanha (2011a; 2011b) distingue outra, estilo como "atributo textual", que se ocupa do estudo do estilo do(s) texto(s) traduzido(s) (TTs). Sendo assim, seguindo a pesquisadora, estudamos o estilo numa perspectiva combinada, levando-se em conta ambos os atributos.

Seguindo os preceitos de Baker (2000) e Saldanha (2011a; 2011b), acreditamos que o estilo do tradutor possa ser identificado a partir da comparação de vários trabalhos feitos por tradutores distintos, quando se analisam, apesar da diversidade de traduções, as mesmas características estilísticas compartilhadas entre si. Portanto, o foco da pesquisa aqui apresentada, ainda em andamento, recai sobre o trabalho profissional de duas tradutoras que legendam obras audiovisuais disponibilizadas pela Netflix. Buscamos comparar e contrastar as escolhas linguísticas e tradutórias semelhantes que ambas fazem na tradução de seus respectivos documentários, na tentativa de construir um perfil estilístico para cada uma delas. Entendemos aqui escolhas linguísticas como aquelas imediatamente perceptíveis na tela, como a escolha de palavras e expressões e a sua apresentação sintática, enquanto escolhas tradutórias como estratégias de tradução e procedimentos técnicos utilizados pelas tradutoras.

Levando-se em conta essas discussões, o objetivo deste capítulo é apresentar um recorte teórico e metodológico da dissertação de mestrado sobre a temática de estilo da tradução e do tradutor, no âmbito da legendagem. Toda discussão feita neste capítulo é baseada em pesquisa em andamento no Programa de Pós-Graduação em Estudos de Tradução (POSTRAD) da Universidade de Brasília (UnB)[3], cujo foco recai sobre o estilo de duas legendistas que traduzem

[3] Pesquisa de autoria de Janailton Mick Vitor da Silva, em andamento no POSTRAD/UnB, com bolsa da CAPES, e orientada pela profa. dra. Alessandra Ramos de Oliveira Harden.

documentários e séries de TV para a Netflix, Carla A. A. Prado e Karina C. Alves[4]. A pesquisa mencionada busca analisar estilo a partir da identificação e comparação de padrões estilísticos (linguísticos e/ou tradutórios) de duas legendistas distintas, a partir das legendas de quatro documentários.

As pesquisas de estilo de Baker (2000) e Saldanha (2011a; 2011b) citadas, bem como algumas outras que igualmente se debruçam sobre o tema, filiam-se aos Estudos da Tradução Baseados em *Corpus* (ETBC) e fazem uso de ferramentas da Linguística de *Corpus* (LC). Trata-se de recursos que permitem operar com grande quantidade de textos e, assim, rastrear padrões de uso numa mesma língua ou em mais de uma língua envolvida nesse processo (BAKER, 1999; 2000; 2004; BARCELLOS, 2016a; 2016b; BERBER-SARDINHA, 2004; 2009; CAMARGO, 2007; LEECH; SHORT, 2007; MALMKJAER; CARTER, 2002; MUNDAY, 2008; SALDANHA, 2011a; 2011b). Na pesquisa aqui introduzida, igualmente fazemos uso de um programa da LC, o Word Smith Tools® (WST), versão 7, para que se possam identificar e analisar aspectos estilísticos nos *corpora* escolhidos.

Este capítulo traz outras três partes além desta. A próxima seção realiza uma aproximação teórica dos conceitos de estilo dentro do campo da legendagem. Na seção seguinte, apresentam-se aspectos metodológicos da pesquisa em curso, ainda que embrionários e passíveis de futuras alterações. Em seguida, tecem-se algumas considerações. Por fim, apresentam-se as referências utilizadas no trabalho.

2 Os estudos sobre estilo e a legendagem: aproximações teóricas

A noção de estilo, segundo Leech e Short (2007, p. 9), refere-se à "forma na qual a língua é usada num dado contexto, por certa pessoa, para um fim específico etc."[5]. Tradicionalmente, estilo está associado a textos literários escritos, mas também pode ser aplicado à língua falada e a outros gêneros (LEECH; SHORT, 2007; MALMKJAER; CARTER, 2002; MALMKJAER, 2003; 2004; MUNDAY, 2008). Além disso, para se falar em estilo é preciso também considerar o domínio extralinguístico, que se relaciona com a língua em uso, o escritor e também o tradutor, levando-se em conta certos questionamentos, tais como: i) quem são o autor e o tradutor?; ii) quais os períodos de escrita e tradução?; iii) quais os contextos socioculturais do autor e do tradutor? (BAKER, 2000; BOASE-BEIER, 2014; LEECH; SHORT, 2007; MALMKJAER, 2003; 2004; MUNDAY, 2008; SALDANHA, 2011b).

[4] Ambas as tradutoras permitiram o uso de seus nomes em publicações relacionadas à pesquisa, conforme comunicação por *e-mail* em 17/10/2017 e 25/10/2017.

[5] "[…] *it refers to the way in which language is used in a given context, by a given person, for a given purpose, and so on"* (LEECH e SHORT, 2007, p. 9).

A tradução está associada à estilística, área que se ocupa do estudo (linguístico) do estilo, uma vez que busca, como a estilística, compreender como os textos fazem sentido, concentrando-se em aspectos como estilo, escolha e efeito, como afirma Boase-Beier (2011; 2014). A partir dessa relação, Malmkjaer (2003) cunhou o termo "estilística tradutória" (*translation stylistics*) para se referir ao sub-ramo da estilística que se ocupa em estudar "os motivos pelos quais uma tradução, *dado o texto fonte*, foi feita de determinada maneira que passa a significar do modo que significa"[6] (MALMKJAER, 2003, p. 39, grifos da autora). Por outro lado, "estilo", noção com a qual essa autora já trabalhava anteriormente, seria a "ocorrência consistente e estatisticamente significativa de itens e estruturas, ou tipos de itens e estruturas [num texto], dentre outras oferecidas na língua"[7] (MALMKJAER; CARTER, 2002, p. 510).

A compreensão de estilo da tradução por Boase-Beier (2011; 2014) e Malkmjaer (2003; 2004) é, no entanto, dependente do trato com o TF (texto fonte) em relação ao TT (texto traduzido), mais do que com o tradutor. Apesar de considerarem que o estilo da tradução é influenciado pela interpretação subjetiva do tradutor, ainda dão ênfase ao texto de partida e à reprodução do seu estilo no texto de chegada (SALDANHA, 2011b). Para Saldanha (2011b), Boase-Beier (2011; 2014) coloca a responsabilidade do estilo do TT nas mãos do tradutor, o qual o alcançaria a partir da recriação do estilo e significado do TF. A tarefa de recriar esse estilo do texto impediria qualquer criação artística por parte do tradutor. É nesse sentido que Saldanha (2011b) diferencia duas abordagens de estilo, uma como atributo textual (*translation style*), ou seja, estilo do TT, e outra como atributo pessoal (*translator style*), isto é, estilo do tradutor. A autora sugere uma proposta combinada das duas e admite que, em qualquer uma dessas abordagens, não se pode buscar reproduções estilísticas do TF. Assim, "considerar estilo como atributo pessoal e textual é atribuir responsabilidade pelas escolhas estilísticas e ir além do texto de partida como fonte de motivação"[8] (SALDANHA, 2011b, p. 28).

É Mona Baker quem inicia os estudos sobre estilo do tradutor. Em 2000, data da publicação de seu artigo sobre o tema na revista *Target*, Baker afirmou ter havido pouco ou nenhum interesse em se estudar o estilo do tradutor ou de um grupo de tradutores em um *corpus* de material traduzido pertencente a um período histórico particular, e explica que esse desinteresse pode ter resultado do fato de a tradução ser tradicionalmente vista como uma atividade derivativa, e

[6] "*why, **given the source text**, the translation has been shaped in such a way that it comes to mean what it does*" (MALMKJAER, 2003, p. 39, grifos da autora).

[7] *"By **style** is meant a consistent occurrence in the text of certain items and structures, or types of items and structures, among those offered by the language as a whole."* (MALMKJAER; CARTER, 2002, p. 510, grifos dos autores).

[8] *"Considering style as a personal attribute, as well as a textual one, allows us to attribute responsibility for stylistic choices and to go beyond the source text in search for motivation"* (SALDANHA, 2011b, p. 28).

não criativa (BAKER, 2000). Essa suposição acerca das características da atividade tradutória teria levado à conclusão errônea de que o tradutor não pode nem deve ter um estilo próprio, sendo apenas reprodutor do estilo do TF. Contudo, a autora afirma ser impossível a total impessoalidade ao se produzir linguagem, da mesma forma que não se pode evitar deixar impressões digitais quando se toca num objeto (BAKER, 2000).

Assim, a pesquisadora define estilo "como um tipo de impressão digital, expressa em uma variedade de características linguísticas e não linguísticas"[9] (BAKER, 2000, p. 245). No caso do tradutor literário, a definição de estilo do tradutor deveria incluir:

- a preferência do tradutor por traduzir um tipo de material específico, quando aplicável, da mesma forma que seu uso consistente de estratégias tradutórias, como o uso de prefácios, posfácios, notas de rodapé, comentários ao longo do texto, entre outros;
- a forma de expressão típica do tradutor, em vez de apenas intervenções claras;
- o uso característico da língua feito pelo tradutor, com base na criação de um perfil individual de hábitos linguísticos que se revelam quando comparado a outros tradutores;
- a descrição do comportamento linguístico do tradutor, observado a partir de sua preferência consistente por itens lexicais, padrões sintáticos, dispositivos coesivos ou até mesmo estilo de pontuação, quando outras opções estariam igualmente disponíveis na língua.

Saldanha (2011b) segue adiante no estudo do estilo do tradutor, iniciado por Baker (2000), e define estilo do tradutor (*style as personal attribute*) como uma "forma de traduzir", que:

i. é reconhecível em mais de uma tradução pelo mesmo tradutor;
ii. distingue o trabalho de um tradutor do trabalho de outros tradutores;
iii. constitui um padrão coerente de escolha;
iv. é "motivada", no sentido de possuir uma ou mais funções discerníveis;
v. não pode ser puramente explicada com referência ao estilo do autor ou do texto fonte, ou como resultado de limitações linguísticas[10]. (SALDANHA, 2011b, p. 31)

[9] *"I understand style as a kind of thumb-print that is expressed in a range of linguistic — as well as non-linguistic – features*" (BAKER, 2000, p. 245).
[10] "*1. is felt to be recognizable across a range of translations by the same translator, 2. distinguishes the translator's work from that of others, 3. constitutes a coherent pattern of choice, 4. is 'motivated', in the*

Ambos os conceitos de estilo do tradutor, de Baker (2000) e de Saldanha (2011a; 2011b), dão visibilidade ao tradutor, pois reconhecem que esse profissional pode ocupar espaços nas suas traduções por meio de intervenções linguísticas e tradutórias consistentes. Especificamente sobre a proposta de Saldanha (2011b), a pesquisadora lembra que não se devem buscar justificativas para o estilo do TT e do tradutor no estilo do TF, pois, se assim o for, o trabalho feito pelo tradutor seria uma reprodução do estilo do TF. Assim, não poderia haver atribuição de estilo nem ao TT nem ao tradutor. Desse modo, o estilo como atributo textual e pessoal existe para além do estilo do TF. E, especialmente para o tradutor, não se devem buscar influências do autor do TF e de diferenças linguísticas em seu trabalho. No âmbito de nossa pesquisa de mestrado, acreditamos que o estilo do tradutor seria construído no TT e através dele, texto esse que, *a priori*, já tem um estilo próprio que independe do estilo do TF. Seu estilo também seria identificável a partir da recorrência de padrões em mais de uma tradução.

A presença do tradutor poderia ser entendida aqui a partir da diferenciação entre "estilo" e "voz" feita por Munday (2008). Para ele, voz "se refere ao conceito abstrato de presença autoral, narratorial ou translatorial", enquanto estilo "é a manifestação linguística dessa presença no texto"[11]. Especificamente a respeito de voz, Munday (2008) defende que não é apenas o autor que deixa sua presença no texto, mas também o tradutor. De fato, quando se estuda estilo do TT com base em padrões linguísticos do tradutor, observa-se como esses padrões afetam ou não a voz narrativa do autor do TF, aquela que ecoa através da voz do tradutor. A presença discursiva do tradutor poderia ser medida, por exemplo, a partir de sua criatividade nas suas escolhas e seleções linguísticas recorrentes.

No que se refere à legendagem, pode-se dizer que essa forma de tradução tem um estilo próprio, pois as legendas são produzidas de acordo com critérios técnicos, linguísticos e tradutórios[12]. Sob uma perspectiva técnica, elas devem aparecer em sincronia com imagem e diálogo, e permanecer na tela o suficiente para sua completa visualização, obedecendo a um número de caracteres e tempo de apresentação definidos pela empresa legendadora. Num viés linguístico, a legendagem é uma forma de reescrita que passa por redução na forma de condensação, reformulação e omissão no(s) nível(is) da palavra e/ou sentença. No nível tradutório, o legendista lida com a tradução de elementos da oralidade, variações linguísticas e nuances

---

sense that it has a discernable function or functions, 5. cannot be explained purely with reference to the author or source-text style, or as the result of linguistic constraints" (SALDANHA, 2011b, p. 31).

[11] "*Whereas we shall use voice to refer to the abstract concept of authorial, narratorial, or translatorial presence, we consider style to be the linguistic manifestation of that presence in the text"* (MUNDAY, 2008, p. 19).

[12] Esses critérios são mais amplamente discutidos em Díaz Cintas, 2005; 2013; Díaz Cintas e Remael, 2007; Georgakopoulou, 2009; Gambier, 2003; Gottlieb, 1994; 1998; 2005a; 2005b; Ivarsson e Carroll, 1998; Naves et al., 2016; Perego, 2003; 2009.

como o contexto e o cotexto das cenas, o gênero da obra audiovisual, as culturas envolvidas, entre outros aspectos. Nos Quadros[13] 1, 2 e 3, apresentamos o que entendemos aqui por características estilísticas das legendas, numa perspectiva de sua criação e apresentação na tela, em forma resumida e com sua respectiva indicação de fontes. Essas características são resumidas a partir de pesquisas dos autores antes citados, bem como dos Guias de Estilo da empresa Netflix.

O Quadro 1 lista características técnicas das legendas, que se referem ao modo de confecção das legendas, principalmente no que diz respeito à quantidade de caracteres por linha de legenda, ao tempo de exibição na tela e ao uso de recursos tipográficos.

Quadro 1 – Características estilísticas das legendas: aspectos técnicos

DÍAZ CINTAS e REMAEL (2007)	**NETFLIX (2016a; 2016b)**
• Tamanho: 2 linhas; • Caracteres/linha: 40-41 (DVD e cinema); 28-37 (TV); • Tempo: 1-6 segundos; • Espaço de tempo de 2 a 4 *frames* entre as legendas; • Posição: topo e base da tela e vertical; • Fonte Arial, Helvetica ou Times New Roman, e cor branca; • Uso de sinais tipográficos por extenso; • Conversão de moeda a depender do caso; • Itálico: palavras em outras línguas; referências bibliográficas e literárias; títulos de publicações, livros, *shows* etc.; vozes em *off*; letras de música; • Uso de abreviações por processos como redução, acrônimo, contração e combinação; • Números escritos em dígito ou por extenso; • Letras maiúsculas têm sido usadas para designar: título do programa audiovisual, sinais de trânsito, grafite, manchetes de jornais, nomes em roupas, *banners*, mensagens em telas de computador, entre outras inserções que merecem ênfase.	• Tamanho: 2 linhas; • Caracteres/linha: 42; • Tempo: 5-7 segundos; • Espaço de tempo de 2 *frames* entre as legendas; • Posição: centro, topo e base da tela e lado vertical; • Fonte Arial, cor branca, tamanho a variar de acordo com o tamanho da tela; • Sincronização com áudio e mudança de plano; • Não conversão de moeda; • Itálico: palavras em outras línguas; títulos de álbum, livro, filme e programa; vozes em *off*; letras de música; • Uso de abreviações por processos como redução, acrônimo e contração; • Números escritos em dígito ou por extenso; • Uso de aspas duplas para indicar a voz de outros falantes e aspas simples para citações dentro de citações.

O Quadro 2 enfatiza a caracterização linguística das legendas, ou seja, as mudanças e adaptações lexicais e sintáticas que precisam ser feitas no TT. Vale ressaltar que os manuais da Netflix não indicam explicitamente esses aspectos.

[13] Todos os quadros aqui elaborados são de nossa autoria.

Quadro 2 – Características estilísticas das legendas: aspectos linguísticos

DÍAZ CINTAS; REMAEL (2007)	PEREGO (2009)	NETFLIX (2016a; 2016b)
• Simplificação de perífrases verbais; • Generalização de enumerações; • Utilização de sinônimos pequenos ou expressões equivalentes; • Utilização do tempo simples ao composto; • Mudança de classe de palavras; • Uso de formas curtas ou contrações; • Mudança de sentenças negativas para afirmativas, indiretas para diretas, retóricas para afirmativas etc.; • Simplificação de modais e marcadores de modalidade; • Mudança de: discurso direto para indireto; sujeito da sentença ou oração; sentenças complexas para simples; sentenças negativas para afirmativas; sentenças indiretas para diretas; sentenças passivas para ativas; • Manipulação de tema e rema; • Substituição de substantivos ou sintagmas nominais por pronomes e outros dêiticos; • Combinação de duas ou mais orações/sentenças em uma; • Segmentação linguística, retórica e visual.	• Concisão; • Menos redundância; • Alto grau de organização textual, informatividade, planejamento de informações, coesão e coerência; • Especificação de referentes; • Explicitação no desenvolvimento de argumentos; • Desambiguação de formas pronominais; • Escolhas lexicais específicas; • Reconstrução de formas elípticas.	• Segmentação linguística, retórica e visual.

Por fim, o Quadro 3 apresenta normas que vão além da técnica e da adaptação linguística, trazendo elementos associados à filtragem do que pode ou não ser traduzido, às ferramentas de auxílio ao tradutor e à visibilidade tradutória[14].

[14] Vale ressaltar que, segundo Naves et al. (2016), questões tradutórias também envolvem a operacionalização das características técnicas e linguísticas.

Quadro 3 – Características estilísticas das legendas: aspectos tradutórios

DÍAZ CINTAS; REMAEL (2007)	NETFLIX (2016a; 2016b)
• Exclusão total e/ou parcial de: *Question tags,* modificadores (adjetivos e advérbios), elementos interpessoais (cumprimentos, interjeições, vocativos), outras palavras fáticas (*anyway, you know*), hesitações, repetições e falsos começos. • Correção provável de trechos agramaticais e neutralização de dialetos, socioletos e idioletos na língua-alvo; • Uso de estratégias de tradução[15] (empréstimo, calque, explicitação, substituição, transposição, (re)criação lexical, compensação, adição, omissão); • Tradução de humor mediante interação palavra-imagem; jogo com as palavras ou com uso de características próprias do gênero; intertextualidade.	• Proibição de censura a diálogo (nunca deve ser censurado); • Aceitabilidade de ambos os registros da língua (exemplo: norma culta e coloquial); • Uso de quaisquer formas da segunda pessoa singular (exemplo: você e tu); • Tradução de todo e qualquer texto disponível na tela que seja relevante para a trama; • Não tradução do título principal, a menos que uma tradução aprovada seja fornecida pela Netflix; • Legendagem apenas de músicas pertinentes à trama e se os direitos tiverem sido concedidos. • Utilização de glossários com termos estabelecidos ao longo do filme e série; • Manutenção de palavras ou frases que se repetem mais de uma vez na língua do TF; • Manutenção de nomes próprios na língua do TF, com exceção de traduções oficiais fornecidas pela Netflix; • Tradução de apelidos apenas quando transmitirem algum significado específico ou especial; • Uso de traduções específicas da língua-alvo para personagens históricos/místicos; • Exclusão (na tradução) de erros ortográficos e problemas de pronúncia intencionais, a menos que seja pertinente ao enredo; • Inclusão de créditos em letra maiúscula em obras Originais da Netflix, de acordo com o gênero, (por exemplo: 'UMA SÉRIE ORIGINAL NETFLIX'); • Inclusão de crédito a apenas um/a tradutor/a, como a última legenda da obra, de acordo com o *Original Credits Translation Document;* • Tradução de diálogo em língua estrangeira apenas se for necessário, para a compreensão da obra como um todo; • Verificação de ortografia, sotaque e pontuação corretas, conforme o caso, se houver uso de palavras estrangeiras.

[15] Baseados em Díaz Cintas (2003 apud DÍAZ CINTAS; REMAEL, 2007) e Santamaria Guinot (2001 apud DÍAZ CINTAS; REMAEL, 2007), Díaz Cintas e Remael (2007) propuseram algumas estratégias que podem auxiliar o legendista.

As características apresentadas nos quadros 1, 2 e 3 apresentam o modo de construção e apresentação das legendas na tela. Por um lado, essas normas refletem a "forma" (LEECH; SHORT, 2007) como a língua é apresentada na tela e, por outro, como ela é empregada num determinado contexto para seus devidos fins e efeitos por um legendista. Diante do exposto, e tomando como base outras definições sobre estilo[16] anteriormente apresentadas, chegamos à seguinte definição de estilo empregada em nossa pesquisa:

- O estilo da legendagem é a forma como a língua é apresentada e empregada nas legendas, resultante de influências técnicas, linguísticas e tradutórias, e condicionada ao aparato polissemiótico do produto audiovisual, para determinados propósitos e efeitos.

Essa definição, contudo, não demarca o lugar específico que o tradutor ocupa nas legendas. Ela está, nos dizeres de Díaz Cintas e Remael (2007), relacionada à tarefa, um tanto paradoxal e contraditória, do legendista, que deve criar uma legenda pós-produção da obra, que apareça e desapareça da tela, mas que aparente não estar lá. Nesse sentido, é preciso fixar os olhares na figura do legendista.

Diante de normas estabelecidas por outrem, o tradutor produz suas traduções e (re)cria discursos, o que indica que o seu estilo pode ser construído no TT e através dele. Nesse sentido, tomando como base as definições de estilo do tradutor de Baker (2000), Munday (2008) e Saldanha (2011a; 2011b), elaboramos a definição que segue:

- O estilo do legendista é o conjunto de padrões e hábitos linguísticos e não linguísticos, influenciados e/ou determinados pelo estilo da legendagem enquanto atributo textual, que refletem um padrão consistente de escolhas linguísticas e não linguísticas feitas pelo legendista em mais de uma tradução e para determinados propósitos e efeitos.

É fato que as definições sobre estilo do escritor e do tradutor apresentadas anteriormente fazem referência a escolhas feitas dentro um leque maior de opções disponíveis na língua. No caso da legendagem, esse leque torna-se ainda mais restrito, pois as opções do legendista estão sujeitas – mais do que simplesmente disponíveis e passíveis de uso irrestrito – às condições técnicas, linguísticas e tradutórias do meio. Dessa forma, essas escolhas seriam também fruto de uma influência/determinação da própria legendagem, e não apenas da língua. Não obstante, buscamos analisar o estilo de duas legendistas, a partir da identificação

[16] Foram usadas as definições de Baker (2000), Boase-Beier (2011; 2014), Leech e Short (2007), Malkmjaer (2003; 2004), Malkmjaer e Carter (2002), Munday (2008) e Saldanha (2011a; 2011b).

de padrões estilísticos que possam indicar um perfil individual tradutório para cada uma delas.

3 Os *corpora* da pesquisa: apontamentos teóricos e metodológicos nos ETBC

Nesta pesquisa, afiliamo-nos aos ETBC, tendo em vista a necessidade de lidarmos com a manipulação eletrônica de TFs e TTs, a partir do uso de programa computacional para fins de análise linguística sob uma perspectiva descritiva (BAKER, 1995; 1996; CAMARGO, 2007). *Corpus* é aqui entendido como uma coleção de textos em formato eletrônico e passíveis de análises automáticas ou semiautomáticas (BAKER, 1996).

Os *corpora* da pesquisa foram divididos em quatro: i) um *corpus* de TTs por Carla Prado (CTCP); ii) um *corpus* de TTs por Karina Curi (CTKC); iii) dois *corpora* paralelo: a) um com TFs e TTs por Carla Prado (CPCP: *Corpus* Paralelo Carla Prado); e b) um com TFs e TTs por Karina Curi (CPKC: *Corpus* Paralelo Karina Curi). Os TTs englobam dois documentários para cada tradutora, cada um com uma temática distinta da outra, escrito por pessoas diferentes e lançados em anos distintos. O Quadro 4, a seguir, detalha os nossos *corpora* sob estudo.

Quadro 4 – *Corpora* de estudo

LEGENDISTAS	DOCUMENTÁRIOS[17]	TEMÁTICAS[18]	AUTORES[19]
CARLA A. A. PRADO	*Amanda Knox (2016)* 1h 32 min	Crime, Policial	Roteiro: Brian McGinn, Rod Blackhurst, Matthew Hamachek. Produção: Brian McGinn, Mette Heide, Rod Blackhurst, Stephen R. Morse. Direção: Rod Blackhurst, Brian McGinn.
	Audrie & Daisy (2016) 1h 38 min	Biográfico, Crime	Roteiro: Michael Goodier. Produção: Richard Berge, Bonni Cohen, Sara Dosa. Direção: Bonni Cohen, Jon Shenk.
KARINA C. ALVES[20]	*Tony Robbins: I'm not your guru (2016)* 1h 56 min	Motivacional	Direção e Roteiro: Joe Berlinger. Produção: Joe Berlinger, Kevin Huffman, Lisa Gray.
	Get me Roger Stone (2017) 1h 41 min	Biográfico, Sociocultural	Direção e Roteiro: Daniel DiMauro, Dylan Bank, Morgan Pehme. Produção: Daniel DiMauro, Frank Morano, Fredrik Stanton, Kara Elverson, Morgan Pehme, Shirel Kozak.

A nomenclatura a ser usada nos arquivos eletrônicos é a seguinte:

Quadro 5 – Nomenclatura nos arquivos eletrônicos dos *corpora* da pesquisa

DOCUMENTÁRIOS	NOMENCLATURA TTs	NOMENCLATURA TFs
Amanda Knox	AK_PT_PRADO	AK_EN_PRADO
Audrie & Daisy	AD_PT_PRADO	AD_EN_PRADO
Tony Robbins: I'm not your guru	TB_PT_ALVES	TB_EN_ALVES
Get me Roger Stone	RS_PT_ALVES	RS_EN_ALVES

[17] Os títulos dos documentários foram informados por *e-mail* em 29/04/2017 (Carla) e 10/05/2017 (Karina).

[18] As temáticas aqui apresentadas são classificadas pela Netflix e por nós.

[19] Nesse contexto, "autores" designam as pessoas envolvidas na parte mais técnica da produção dos documentários, como produtores, diretores e roteiristas. Não obstante, reconhecemos que, diferentemente de outros gêneros cinematográficos, que dependem, em grande parte, de roteiros quase que completamente prontos, os entrevistados igualmente participam da autoria dos documentários em que atuam.

[20] A tradutora informou por *e-mail* que geralmente assina as traduções como "Karina Curi". Quando é a Netflix quem assina, coloca-se "Karina Alves".

A seleção desses *corpora* seguiu os pré-requisitos básicos para a construção de qualquer *corpus* na área da LC (BERBER-SARDINHA, 2004). Os *corpora* são compostos de textos autênticos, as transcrições de áudio e as legendas, ambos construídos para fins de entretenimento das obras, e não fins acadêmicos. Eles foram escolhidos por critérios ligados ao objetivo da pesquisa, que é, como já mencionado, analisar o estilo de duas legendistas. Ademais, foram combinados parâmetros[21] sugeridos por pesquisadores que trabalham com estilo dentro dos ETBC (BAKER, 2000; SALDANHA, 2011a; MUNDAY, 2008), tais como ter textos de temáticas distintas, escritos por pessoas distintas e lançados em anos diferentes. Além disso, a representatividade dos *corpora* também se relaciona com os critérios acima elencados e com as palavras dos pesquisadores mencionados. Na nossa pesquisa em andamento, a representatividade também é entendida a partir do tamanho dos *corpora*. Neles, temos documentários que apresentam uma quantidade um tanto similar de minutos, o que nos leva a pensar que possuem, por conseguinte, uma quantidade balanceada de caracteres por áudio transcrito e por tradução em legendas.

Os *corpora* paralelos mencionados são compostos por transcrições de áudio em língua inglesa dos documentários com suas respectivas traduções em língua portuguesa, na forma de legendas. Para a pesquisa de estilo, a utilidade desses *corpora* está na possibilidade de "pesquisar traduções consagradas de certos itens lexicais ou estruturas sintáticas, peculiaridades de determinado(s) tradutor(es), diferenças entre traduções de um mesmo texto [...] etc." (CAMARGO, 2007, p. 18). Nesse sentido, o estudo das peculiaridades linguísticas e tradutórias de duas tradutoras pode apontar para a constituição de estilos próprios.

É preciso reconhecer dificuldades metodológicas para o estudo do estilo de modo geral e, em especial, no âmbito da legendagem. Para Baker (2000), há inúmeras variáveis que dificultam a atribuição de estilo para o tradutor, tais como as línguas de partida e chegada; o estilo do autor; os socioletos do autor e do tradutor; o estilo do TF; as culturas de partida e de chegada; o contexto sócio-histórico, profissional e cultural do autor e do tradutor; e a natureza do material envolvido na tradução (exemplo: nível de dificuldade; público de chegada e público alvo; tipo de texto; gênero textual). A essas variáveis, Saldanha (2011a) adiciona outras: as características específicas do subgênero do texto a ser traduzido; a variedade linguística com a qual o tradutor lida; as limitações linguísticas; entre outras.

[21] O *corpus* / os *corpora* deveria(m) ter: i) certa variedade de autores, gêneros e subgêneros, datas e locais de publicação, variedades linguísticas e línguas (SALDANHA, 2011a); ii) textos de tamanhos parecidos e do mesmo gênero textual (BAKER, 2004); iii) uma vasta gama de traduções que mantenham relações entre si, feitas pelo mesmo tradutor e escritas pelo mesmo autor, de gêneros parecidos ou distintos; ou análise de várias traduções da mesma obra por tradutores distintos (MUNDAY, 2008).

Saldanha (2011a) sugere que, para minimizar algumas dessas variáveis, pode-se ter um *corpus* com certa variedade de autores, gêneros e subgêneros, datas e locais de publicação, variedades linguísticas e de línguas. Um *corpus* de referência também pode ser importante como parâmetro de análise e comparação. A autora propõe uma abordagem combinada, a partir da noção de estilo como atributo pessoal e textual, e defende a análise de vários trabalhos, de diferentes gêneros e tipos textuais, feitos pelo mesmo tradutor. Por outro lado, Baker (2000) indicou que seria adequado comparar várias traduções diferentes do mesmo TF por tradutores diversos, mantendo, ao menos nesse caso, as variáveis autor-língua-fonte constantes. Outra forma de estudar estilo seria a comparação feita entre textos originalmente escritos pelo tradutor e textos por ele traduzidos (BARCELLOS, 2016a; 2016b).

Todas essas possibilidades e dificuldades se aplicam ao estudo do estilo do tradutor literário. No caso de análise focada no trabalho de legendistas, não se sabe ao certo se as mesmas variáveis teriam igual impacto, nem se outras variáveis deveriam ser levadas em conta, como a atuação de revisores e, conforme já aventado por Díaz Cintas (2004) e Munday (2008), as pressões institucionais e ideológicas enfrentadas por tradutores.

Diante das dificuldades elencadas, compreendemos aqui que o grande desafio para pesquisas na área de estilo "é desenvolver metodologias cada vez mais robustas que permitam identificar padrões de escolhas linguísticas de tradutores" (BARCELLOS, 2016a, p. 2). As metodologias alinhadas aos preceitos da LC podem ajudar nesse sentido, pois permitem, a partir da utilização de programas computacionais, rastrear padrões de uso nas línguas envolvidas nesse processo.

Um grande facilitador de pesquisas no campo dos ETBC, e aqui mais especificamente nos estudos de estilo, tem sido o console WST® em suas várias versões. Observando que as pesquisas têm apontado para a caracterização do que seria estilo em tradução, como atributo textual e/ou pessoal, optamos por igualmente fazer uso dessas ferramentas da LC no âmbito da legendagem. Baños, Bruti e Zanotti (2013) afirmam também que as metodologias oferecidas pela LC podem auxiliar no estudo de TAV, desde que lidem, também, com uma análise multimodal dos dados, e não apenas verbal.

Embora seja reconhecida a necessidade de se analisar os *corpora* desta pesquisa além da materialidade verbal e linguística, é preciso, antes, manipulá-los com o WST®. Até o presente momento, nossa pesquisa tem o caminho metodológico descrito a seguir.

- Obtenção do TF e do TT de cada documentário.

O TF é a transcrição de áudio em inglês e o TT é formado pelas legendas em português para cada um dos documentários. Esse passo será alcançado pela

transcrição do TF e seu respectivo TT. Cada transcrição será feita manualmente, e cada bloco de fala, com sua respectiva tradução, será salva em uma tabela no Microsoft Office Word 2007®, com três colunas, uma para cada dado necessário (personagem / TF / TT). Apesar de ser um processo demorado, ele facilita o alinhamento dos textose a criação dos dois *corpora* paralelos: CPCP e CPKC. Esse tipo de *corpus* será bastante útil na análise entre elementos dos TFs e TTs, pois permite investigar as (possíveis) influências do TF no TT. Em seguida, após término da transcrição, os arquivos em Word serão salvos em formato *.txt. Até o momento da pesquisa, os *corpora* ainda não foram totalmente compilados nem a licença do WST® foi adquirida. Todavia, quando essas etapas anteriores tiverem sido concluídas, intencionamos seguir os passos[22] abaixo.

- Carregamento dos *corpora* no WST® para exploração individual.

Essa exploração se dará inicialmente usando a Wordlist com base no CTPC e CTKC, para verificação da razão forma/item, que indica a variedade lexical dos TFs e TTs nos *corpora*. Para Baker (2000) e Saldanha (2011a; 2011b), essa variedade pode servir como indício do estilo do tradutor. Serão criadas listas de palavras individuais e de agrupamentos lexicais e checadas suas frequências de uso. Buscas no Concord serão feitas de modo a observar o cotexto e contexto de uso das palavras e agrupamentos. Os resultados podem apontar para o estudo de vários aspectos linguísticos, tais como colocações lexicais, expressões idiomáticas, itens culturais específicos, sufixação, sinonímia, prosódia semântica, hiperonímia, entre outros. Uma exploração manual dos *corpora paralelos* pode ser feita para identificar recursos não linguísticos, como o uso de estratégias e procedimentos tradutórios, que serão marcados por etiquetas em cada arquivo por meio de edição.

- Estabelecimento de categorias de análise para o estudo de estilo no*s corpora.*

Serão utilizadas as ferramentas Wordlist e o Concord para detectar padrões estilísticos comuns e distintos às duas tradutoras, tanto no CTCP, quanto no CTKC, o que permitirá a determinação das categorias de análise para o estudo do estilo. Em seguida, será feita a observação de como tais padrões estilísticos constroem os perfis individuais de cada tradutora ao longo de suas traduções, contrastando estatisticamente os padrões entre ambas as legendistas e recorrendo, também, aos *corpora* paralelos para checar os TFs. Nesse momento, serão também levados em conta alguns pontos, tais como: o TF, para checar se o estilo do TT é responsivo a ele, texto esse composto por múltiplas vozes; os guias de estilo da

[22] Reiteramos que outros passos podem igualmente ser seguidos a depender do desenrolar da pesquisa.

Netflix, de modo a perceber se as escolhas das tradutoras respondem apenas às normas linguísticas, técnicas e tradutórias; os padrões estilísticos das tradutoras, observando como eles (re)criam os significados do TF no TT; o construto polissemiótico dos documentários, para verificar se e como as escolhas das tradutoras interagem com os vários signos das obras; entre outros.

Por fim, seja no âmbito da tradução audiovisual, seja no da tradução literária, o estudo de estilo, contudo, não deve ser um fim em si mesmo, pois precisa se propor a algo mais. Segundo Baker (2000), estudar hábitos linguísticos e padrões estilísticos é relevante se "nos contar sobre o posicionamento cultural e ideológico de um tradutor ou de vários tradutores, ou sobre os processos e mecanismos cognitivos que contribuem para moldar nosso comportamento tradutório"[23] (BAKER, 2000, p. 258). Saldanha (2011a) pontua que o estudo do estilo do tradutor se torna relevante quando podemos verificar quais padrões estilísticos refletem a arte do tradutor (*translator's art*) e como os tradutores compreendem seu papel como mediadores culturais, bem como em que medida a tradução pode apresentar um valor estilístico próprio. Sendo assim, o estudo do estilo do tradutor audiovisual pode nos indicar, a partir da sua capacidade criativa e tradutória, seu posicionamento cultural e ideológico e sua função de mediador cultural entre diferentes povos e línguas.

4 Considerações finais

Este capítulo buscou divulgar os recortes teóricos e metodológicos de uma pesquisa de mestrado, ainda em andamento, sobre estilo na tradução de legendas, junto ao POSTRAD na UnB. A intenção foi compartilhar experiências vivenciadas em um momento de escolhas de rumos da investigação. Num primeiro momento, fizemos uma aproximação teórica dos estudos de estilo para o âmbito da legendagem, e, num segundo momento, foram apresentados os *corpora* desta pesquisa e apontamentos teóricos e metodológicos sobre estilo dentro dos ETBC.

Os *corpora* aqui utilizados estão sendo construídos e ainda serão manipulados. Não obstante, já se toma como utilização o WST® como suporte metodológico para investigação do estilo. A expectativa é que seja possível identificar de quais recursos estilísticos, sejam linguísticos, sejam tradutórios, as legendistas fizeram uso, para então avaliar as escolhas por elas feitas na tentativa de determinar seu estilo.

Acreditamos que os padrões estilísticos a serem identificados nos *corpora* podem indicar a criatividade das legendistas e seus posicionamentos culturais e

[23] "[…] *it is only worthwhile if it tells us something about the cultural and ideological positioning of the translator, or of translators in general, or about the cognitive processes and mechanisms that contribute to shaping our translational behaviour"* (BAKER, 2000, p. 258).

ideológicos, como demarcou Baker (2000). Além disso, também podem desvelar a sua arte tradutória de mediadoras culturais, conforme referiu Saldanha (2011a). A intenção com esta pesquisa é, portanto, verificar quais os espaços as legendistas, diante de todas as limitações e variáveis impostas, ocupam nas legendas que produzem.

Agradecimentos

À CAPES, por concessão de bolsa de pós-graduação para realização da pesquisa de mestrado apresentada neste capítulo. Ao POSTRAD/UnB, pelo custeio da viagem para participação do autor nos IX EBRALC e XIV ELC em São Leopoldo (RS), realizados de 15 a 18 de agosto de 2017. Agradecemos às legendistas Carla Prado e Karina Alves por sua participação na pesquisa.

Referências

BAKER, M. *Corpora* in translation studies: an overview and some suggestions for future research. *Target*, Amsterdam, v. 7, n. 2, p. 223-243, 1995.

_____. *Corpus*-based translation studies: the challenges that lie ahead. In: SOMERS, H. (Ed.). *Terminology, LSP and translation:* studies in language engineering in honour of Juan C. Sager. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1996. p. 177-186.

_____. The role of *corpora* in investigating the linguistic behaviour of professional translators. *International Journal of* Corpus *Linguistics*, v. 4, n. 2, p. 281-298, 1999.

_____. Towards a methodology for investigating the style of a literary translator. *Target*, Amsterdam, v. 12, n. 2, p. 241-266, 2000.

BAÑOS, R.; BRUTI, S.; ZANOTTI, S. *Corpus* Linguistics and audiovisual translation: in search of an integrated approach. *Perspectives: Studies in Translatology*, v. 21, n. 4, p. 483-490, 2013.

BARCELLOS, C. P. Estudo de caso sobre a relação entre características dos textos traduzidos e estilo da tradução. *ARTEFACTUM* – Revista de Estudos em Linguagem e Tecnologia, Rio de Janeiro, n. 1, p. 1-12, 2016a.

_____. *Estilo da tradução, convencionalidade e mudanças na tradução:* um estudo de caso sobre os padrões de escolhas do tradutor Paulo Henriques Britto. 196f. Tese (doutorado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Geral, Belo Horizonte, 2016b.

BERBER SARDINHA, T. *Lingüística de* Corpus. Barueri: Manole, 2004.

_____. *Pesquisa em Lingüística de* Corpus *com Word Smith Tools*. Campinas: Mercado das Letras, 2009.

BOASE-BEIER, J. Stylistics and translation. In: GAMBIER, Y.; VAN DOORSLAER, L. *Handbook of translation studies.* Amsterdam/Birmingham: John Benjamins Publishing Company, 2011, p. 153-156.

_____. *Stylistic Approaches to Translation*. Manchester: St. Jerome, 2014.

CAMARGO, D. C. *Metodologia de pesquisa em tradução e linguística de* corpus. São Paulo: Cultura Acadêmica / São José do Rio Preto: Laboratório Editorial do IBILCE/UNESP, 2007.

CHAUME, F. Film Studies and Translation Studies: two disciplines at stake in audiovisual translation. *Meta: Translators' Journal,* Montréal, v. 49, n. 1, p. 12-24, 2004.

DÍAZ CINTAS, J.; REMAEL, A. *Audiovisual translation:* subtitling. Manchester: St. Jerome, 2007.

DÍAZ CINTAS, J. In search of a theoretical framework for the study of audiovisual translation. In: ORERO, P. (Org.). *Topics in audiovisual translation.* Amsterdam/Birmingham: John Benjamins Publishing Company, 2004. p. 21-34.

_____. Back to the future in subtitling. In: MUTRA: CHALLENGES OF MULTIDIMENSIONAL TRANSLATION, 1, 2005, Saarbrücken. *Conference proceedings.* Saarbrücken: Saarland University, 2005, p. 1-17. Disponível em: <http://www.euroconferences.info/proceedings/2005_Proceedings/2005_DiazCintas_Jorge.pdf>. Acesso em: 22 fev. 2017.

_____. The technology turn in subtitling. In: INTERNATIONAL MAASTRICHT-ŁÓDŹ DUO COLLOQUIUM ON "TRANSLATION AND MEANING", PART 9, 5, 2010, Maastricht. *Proceedings…* Maastricht: Maastricht School of Translation and Interpreting Zuyd University of Applied Sciences, 2013, p. 119-132.

_____. *Teoría y práctica de la subtitulación:* inglés-español. Barcelona: Ariel, 2003.

GAMBIER, Y. Introduction. Screen transadaptation: perception and reception. *The Translator*, v. 9, n. 2, p. 171-189, 2003.

GEORGAKOPOULOU, P. Subtitling for the DVD industry. In: DÍAZ CINTAS, J.; ANDERMAN, G. (Org.). *Audiovisual translation:* language transfer on screen. Great-Britain: Palgrave Macmillan, 2009, p. 21-36.

GOTTLIEB, H. Subtitling: diagonal translation. *Perspectives: Studies in Translatology*, v. 2, n.1, p. 101-121, 1994.

_____. Subtitling. In: BAKER, M. (Ed.). *Routledge Encyclopedia of Translation Studies.* London/New York: Routledge, 1998, p. 244-248.

_____. Multidimensional translation: Semantics turned Semiotics. In: MUTRA: CHALLENGES OF MULTIDIMENSIONAL TRANSLATION, 1, 2005, Saarbrücken. *Conference proceedings...* Saarbrücken: Saarland University, 2005a, p. 1-29. Disponível em: <http://www.euroconferences.info/proceedings/2005_Proceedings/2005_Gottlieb_Henrik.pdf>. Acesso em: 22 fev. 2017.

_____. Texts, translation and subtitling – in theory, and in Denmark. In: _____. (Ed.). *Screen Translation. Eight studies in subtitling, dubbing and voice-over*. Copenhagen: University of Copenhagen, 2005b, p. 1-40.

MALMKJAER, K. What happened to God and the angels: an exercise in translational stylistics. *Target*, Amsterdam/Birmingham, v. 15, p. 37-58, 2003.

_____. Translational stylistics: Dulcken's translations of Hans Christian Andersen, *Language and Literature*, London/Thousand Oaks/New Delhi, v. 13, n. 1, p. 13-24, 2004.

MALMKJÆR, K.; CARTER, R. Stylistics. In: MALMKJÆR, K. (Ed.). *The Linguistics Encyclopedia.* 2. ed. London and New York: Routledge, 2002, p. 510-520.

NAVES, S. B. et al. *Guia para produções audiovisuais acessíveis.* Brasília: Ministério da Cultura/Secretaria do Audiovisual, 2016.

NETFLIX. *Timed text style guide:* general requirements. Scotts Valley: Netflix, 2016a. Disponível em: <https://backlothelp.netflix.com/hc/en-us/articles/215758617-Timed-Text-Style-Guide-General--Requirements>. Acesso em: 03 nov. 2017.

_____. *Brazilian Portuguese timed text style guide.* Scotts Valley: Netflix, 2016b. Disponível em: <https://backlothelp.netflix.com/hc/en-us/articles/215600497-Brazilian-Portuguese-Timed-Text--Style-Guide>. Acesso em: 03 nov. 2017.

IVARSSON, J.; CARROLL, M. *Subtitling.* Simrishamn: TransEdit, 1998.

LEECH, G.; SHORT, M. *Style in fiction:* a linguistic introduction to English fictional prose. Harlow: Pearson/Longman, 2007.

OLIVEIRA, A. R. Equivalência: sinônimo de divergência. *Cadernos de Tradução*, Florianópolis, v. 1, n. 19, p. 97-114, 2007. Disponível em: <https://periodicos.ufsc.br/index.php/traducao/article/view/6994>. Acesso em: 16 out. 2017.

PEREGO, E. Evidence of explicitation in subtitling: towards a categorization. *Across Languages and Cultures*, Budapeste, v. 4, n. 1, p. 63-88, 2003.

_____. The codification of nonverbal information in subtitled texts. In: DÍAZ CINTAS, J. (Org.). *New trends in audiovisual translation.* Bristol: Multilingual Matters, 2009, p. 58-69.

PYM, A. *On translator ethics*: principles for mediation between cultures. Amsterdam: Benjamins, 2012.

_____. *Method in translation history*. Manchester: St. Jerome, 1998.

SALDANHA, G. Translator Style: methodological considerations. *The Translator*, v. 17, n. 1, p. 25-50, 2011a.

_____. Style of translation: the use of foreign words in translations by Margaret Jull Costa and Peter Bush. In: KRUGER, A.; WALLMACH, K.; MUNDAY, J. (Ed.). Corpus *Based Translation Studies*: Research and Applications. London/New York: Continuum, 2011b, p. 237-258.

SANTAMARIA GUINOT, L. *Subtitulació i referents culturals.* La traducció com a mitjà d'adquisició de representations socials. 2001. Tese (doutorado). Departamento de Tradução e de Interpretação, Universidade Autônoma de Barcelona, Barcelona, 2001.

Elaboração de um protótipo de glossário bilíngue (português-inglês) de treinamento de força: subsídios para o tradutor

Development of a prototype bilingual (Portuguese-English) glossary of strength training: an aid for translators

Márcia dos Santos Dornelles
Maria José Bocorny Finatto

Resumo: Este artigo sintetiza os principais pontos de nossa pesquisa de mestrado (DORNELLES, 2015), na qual produzimos um protótipo de glossário bilíngue da terminologia do Treinamento de Força (TF), uma subárea da Educação Física. O glossário, na direção português-inglês, foi especialmente desenvolvido para tradutores. As bases teóricas da pesquisa foram a Teoria Comunicativa da Terminologia e a Linguística de *Corpus*. Caracterizamos as bases teóricas e a metodologia utilizadas na elaboração do protótipo considerando que essa experiência possa ser replicada por outros investigadores. Ainda, discutimos as dificuldades e as soluções encontradas no trabalho terminográfico. Por fim, apresentamos os principais resultados da pesquisa, amostras das partes que compõem o glossário e uma breve descrição do comportamento da terminologia do TF, nas duas línguas.

Palavras-chave: Teoria Comunicativa da Terminologia. Linguística de *Corpus*. Terminografia bilíngue. Treinamento de Força.

Márcia dos Santos Dornelles – Servidora da Universidade Federal do Rio Grande do Sul, mestre em Letras pela UFRGS – marcia@esef.ufrgs.br.
Maria José Bocorny Finatto – Professora do Dep. de Linguística, Filologia e Teoria Literária da UFRGS, doutora em Letras pela UFRGS, docente do PPG-Letras-UFRGS – mariafinatto@gmail.com.

Abstract: This article synthesizes the main points of our Master's research (DORNELLES, 2015), in which we developed a prototype bilingual glossary of Strength Training (ST), a subarea of Physical Education. This Portuguese-English glossary was specially designed for translators. The research followed the foundations of the Communicative Theory of Terminology and *Corpus* Linguistics. We describe these theoretical foundations and the methodology used in building the prototype considering that this experience can be replicated by other researchers. Furthermore, we discuss the difficulties and solutions found in the terminographical work. Finally, we present the main results of the study, including samples of the integral parts of the glossary and a brief description of the behavior of the ST terminology, in both languages.

Keywords: Communicative Theory of Terminology. *Corpus* Linguistics. Bilingual Terminography. Strength Training.

1 Introdução

Este artigo busca sintetizar os principais pontos de nossa pesquisa de mestrado (DORNELLES, 2015), na qual apresentamos um protótipo de glossário bilíngue da terminologia do Treinamento de Força (TF), na direção português-inglês, especialmente desenvolvido para uso de tradutores. Publicar esta síntese neste livro visa a ampliar a divulgação da pesquisa e a possibilitar que outros investigadores – inclusive profissionais de Tradução – repliquem nossa experiência, em outras áreas do conhecimento, naquilo que julgarem ser mais adequado e pertinente. Do mesmo modo, buscamos dar a conhecer, especialmente para a comunidade de interessados em Linguística de *Corpus* do Brasil, o modelo de tratamento terminológico que adotamos.

As bases teóricas da pesquisa foram a Teoria Comunicativa da Terminologia e a Linguística de *Corpus*. Assim, aqui discutimos algumas das dificuldades com as quais nos deparamos e apresentamos as decisões por nós tomadas nas diferentes etapas da elaboração do glossário. Além disso, trazemos uma breve descrição do comportamento da terminologia do TF identificada, nas duas línguas.

O terminógrafo[1], ao elaborar um glossário terminológico bilíngue, baseado em *corpus* e direcionado a tradutores, deve preocupar-se não só em repertoriar, nas duas línguas, os termos próprios de uma (sub)área do conhecimento, mas também em apresentá-los inseridos em suas combinatórias típicas, ou seja, associados aos elementos que a eles se combinam em nível sintagmático, de forma recorrente nos textos daquela especialidade. Isso porque o tradutor precisa produzir um texto de chegada adequado ao padrão de linguagem em foco, de forma a espelhar o *modus dicendi* daquele campo. Assim, seu texto, com as terminologias devidamente

[1] Para mais detalhes sobre estudos e pesquisas em Terminologia e Terminografia, sugerimos acessar <http://www.ufrgs.br/termisul/>, especialmente a parte <http://www.ufrgs.br/termisul/publica.php>.

inseridas em "fraseamentos" convencionalizados, soará natural à comunidade de leitores, evitando-se ruídos na comunicação.

Para tanto, assim como um biólogo precisa explorar o meio em que vive seu espécime de estudo para entender o comportamento deste, também o terminógrafo precisará conhecer o *habitat* das terminologias com que lida: o texto especializado. Nesse sentido, o conhecimento das propriedades do gênero textual em estudo qualifica um produto terminográfico, considerando que os termos e demais elementos a ele incorporados, como as fraseologias especializadas, os contextos definitórios e os exemplos de uso, extraídos de seu âmbito natural de emprego, ajudam a compor os modos de dizer desse gênero. Dessa forma, um produto terminográfico terá as melhores chances de ser bem aceito pela comunidade de usuários tradutores. Em síntese, se o gênero textual e discursivo é um elemento condicionante do perfil das terminologias – e de seu uso –, então seus traços "textuais" também devem estar contemplados no produto oferecido ao profissional tradutor.

Com esses pressupostos e diante da falta de produtos terminográficos bilíngues no âmbito do TF, especialmente dirigido a tradutores brasileiros, nossa pesquisa de mestrado (DORNELLES, 2015) teve como objetivo central apresentar bases teórico-metodológicas consistentes para a elaboração de um glossário específico de TF na direção português→inglês, destinado especialmente a tradutores. Entretanto, acreditamos que ele também seja útil para pesquisadores e estudantes dessa temática que precisem produzir artigos científicos em inglês.

Outros objetivos do nosso trabalho de mestrado, todos alcançados, foram (a) oferecer um protótipo do glossário, composto de guia do usuário, uma árvore de domínio em português do TF, lista de termos em português e 30 exemplares de fichas terminológicas em formato estendido; e (b) oferecer uma descrição do comportamento das unidades terminológicas em português e inglês, e das unidades fraseológicas especializadas (UFEs) eventivas (BEVILACQUA, 2003; 2004) em português em artigos sobre TF.

Nosso *corpus* de estudo foi constituído de 70 artigos de periódicos científicos de destaque no âmbito do TF, metade escrita originalmente em português e metade originalmente em inglês. São, portanto, dois *subcorpora* comparáveis. Para exploração e análise do *corpus*, utilizamos o *software* livre AntConc (ANTHONY, 2011).

2 Bases e concepções teóricas de partida

Conforme referimos, nossa pesquisa apoiou-se nos princípios da Teoria Comunicativa da Terminologia (TCT) e nos fundamentos e diretrizes da Linguística de *Corpus* (LC). Seguir a TCT (CABRÉ, 1999a; 1999b; 2001a; 2001b; 2003; 2009) implica adotar as terminologias como objeto central de

estudo e concebê-las, antes de tudo, como unidades lexicais da língua natural que adquirem valor especializado dentro de um contexto especializado, segundo critérios semânticos, discursivos e pragmáticos. Assim, na TCT, uma terminologia é, antes de tudo, um valor de significação ativado em meio a um discurso.

A LC (BERBER SARDINHA, 2004; BIBER, 2012) tem como principal fundamento "a visão da linguagem como sistema probabilístico, [a qual] pressupõe que, embora muitos traços linguísticos sejam possíveis teoricamente, não ocorrem com a mesma frequência" (BERBER SARDINHA, 2004, p. 30-31). Assim, ao explorarmos um *corpus* à luz da LC, será possível depreender como se dá, no tocante a padrões de uso, essa "ativação do valor de termo".

Ora, em Terminologia, trabalhar com probabilidades em vez de possibilidades faz todo sentido. Afinal, somente após a aceitação e a *repetição* de um candidato a termo pelos especialistas do campo é que ele adquire valor especializado e é incorporado à terminologia desse campo. Em outras palavras, as terminologias passam a ser reconhecidas à medida que se estabelecem em usos recorrentes, adotados pela comunidade discursiva. Portanto, conforme Finatto (2014b, p. 453), "se a condição terminológica é um valor ativado pelos discursos/textos, como se defende na TCT, espera-se depreender os traços constitutivos desse valor ao longo de diferentes textos/*corpora*". É a análise da distribuição e, principalmente, da repetição dessas terminologias e combinatórias e de seus traços que permite ao terminólogo descrever padrões de emprego de termos e de unidades fraseológicas.

Nosso protótipo de glossário repertoriou unidades terminológicas (UT), monolexicais ou polilexicais, e destacou, em campo próprio na microestrutura das fichas, unidades fraseológicas especializadas (UFEs) eventivas. São exemplos de UFEs incluídas no protótipo de glossário *prescrição do treinamento de força, executar séries, repetições completadas*. Essas fichas, por sua vez, correspondem às informações completas sobre um dado termo e são utilizadas para embasar a apresentação de verbetes para um glossário ou dicionário.

As UFEs eventivas são formadas necessariamente por um núcleo terminológico e um núcleo eventivo; sendo esse núcleo "eventivo" assim denominado por ser constituído ou derivado de verbo (verbo, nominalização ou particípio) e denotar processos e ações próprios de uma área de conhecimento ou temática (BEVILACQUA, 2004). Esse último aspecto é o que destaca as UFEs eventivas no âmbito dos artigos científicos do TF em detrimento de outras combinatórias identificadas, motivo pelo qual escolhemos esse tipo de fraseologia para especialmente compor nosso modelo de glossário.

E o que têm em comum uma UT e uma UFE eventiva? De acordo com os pressupostos da TCT, ambas são estruturas integrantes do sistema da língua, portadoras de conhecimento específico de uma área ou temática especializada, e utilizadas em uma situação comunicativa especializada. Em decorrência de serem,

antes de tudo, signos da língua natural, também ambas são suscetíveis a toda gama de fenômenos que nesta ocorrem, dentre eles a variação conceitual (polissemia) e denominativa (sinonímia), considerando a essência comunicativa e discursiva dessas unidades. Assim, é possível imaginar que um segmento como *executar séries* possa ser encontrado em um artigo científico de TF em português, por exemplo, como *realizar séries,* o que nos alerta para um caso de variação de construção terminológica. Casos como esse fazem com que um tradutor se questione sobre como agir para escolher um equivalente na língua estrangeira. O profissional, então, procurará certificar-se se a variabilidade *realizar/executar* também existiria na outra língua e num contexto de comunicação correspondente, no mesmo gênero textual.

Sobre esse princípio de variabilidade, amplamente verificado em textos científicos, Faulstich (2001, p. 20) é incisiva: "Variação e terminologia não se confrontam na abordagem atual. Pelo contrário, defendemos que a terminologia é passível de variação porque faz parte da língua, porque é heterogênea por natureza, e porque é de uso social". Para classificar os tipos de variação terminológica incluídos no nosso modelo de glossário, sejam variações de formas de termos ou de formas de suas construções fraseológicas, recorremos à tipologia de Freixa (2002; 2014).

Um estudo, em especial, contribuiu como um guia para a inclusão das UTs nas fichas terminológicas que integram o nosso glossário. Esse trabalho, feito por Maciel (2001), também nos ajudou a organizar o desenho da nossa árvore de domínio. Essa árvore é um esquema hierárquico, semelhante a um organograma, que busca espelhar a organização conceitual de uma (sub)área. Para além do critério básico de frequência e de distribuição da terminologia em um *corpus* de estudo, Maciel introduziu os conceitos de *pertinência temática* e *pertinência pragmática* para a seleção dos itens que devem figurar em um glossário ou obra afim. Unidades com alto grau de especificidade ao âmbito investigado – neste caso o TF – revelam *pertinência temática*. Outras unidades, embora mais específicas de outras áreas e até menos frequentes, estão presentes nos contextos definitórios de termos importantes do âmbito do TF e, por isso, precisam ser incorporadas como itens do produto terminológico para uma melhor compreensão desses conceitos. Essas unidades "correlatas" revelam o que Maciel (2001) denominou *pertinência pragmática*. Vejam-se exemplos dessas categorias de pertinência de termos na subseção 3.3, na descrição da parte III da árvore de domínio.

No tocante às concepções teóricas de tradução para o tratamento da equivalência no protótipo de glossário, partimos da noção de tradução como um ato de comunicação, uma operação entre textos e um processo mental (HURTADO ALBIR, 2008). Levamos em conta, também, as subcompetências tradutórias de que o tradutor lança mão no seu trabalho, a saber, subcompetências bilíngue, extralinguística, de conhecimentos sobre a tradução, instrumental e estratégica, além de componentes psicofisiológicos (PACTE, 2011). Ademais, adotamos a

noção de equivalência funcional de Gémar (1998), para quem são equivalentes as estruturas que expressam a mesma relação semântica e o mesmo efeito pragmático nos textos de partida e de chegada, ou seja, aquelas que "funcionam", em termos comunicativos, de forma equivalente nos dois cotextos.

Por fim, dentre outros referenciais teóricos e metodológicos que nortearam as decisões tomadas nas diferentes etapas de elaboração do nosso modelo de glossário, destacamos os manuais de Terminologia de Barros (2004) e de Krieger e Finatto (2004); e os trabalhos terminográficos de Fromm (2007), Teixeira (2008), Almeida (2000), e Silva e Teixeira (2010). Suas contribuições são mencionadas ao longo da próxima seção.

3 Materiais e métodos

Esta seção enumera as etapas de elaboração do protótipo de glossário, expondo dificuldades enfrentadas e soluções encontradas. São descritos os procedimentos adotados para a construção do *corpus* de estudo (3.1); o material de apoio utilizado complementarmente ao *corpus* (3.2); a construção da árvore de domínio (3.3); os recursos utilizados para o reconhecimento das UTs (3.4); os critérios e procedimentos de seleção das UTs para inclusão na árvore de domínio e fichamento (3.5); os critérios e procedimentos para reconhecimento dos equivalentes em inglês (3.6) e das UFEs eventivas (3.7); as informações constantes no Guia do Usuário do Glossário (3.8); a forma de apresentação da lista de termos (3.8); e o desenho de ficha terminológica (3.9). Mais detalhes podem ser conferidos em Dornelles (2015).

3.1 O corpus *de estudo*

Nosso *corpus* de estudo foi constituído de 70 artigos de periódicos científicos em formato digital, nas línguas português brasileiro e inglês (esta sem uma variedade específica), que tratam do assunto Treinamento de Força. Ele é dividido em dois *subcorpora* comparáveis, ou seja, são formados por textos originais nesse par de línguas, publicados no período de 2003 a 2014.

Ambos os *subcorpora* precisaram ser construídos, uma vez que não foram encontrados *corpora* já compilados com textos sobre TF e disponíveis para utilização. As revistas foram recomendadas por nossos consultores especialistas[2] na

[2] A pesquisa contou com o auxílio de dois docentes da Escola de Educação Física, Fisioterapia e Dança da UFRGS, doutores em Treinamento de Força, que participaram como consultores em algumas etapas da elaboração do nosso protótipo de glossário. São eles os professores Ronei Silveira Pinto e Eduardo Lusa Cadore.

área, com base na sua relevância no campo investigado, considerando o estrato de classificação na área de Educação Física no sistema WebQualis da CAPES e o fator de impacto na base estatística Journal Citation Reports®.

Consideramos que é um *corpus* equilibrado, pois os artigos que o compõem fazem um apanhado dos principais tópicos abordados no âmbito da pesquisa científica sobre TF, no Brasil e no exterior, na última década. Além disso, ainda que reúna somente artigos científicos, as revistas e os autores dos textos em português são vinculados a diferentes instituições científicas brasileiras; e as revistas e os autores dos textos em inglês são de diferentes países.

Trata-se de um *corpus* especializado, composto por artigos científicos escritos por especialistas para especialistas ou para estudantes das áreas de Educação Física, Fisioterapia e Medicina do Esporte. Não podemos afirmar se todos os artigos foram escritos por falantes nativos; no entanto, importa neste estudo o fato de que todos os textos foram revisados por pares, o que denota que a linguagem empregada, incluindo a terminologia, foi aceita por representantes dessa comunidade de especialistas, entendida aqui como uma comunidade discursiva.

3.1.1 O *subcorpus* em português

O *subcorpus* em português brasileiro constituiu-se de artigos científicos publicados nas seguintes revistas brasileiras *on-line*:

- *Revista Brasileira de Medicina do Esporte* (RBME), da Sociedade Brasileira de Medicina do Exercício e do Esporte (SBMEE) (ISSN 1517-8692);
- *Revista Brasileira de Cineantropometria & Desempenho Humano* (RBCDH), da Universidade Federal de Santa Catarina (UFSC) (ISSN 1415-8426);
- *Motriz: Revista de Educação Física*, da Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP) (ISSN 1980-6574);
- *Revista Brasileira de Atividade Física e Saúde* (RBAFS), da Sociedade Brasileira de Atividade Física e Saúde (SBAFS) (ISSN 1413-3482);
- *Revista da Educação Física*, da Universidade Estadual de Maringá (UEM) (ISSN 0103-3948);
- *Revista Brasileira de Educação Física e Esporte* (RBEFE) da Universidade de São Paulo (USP) (ISSN 1807-5509).

A busca inicial dos artigos foi feita nas páginas das revistas na base de dados SciELO[3], seguindo os seguintes parâmetros:

[3] Acesso em <http://www.scielo.br/scielo.php/script_sci_serial/pid_1517-8692/lng_pt/nrm_iso>.

a) Inclusão do termo *treinamento de força* nas palavras do título ou no assunto.
b) Nas listas resultantes por revista, selecionamos os artigos publicados a partir de 2002, considerando que a versão brasileira da Terminologia Anatômica Internacional (TAI) foi publicada em 2001 (SBA, 2001). Tal decisão baseou-se no fato de que a TAI é muito utilizada na área da saúde, incluindo a Educação Física, pois arrola os termos da anatomia do corpo humano que designam regiões específicas envolvidas, por exemplo, nos movimentos dos exercícios e nas avaliações antropométricas[4].
c) Essa segunda lista resultou em uma população de 46 artigos e foi submetida ao crivo de nosso consultor, que excluiu, a partir da leitura dos títulos, os textos que pareciam descolar o tema do TF do âmbito da Educação Física, ou seja, não focavam o treinamento em si. Foram eliminados, portanto, artigos de pesquisas com cobaias (animais); artigos com enfoque essencialmente clínico, abordando o TF para reabilitação de doenças graves; e artigos mais bem inseridos no campo da Bioquímica.
d) Aplicados esses critérios, procedemos à conversão dos textos que estavam nos formatos *.pdf ou HTML em arquivos não formatados (*.txt), para fins de exploração do *corpus* no AntConc. Os artigos convertidos com sucesso totalizaram 35 e revelaram-se uma amostra suficientemente representativa dos temas abordados no âmbito do TF. Assim, os textos que não puderam ser convertidos foram desprezados.

3.1.2 O *subcorpus* em inglês

O *subcorpus* em inglês foi constituído de artigos científicos publicados nas seguintes revistas *on-line*, oriundas de diferentes países. Não nos limitamos a uma variedade específica da língua, porque os estudantes e pesquisadores dessa especialidade costumam submeter seus artigos a periódicos de diferentes nacionalidades, levando em conta, muitas vezes, o fator de impacto e/ou o Qualis CAPES dos mesmos na área de Educação Física.

– *The Journal of Strength & Conditioning Research* (JSCR), da National Strength and Conditioning Association (NSCA) (ISSN 1064-8011);
– *Isokinetics and Exercise Science Journal* (IES), da European Interdisciplinary Society for Clinical and Sports Application (EISCSA) (ISSN 0959-3020);

[4] Para saber mais sobre o emprego da TAI no âmbito da Educação Física, ver Dornelles (2014).

– *Medicine & Science in Sports & Exercise* (MSSE), do American College of Sports Medicine (ACSM) (ISSN 0195-9131);
– *International SportMedJournal* (ISMJ), da International Federation of Sports Medicine (FIMS) (ISSN 1528-3356);
– *European Journal of Applied Physiology* (EJAP) (ISSN 1439-6319);
– *European Journal of Sport Science* (EJSS), do European College of Sport Science (ECSS) (ISSN 1746-1391);
– *British Journal of Sports Medicine* (BJSM) do BMJ Group (ISSN 0306-3674);
– *Scandinavian Journal of Medicine & Science in Sports* (SJMSS) (ISSN 0905-7188).

A busca inicial dos artigos foi feita no Portal de Periódicos da CAPES[5], que dá acesso à base de dados *SPORTDiscus with Full Text*, entre outras. A pesquisa dos artigos e as etapas de seleção seguiram parâmetros semelhantes aos dos textos em português, considerando que os *subcorpora* são comparáveis. Filtros de busca:

a) Modo de pesquisa: Booleano, com inclusão do termo *strength+training* (equivalente preferencial do termo *treinamento de força*) no título ou nas palavras-chave ou no assunto (descritores);
b) Texto completo;
c) Data de publicação: janeiro de 2002 a abril de 2014;
d) Analisado por especialistas;
e) País: tudo;
f) Idioma: inglês;
g) Tipo de publicação: *journal article*.
h) A lista resultante, com 140 textos, foi submetida ao crivo de nosso consultor, que excluiu artigos de pesquisas com cobaias (animais); artigos com enfoque essencialmente clínico, abordando o TF para reabilitação de doenças graves; e artigos mais bem inseridos no campo da Bioquímica Básica.
i) Aplicados esses critérios, convertemos os textos em *.pdf ou HTML para *.txt. Os textos que não puderam ser convertidos foram desprezados. Essa etapa resultou em 65 artigos em inglês, quase o dobro do *subcorpus* em português.

Considerando nosso objetivo de utilizar *subcorpora* comparáveis, que requerem uma correspondência quantitativa aproximada entre eles, tanto em número de textos como em número de *tokens* (palavras ou itens), bem como

[5] Acesso em <http://www.periodicos.capes.gov.br/>.

uma correspondência de conteúdo (assuntos), foi preciso proceder a uma última seleção, a saber:

j) tomamos a lista integral dos 35 artigos em português e, para cada um, buscamos, na lista em inglês – bem mais extensa (65 textos) – um artigo que contivesse termos em comum no título, e/ou nas palavras-chave, e/ou nos assuntos (descritores) (ver Quadro 2). Nesse cotejo, os textos restantes em inglês foram desprezados.

3.1.3 Configuração final do *corpus* comparável

A seguir, o Quadro 1 mostra nosso *corpus* de estudo em números.

Quadro 1 – O *corpus* de estudo em números

	N. artigos científicos	**Types (formas)**	**Tokens (itens)**	**Densidade lexical (razão types/tokens)**
Subcorpus Port.	35	7.463	122.502	0,061
Subcorpus Ingl.	35	7.501	164.619	0,045
Corpus completo	70		287.121	

Fonte: Elaborado pelas autoras

O Quadro 2 mostra um extrato da configuração final do *corpus*, com os artigos correspondentes postos lado a lado e os termos em comum salientados em azul. A primeira coluna registra os códigos que atribuímos aos artigos em português e os anos de publicação. A quarta coluna, além desses registros, também informa os países das instituições de origem de seus autores.

3.2 Material de apoio

Maciel (2013, p. 42) faz a seguinte observação sobre a incompletude de todo *corpus* e a necessidade de lançar mão de recursos externos:

> O *corpus* será sempre um material incompleto, um artefato preparado em função dos critérios preestabelecidos pelo pesquisador em vista de seus propósitos. Além disso, o *corpus* não é um material inteiramente neutro; reflete, de um lado, a subjetividade do seu compilador; de outro, as opiniões dos autores dos textos e o pensamento de uma dada comunidade em determinada época. Os *corpora* estão sujeitos aos efeitos da passagem do tempo e da evolução das ideias; nenhum deles contém tudo o que é necessário para entender a área ou dominar-lhe um recorte. **Diante da impossibilidade de abarcar a plenitude da informação, construir ou compilar uma terminologia exige também recursos externos.** Esse aspecto não

invalida a utilização de *corpora*, mas alerta para sua relativização que se agrava pelo confronto da interpretação do significado instaurado no texto e da estabilidade do termo no sistema conceitual da respectiva área. (Grifo nosso.)

Quadro 2 – Extrato da lista dos artigos comparáveis português-inglês

Cód. art. port./ ano public.	Títulos dos artigos por revista brasileira	Cód. art. ing./ ano public./ países origem	Títulos dos artigos comparáveis em inglês
1. Revista Brasileira de Medicina do Esporte (ISSN 1517-8692)			
RBME 02 2010	A percepção de esforço no treinamento de força. **Palavras-chave:** índice de esforço percebido, treinamento de força, exercício resistido.	JSCR 12 2005 Brasil e EUA	Influence of exercise order on the number of repetitions performed and perceived exertion during resistance exercises. **Palavras-chave:** strength; strength training; exercise; Borg scale. **Assuntos:** exercise; isometric exercise; physical education; physical fitness; hygiene; sports sciences.
RBME 03 2010	Efeitos de 24 semanas de treinamento resistido sobre índices da aptidão aeróbia de mulheres idosas. **Palavras-chave:** envelhecimento, capacidade aeróbia, treinamento de força, sarcopenia.	SJMSS 02 2010 **Revisão** Dinamarca	Role of the nervous system in sarcopenia and muscle atrophy with aging; strength training as a countermeasure. **Palavras-chave:** motor neurons; CNS; muscle power; RFD; strength training **Assuntos:** physical fitness; cytokines; cell death; motor neurons; growth factors; oxidative stress, muscle strength; strength training; isometric exercise; immune response -- regulation; hypoglycemic agents.
RBME 04 2010	Influência do estado de treinamento sobre o comportamento da pressão arterial após uma sessão de exercícios com pesos em idosas hipertensas. **Palavras-chave:** hipotensão pós-exercício; envelhecimento; treinamento de força.	IES 05 2005 Bélgica	Influence of different resistive training modalities on blood pressure and heart rate responses of healthy subjects. **Palavras-chave:** hemodynamics; blood pressure; resistive exercise; strength training. **Assuntos:** isometric exercise; exercise; rehabilitation; blood pressure; heart beat; muscles; physical education; health.
RBME 05 2010	Efeito de 12 semanas de treinamento com pesos sobre a força muscular, composição corporal e triglicérides em homens sedentários. **Palavras-chave:** treinamento de força; triglicerídeo plasmático, percentual de gordura.	SJMSS 08 2007 **Revisão** Suécia	Strength training effects of whole-body vibration? **Palavras-chave:** systematic review; exercise; muscle force; muscle strength; oscillation; neural mechanisms; jumping; strength training. **Assuntos:** muscle strength; weight training; medical literature; physical fitness testing; jump & reach tests; vibration (mechanics) -- research; library information networks; reviews.
RBME 07 2009 **Revisão**	Força muscular *versus* pressão arterial de repouso: uma revisão baseada no treinamento com pesos **Palavras-chave:** resposta cardiovascular; hipertensão; treinamento de força.	EJAP 01 2013 Japão e EUA	Effects of high-intensity and blood flow-restricted low-intensity resistance training on carotid arterial compliance: role of blood pressure during training sessions. **Palavras-chave:** arterial stiffness; strength training; vascular occlusion; muscle hypertrophy. **Assuntos:** strength training; muscle strength; carotid artery; systolic blood pressure, regional blood flow; vascular resistance.

Fonte: Elaborado pelas autoras

Barros (2004, p. 202) já destacava a necessidade da utilização de "textos de apoio, que servem para a complementação de informações" – para além de um *corpus* de estudo que se reúne, sobretudo quando se pretende alimentar um glossário de termos. No tocante às fontes de documentação sobre os termos, a autora explica que

> Além da busca de dados de cunho semântico-conceptual sobre as unidades de tratamento (os termos), o terminólogo pode vir a ter outros tipos de necessidade, tais como encontrar alternativas de denominação, conhecer a equivalência em outras línguas, resolver dúvidas sobre o comportamento gramatical da unidade terminológica na língua em questão e até saber se já existem levantamentos terminológicos feitos sobre um determinado domínio. (BARROS, 2004, p. 206)

Em face dessas constatações, passamos a descrever nosso material de apoio.

3.2.1 Livros-texto

O contexto ou enunciado definitório, de acordo com Finatto (2003, p. 198-199), "é um elemento-chave na constituição e na veiculação do conhecimento especializado, tecnológico ou científico, uma vez que expressa um segmento

de relações de significação de uma dada área do saber". Além disso, na condição de textos particularizados, as definições "revelam facetas de compreensão de fenômenos no seio de uma determinada ciência" (FINATTO, 2003, p. 199).

Pearson (2004 [1999]), analisando a densidade de contextos definitórios em diferentes tipos de comunicação técnico-científica, já alertava que, nos textos redigidos de especialistas para especialistas, como é o caso do nosso *corpus*, há "uma densidade muito alta de termos, mas provavelmente muito poucos elementos definitórios. A explicação é simples: supõe-se que o leitor conhece e entende os termos utilizados" (p. 55). Barros (2004, p. 209), a propósito, complementa: "As obras de cunho didático ou explicativo são, em geral, de grande auxílio ao terminólogo, uma vez que costumam ter uma preocupação em expor de modo claro os conceitos e a terminologia do domínio".

Cientes disso, como material de apoio para a elaboração da árvore de domínio do TF e para a composição dos contextos definitórios dos termos nas fichas terminológicas do protótipo de glossário, utilizamos livros-textos da área de Educação Física que abordam o TF. Algumas dessas obras foram traduzidas do inglês ao português e revisadas por especialista da área. Quase todas são referências utilizadas na disciplina de TF em cursos de Educação Física no Brasil. Esse material não foi digitalizado, e as consultas foram feitas por manuseio.

Nas obras traduzidas ao português, duas seções muito úteis foram o Sumário e o Índice. O Sumário auxiliou na organização hierárquica da árvore de domínio, uma vez que apresenta os tópicos de estudo seguindo certa ordem conceitual. O Índice, pelo fato de incluir boa parte da terminologia da área de forma organizada e isolada, seguida das páginas em que o assunto é tratado, auxiliou na confirmação do valor terminológico das unidades e no reconhecimento de UTs não ocorrentes no *corpus* de estudo.

3.2.2 Artigos científicos de referência

Artigos de referência na área, cinco em português e três em inglês, foram indicados por nossos consultores. Alguns deles trazem posicionamentos oficiais de entidades ligadas ao campo do TF, bem como fundamentos, terminologia e procedimentos dessa especialidade, que são úteis para o entendimento dos conceitos e para a redação de enunciados definitórios das UTs nas fichas.

3.2.3 Glossário particular

Outro material consultado foi um glossário particular preexistente inglês-português de Educação Física, que foi construído com termos compilados dos livros supracitados e de vários outros textos acadêmicos ao longo de anos de

tradução na área. Assim, conforme se poderá perceber, é possível aproveitar a experiência prévia de um tradutor em equipes que produzam esse tipo de obra, sem contar o fato de os próprios profissionais poderem transformar seu conhecimento em publicações específicas, que também os não tradutores podem aproveitar. Esse glossário "particular" mostrou-se útil especialmente para a consulta de equivalentes em inglês para as UTs selecionadas. Encontrados os equivalentes, estes foram pesquisados e confirmados no *subcorpus* em inglês no AntConc.

3.2.4 Terminologia Anatômica Internacional (TAI)

Barros (2004, p. 208) observa que "Em muitos campos a terminologia empregada já passou por um processo de normalização. Nesses casos, é fundamental consultar os organismos de classe para verificar se já não existem vocabulários normalizados no domínio". Nesse sentido, como fonte de consulta dos termos da anatomia do corpo humano, utilizamos a versão brasileira da TAI (SBA, 2001), traduzida ao português pela Comissão de Terminologia Anatômica (CTA) da Sociedade Brasileira de Anatomia (SBA) e publicada em 2001 pela Editora Manole. Ainda que nosso protótipo de glossário não tenha fins normatizadores, como é o caso da TAI, reconhecemos a importância desse trabalho elaborado por profissionais da área médica e registramos nas fichas terminológicas os termos recomendados pela SBA.

No entanto, como nossa pesquisa é baseada no *corpus*, ou seja, no uso real da terminologia, quando uma UT recorrente continha um termo anatômico discordante da norma de uso oficial prescrita pela TAI, ela também foi registrada na ficha, seguida de um asterisco (*), e foi feita uma observação nesse sentido na Nota. Entre as duas formas – a normatizada e a não normatizada –, foi privilegiada e lematizada – isto é, transformada em entrada do glossário – aquela com maior distribuição no *corpus* de estudo e, em caso de empate, com maior frequência.

3.2.5 Acordo Ortográfico da Língua Portuguesa (AOLP) e Vocabulário Ortográfico da Língua Portuguesa (VOLP)

Com o mesmo propósito da consulta à TAI (SBA, 2001), recorremos também ao AOLP (1990) e ao VOLP (ABL, 2009). Quando uma UT variante apresentou uma grafia divergente da normatizada nessas fontes, ela foi registrada com um asterisco (*); e, na *Nota* explicativa, fizemos uma observação nesse sentido. Aqui mais uma vez, não deixamos de registrar quaisquer variantes e de alertar nosso usuário sobre elas.

3.2.6 Google Acadêmico

O Google Acadêmico (GOOGLE INC., 2011) foi especialmente útil em alguns casos em que não foram encontrados, no *subcorpus* em inglês, equivalentes de UFEs eventivas extraídas do *subcorpus* em português. Nesses casos, ele serviu para a extração de exemplos desses usos fraseológicos em inglês, em fontes de reconhecida qualidade. A busca foi feita pelo gênero textual artigo científico e pelo mesmo período de publicação (2003 a 2014) do *corpus* de estudo.

3.2.7 Wikipédia

Essa enciclopédia livre eletrônica[6] foi utilizada para ajudar a compor a definição simplificada ou a nota explicativa de algumas UTs nas fichas terminológicas. Submetemos os textos extraídos à validação dos consultores especialistas. Em alguns casos, fizemos ajustes; em outros, utilizamos a citação direta com referência à fonte.

3.3 A árvore de domínio do Treinamento de Força

Em Terminologia, basicamente o que distingue um termo de um não termo, para além da simples recorrência em um *corpus* especializado, é a sua "função essencialmente referencial dentro de um sistema de conceitos" (MACIEL, 2001, p. 276). De acordo com a norma técnica ISO 1087[7] de 1990, que traz diretrizes gerais ou relativas à composição de glossários e à organização do trabalho de pesquisa terminológica, um sistema de conceitos é um "conjunto estruturado de conceitos construído com base nas relações estabelecidas entre esses conceitos e no qual cada conceito é determinado por sua posição nesse conjunto" (ISO 1087, 1990, p. 4).

Barros (2004) salienta a importância da organização desse sistema de conceitos em diversas etapas da elaboração de obras terminográficas: na escolha da nomenclatura, no tratamento dos dados, na organização do sistema de remissivas, no aprofundamento da pesquisa terminológica, entre outras.

Em nossa pesquisa de mestrado, optamos por representar nossa "leitura" da organização do domínio do TF com uma árvore de domínio. A vantagem desse formato é a facilidade de visualização das divisões e conexões feitas; a desvantagem é a dificuldade para sua elaboração em uma página, seja em versão impressa ou eletrônica. Na definição de Krieger e Finatto (2004, p. 134),

[6] Acesso em: <https://pt.wikipedia.org/wiki/Brasil>.
[7] Essa norma foi substituída pelas normas ISO 1087-1 (2000) e ISO 5127 (2001). No entanto, não tivemos acesso a elas.

> Uma árvore de domínio é um diagrama hierárquico composto por termos-chave de uma especialidade, semelhante a um organograma. [...] Esse tipo de esquema pretende apenas servir como uma organização possível para uma especialidade ou ciência, de modo que o pesquisador possa, baseado nele, compreender algumas de suas hierarquias básicas e também situar um recorte do reconhecimento terminológico para seu dicionário.

Como oferecemos apenas um protótipo de glossário, nossa árvore, longe de ser exaustiva, constitui uma referência que dá conta da amostra de UTs lematizadas nos exemplares de fichas. A árvore foi construída a partir de suas "raízes, tronco e galhos mais grossos", e suas ramificações – que podem ser infinitas conforme a evolução do conhecimento no âmbito do TF. Para uma etapa futura, na qual possamos ampliar o protótipo, em uma edição eletrônica do glossário, a árvore de domínio poderia ser substituída por uma lista sistemática. Para a seleção das UTs que compõem a árvore, descrevemos os procedimentos e critérios adotados na subseção 3.5.

As principais relações conceituais hierárquicas estabelecidas em nossa árvore são a relação genérica ("tipo de") e a relação partitiva ("parte de"), que interferem diretamente na elaboração da definição terminológica. Segundo Barros (2004, p. 116-117), sob a ótica da semântica, a relação genérica equivale à relação hiperonímica-hiponímica: o conceito mais genérico é o *hiperônimo*; os mais específicos são os *hipônimos*; e estes, quando pertencem ao mesmo nível de abstração dentro de um sistema estruturado – neste caso a árvore –, são *co-hipônimos*. Já na relação partitiva, têm-se a relação holonímia-meronímia: a noção superordenada, ou integrante, é o *merônimo*; e a noção subordinada, ou partitiva, é o *holônimo*. Na nossa árvore, outras relações estabelecidas são "aplicação/utilização em". Todas essas relações estão representadas na legenda logo abaixo da árvore.

Para o desenho da árvore, optamos pelo programa Microsoft Office Word (MICROSOFT CORPORATION, 2006), no qual usamos a funcionalidade "inserir formas". Ainda que não seja uma ferramenta específica para esse fim[8], nós a escolhemos pela estética de apresentação. Para driblar a falta de espaço, dividimos a árvore em três partes (Figuras 1 a 3).

Conforme explica a legenda abaixo de cada parte da árvore, os termos incluídos em nosso protótipo de glossário estão numerados e salientados com fundo lilás (▭). Já projetando uma versão eletrônica do glossário, esses termos remetem, via *hiperlink*, à ficha terminológica correspondente; e vice-versa. Os termos numerados com fundo incolor e contorno em linha contínua (▭) serão acrescentados futuramente. Já as células numeradas com fundo incolor e contorno com linha tracejada (▭) encerram palavras ou sintagmas que, segundo nossa avaliação, não possuem valor terminológico, pois não carregam conhecimento especializado; ainda assim, a numeração foi mantida para demonstrar as relações

[8] Um *software* específico para esse fim e gratuito é o CmapTools, disponível em: <http://cmap.ihmc.us/>.

hierárquicas entre os conceitos. As células contendo três pontos demarcam a expansão do conhecimento.

Nossa maior dificuldade na arquitetura da árvore como um todo foi sistematizar de forma hierárquica algumas UTs que, a nosso ver, poderiam situar-se em mais de um ramo da estrutura. Por isso, partes do seu desenho foram refeitas várias vezes. Nesse sentido, reiteramos, a árvore configura *uma* organização possível do domínio do TF, não podendo ser considerada um fim em si mesmo.

A parte I da árvore (Figura 1) oferece uma visão macroestrutural do conhecimento no âmbito do TF. Na parte superior, mostramos onde o TF se insere como *disciplina* ou *temática de estudo* na área de Educação Física, a qual integra a grande área das Ciências da Saúde. Consultamos as áreas que compõem as Ciências da Saúde no *site* do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)[9]. Adotamos os núcleos de conhecimento que integram a área de Educação Física no Plano Pedagógico do curso de Licenciatura em Educação Física da UFRGS[10]. Nossa representação também foi apresentada ao nosso consultor especialista e considerada válida para o fim que estabelecemos: subsidiar o tradutor usuário do glossário.

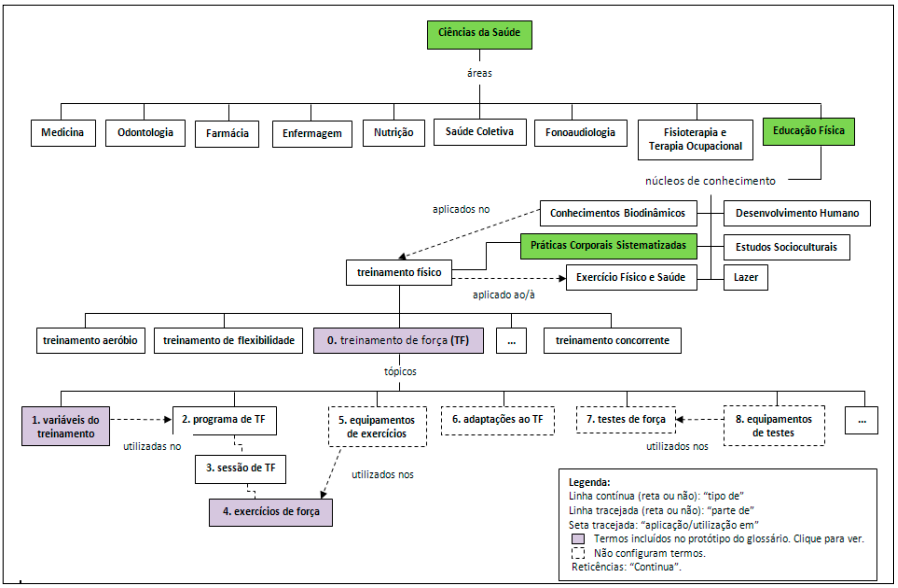


Figura 1 – Árvore de domínio do Treinamento de Força (parte I)
Fonte: Elaborado pelas autoras

[9] Disponível em: <http://www.cnpq.br/documents/10157/186158/TabeladeAreasdoConhecimento.pdf>.
[10] Disponível em: <http://www.ufrgs.br/esef/Arquivos/COMGRAD_EFI/ppc_licenciatura.pdf>.

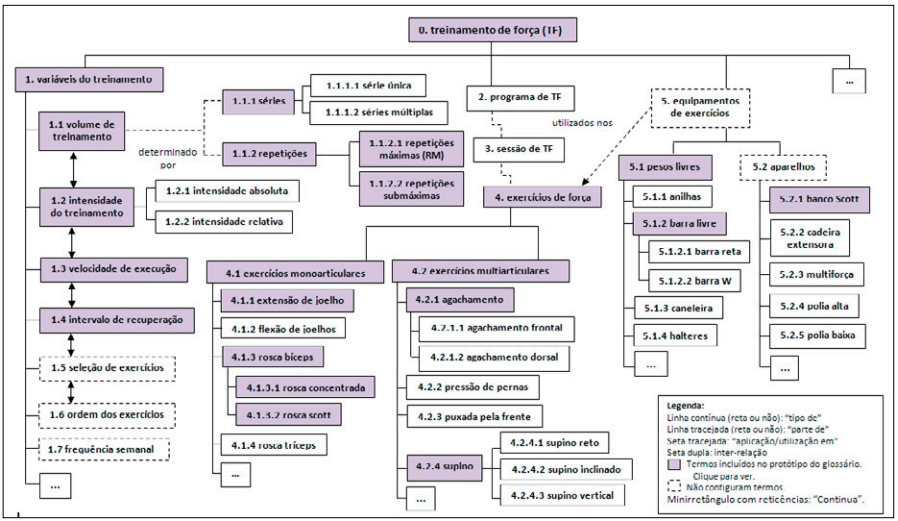


Figura 2 – Árvore de domínio do Treinamento de Força (parte II)
Fonte: Elaborado pelas autoras

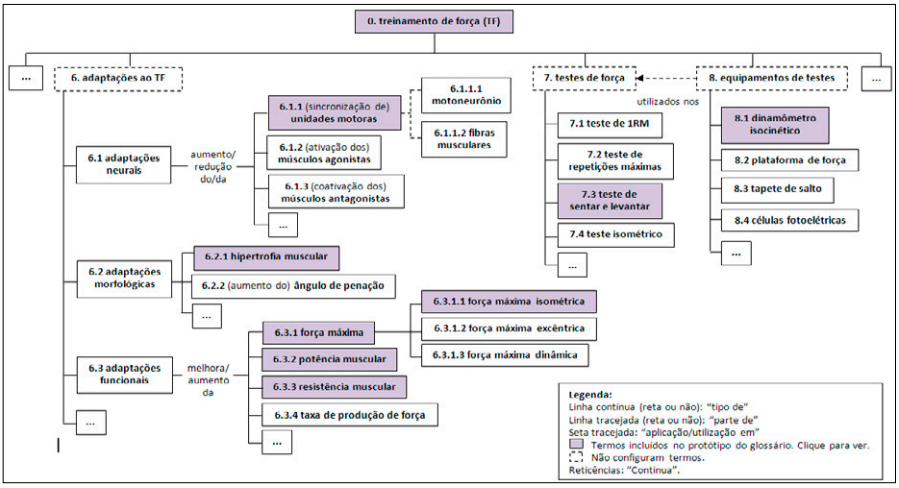


Figura 3 – Árvore de domínio do Treinamento de Força (parte III)
Fonte: Elaborado pelas autoras

Na parte inferior da árvore, abaixo da célula do *treinamento de força* (TF), enumeramos, com o auxílio dos consultores técnicos, os principais tópicos (1-8) abordados nos artigos científicos que compõem o *corpus* de estudo sobre o TF como *tipo de treinamento físico*. O TF é o "termo zero": o ponto de partida para as relações com os demais termos.

Na parte II da árvore (Figura 2), concentramo-nos nos tópicos 1 a 5; e na parte III (Figura 3), nos tópicos 6 a 8. Para estabelecer divisões e subdivisões hierárquicas, valemo-nos do material de apoio já referido, do *corpus* de estudo e do auxílio de nossos consultores. Com essa consultoria especializada, aliada à memória de tradução[11] registrada em nosso glossário particular de Educação Física e à consulta a artigos científicos sobre TF no Google Acadêmico, também foi possível reconhecer e incluir na árvore algumas UTs que não ocorreram no *corpus* de estudo.

Ainda na parte III da árvore, observa-se que *aumento*, *redução* e *melhora* {do/da} são núcleos eventivos que, combinados aos núcleos terminológicos, formam UFEs eventivas que são incluídas nas respectivas fichas terminológicas. Observam-se, também, UTs incluídas pelo critério de pertinência pragmática (MACIEL, 2001). Nos tipos de adaptações neurais, as UTs *unidades motoras* (6.1.1), *músculos agonistas* (6.1.2) e *músculos antagonistas* (6.1.3) são bastante empregadas no âmbito do TF, ainda que sejam mais específicas da área de Biologia, subárea Fisiologia/Neurofisiologia. Também prototípicas dessa subárea são as partes que compõem as unidades motoras: *motoneurônio* (6.1.1.1) e *fibras musculares* (6.1.1.2). Nas adaptações morfológicas, tem-se ângulo de penação (6.2.2), termo mais prototípico da Biologia, subárea Anatomia. Em suma, a decisão de incluir essas UTs na árvore e nas fichas baseia-se nos critérios de frequência e distribuição expressivas no *corpus* de estudo, e à sua pertinência pragmática.

3.4 Reconhecimento das unidades terminológicas

Nosso protótipo de glossário, como já mencionamos, foi construído na direção português-inglês. Assim, ele repertoria UTs mono ou polilexicais que figuram na árvore de domínio em língua portuguesa e encabeçam os exemplares de fichas elaborados. Para o reconhecimento das UTs, primeiramente em português, lançamos mão dos seguintes recursos, nesta ordem:

a) no AntConc, as ferramentas lista de palavras-chave (Keywordlist), n-gramas (N-Grams), concordâncias (Concordances) e Clusters;
b) experiência tradutória (os registros de nossa memória de tradução) na especialidade;
c) consulta ao material de apoio (quando necessário); e
d) consulta aos especialistas consultores (quando necessário).

[11] A memória de tradução refere-se, aqui, ao registro de UTs compiladas ao longo de anos de tradução na área de Educação Física em nosso glossário particular, mencionado na seção 3.2.3.

Para gerar as palavras-chave do *subcorpus* em português (Figura 4), escolhemos como *corpus* de referência o Lácio-Ref (NILC; IME; FFLCH, 2004). O reconhecimento de palavras-chave é um procedimento bastante usual em LC, quando se contrasta um *corpus* de estudo X – geralmente um *corpus* específico – com um amplo *corpus* de caráter geral Y, chamado *corpus* de referência. A chavicidade, assim, é uma medida que apontará os itens lexicais específicos presentes em X frente ao todo da língua que Y representa. Dessa forma, do grande *corpus* do Lácio-Ref, descartamos o grupo de textos dos domínios Ciências Biológicas e Ciências da Saúde, por acreditarmos que parte de sua terminologia seja comum à do TF, e baixamos todo o restante de textos dos outros seis domínios. Já as palavras-chave do *subcorpus* em inglês foram geradas em comparação com a lista de palavras de 100 K disponibilizada pelo COCA – *The Corpus of Contemporary American English* (DAVIES, 2008).

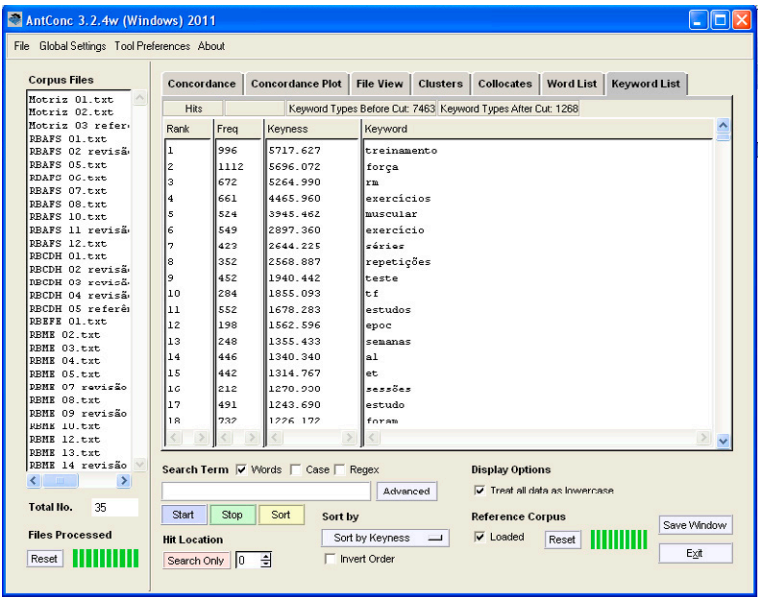


Figura 4 – Extrato da lista de palavras-chave (Keywordlist) do *subcorpus* de estudo em português resultante da comparação com a lista de palavras do Lácio-Ref
Fonte: Elaborado pelas autoras

Geramos, ainda, outras duas listas de palavras-chave, uma para cada *subcorpus* de estudo do TF, desta vez comparando cada um deles com o que chamamos *corpus* de contraste, isto é, um *corpus* na mesma língua de análise composto de textos especializados de um âmbito diferente do investigado. Com esses contrastes, queríamos comprovar se as palavras mais frequentes em nossos *subcorpora* de estudo

eram realmente específicas do âmbito do TF ou eram recorrentes também em outro dos campos da grande área das Ciências da Saúde.

Buscando, então, comprovar a especificidade dos nossos itens de TF frente a outros domínios das Ciências da Saúde, compusemos um *corpus* de contraste em português. Ele foi constituído de 35 artigos científicos da subárea de Dermatologia, selecionados de um *corpus* maior da mesma especialidade construído por estudantes da disciplina Terminologia I do curso de Bacharelado em Letras da UFRGS. Esse material e outros recursos[12] fazem parte da iniciativa denominada "Dermatologia para Tradutores", algo simples mas que foi concebido para auxiliar especialmente professores de tradução que lidam com diferentes línguas de trabalho.

Nossos 35 artigos sobre Dermatologia foram publicados em periódicos brasileiros diversos, e os critérios para seleção foram a atualidade dos textos (2011 a 2013) e a variedade de assuntos abordados. Esse *corpus* de contraste com o material de TF totalizou 103.299 *tokens*.

Nosso *corpus* de contraste em inglês foi constituído também de 35 artigos científicos da subárea de Dermatologia. Dez deles foram publicados no *British Journal of Dermatology*, no período de 2003 a 2013, e estão disponíveis no *site* do mesmo projeto "Dermatologia para Tradutores" da UFRGS. Os outros 25 artigos foram publicados no *Journal of the American Academy of Dermatology*, entre 2013 e 2014. Os critérios para seleção foram os mesmos já descritos. O *corpus* de contraste em inglês totalizou 140.332 *tokens*.

Ao todo, então, foram geradas quatro listas de palavras-chave do âmbito do TF, como mostra o esquema do Quadro 3.

Quadro 3 – Esquema das listas de palavras-chave geradas por *subcorpus* de estudo

***Corpora* de referência/contraste**	***Subcorpus* de estudo TF port.**	***Subcorpus* de estudo TF ingl.**
Corpus de referência port. Lácio-Ref	Lista de palavras-chave port. 1	
Corpus de referência ingl. COCA		Lista de palavras-chave ingl. 1
Corpus de contraste port. 35 artigos Dermatologia	Lista de palavras-chave port. 2	
Corpus de contraste ingl. 35 artigos Dermatologia		Lista de palavras-chave ingl. 2

Fonte: Elaborado pelas autoras

[12] O material está disponível gratuitamente em <http://www.ufrgs.br/textecc/traducao/dermatologia/>.

3.5 Fichamento das UTs

Quatro critérios foram adotados para seleção das UTs-lema, aquelas que encabeçam as fichas terminológicas:

a) distribuição e frequência expressivas no *subcorpus* em português;
b) pertinência temática e pertinência pragmática;
c) possibilidade de encaixe da UT na árvore de domínio; e
d) existência de um equivalente em inglês.

Conforme já frisamos, nosso protótipo de glossário é orientado pela Teoria Comunicativa da Terminologia e baseado em um *corpus* de estudo de TF criteriosamente reunido. Isso implica que um dos parâmetros básicos para a inclusão de uma UT em um produto para tradutores é a sua aceitação e o seu uso pela comunidade de falantes do âmbito especializado sob exame. Geralmente, esse uso é atestado pela sua distribuição e frequência no *corpus* de estudo. O critério de entrada da UT com maior distribuição – frente às suas variantes – é, para nós, mais determinante do que o critério de maior frequência. Isso porque denota que mais especialistas empregam determinada forma do termo, independentemente do número de vezes que ela é repetida.

Em relação às UTs variantes, concordamos com Barros (2004, p. 223), para quem "Se o objetivo é elaborar um dicionário terminológico sem intenções normalizadoras, o registro de toda expressão em relação sinonímica com o termo descrito é importante e recomendável". Assim, optamos por registrar indistintamente todos os tipos de *variantes denominativas* (gráficas, morfossintáticas, lexicais, reduções ou variações complexas diversas), que mantêm entre si uma relação de quase-sinonímia (FREIXA, 2002; 2014), num mesmo campo, genericamente denominado "Variante(s) em português". Na maioria dos casos, elas foram extraídas do *corpus* de estudo e ordenadas por distribuição e frequência; noutras vezes, foram extraídas do material de apoio.

Para extração das candidatas a UTs com maior frequência no *subcorpus* em português, colamos lado a lado numa planilha três listas geradas no AntConc (ver extrato no Quadro 4):

a) os n-gramas (tamanho 2 a 5);
b) as palavras-chave em comparação com o Lácio-Ref; e
c) as palavras-chave em contraste com o *corpus* de Dermatologia em português.

Primeiramente, analisamos os 1.000 (mil) primeiros n-gramas com vistas a reconhecer e extrair UTs polilexicais. Em seguida analisamos as 300 primeiras palavras-chave de cada lista, a fim de reconhecer e extrair UTs monolexicais. Na

análise de todas as listas, valemo-nos, num primeiro momento, somente de nossa experiência tradutória na especialidade – portanto ainda na esfera das "suspeitas" – e adotamos esta classificação por cores:

- Forte candidata a UT: pertinência temática;
- Possível candidata a UT: dúvida sobre pertinência temática ou pragmática;
- Unidade verificada e descartada: sem pertinência temática ou pragmática;
- Candidato a núcleo eventivo de UFE.

Sem destaque de cor: Improvável de ser UT: sem pertinência temática ou pragmática.

Quadro 4 – Extrato de planilhas comparativas dos n-gramas e palavras-chave do *subcorpus* em português

KEYWORDS PORT COM LACIO-REF 6DOMS
Keyword Types Before Cut: 7463
Keyword Types After Cut: 1268 (lowercase)

Rank	Freq	Keyness	Keyword
1	996	5.717.627	treinamento
2	1112	5.696.072	força
3	672	5.264.990	rm
4	661	4.465.960	exercícios
5	524	3.945.462	muscular
6	549	2.897.360	exercício
7	423	2.644.225	séries
8	352	2.568.887	repetições
9	452	1.940.442	teste
10	284	1.855.093	tf
11	552	1.678.283	estudos
12	198	1.562.596	epoc
13	248	1.355.433	semanas
14	446	1.340.340	al
15	442	1.314.767	et
16	212	1.278.930	sessões
17	491	1.243.690	estudo
18	732	1.226.172	foram
19	265	1.224.560	carga
20	208	1.194.319	idosos
21	214	1.107.951	sessão

KEYWORDS PORT COM DERMATO35
Keyword Types Before Cut: 7463
Keyword Types After Cut: 1065 (lowercase)

Rank	Freq	Keyness	Keyword
1	1112	1.345.571	força
2	996	1.192.443	treinamento
3	672	808.434	rm
4	661	784.358	exercícios
5	549	671.456	exercício
6	524	608.578	muscular
7	423	463.584	séries
8	352	430.515	repetições
9	8117	429.427	de
10	284	347.347	tf
11	552	325.318	estudos
12	265	288.758	carga
13	446	276.336	al
14	442	275.186	et
15	452	268.667	teste
16	218	255.416	máxima
17	214	250.561	sessão
18	198	242.165	epoc
19	1515	235.677	para
20	208	234.926	idosos
21	212	232.389	sessões

N-GRAMAS 2-5 PORT
Total No. of N-Grams Types: 343460
Total No. of N-Grams Tokens: 472164 (lowercase)

Rank	Freq	N-gram
1	666	de força
2	442	et al
3	428	de rm
4	417	de treinamento
5	304	treinamento de
6	281	para a
7	258	treinamento de força
8	247	que o
9	240	teste de
10	214	força muscular
11	208	que a
12	204	para o
13	197	número de
14	185	com o
15	183	entre os
16	182	do treinamento
17	180	de repetições
18	172	e a
19	164	da força
20	158	o treinamento
21	146	de recuperação

Fonte: Elaborado pelas autoras

Posteriormente, verificamos os dados seguindo os demais critérios de reconhecimento de UTs (ver subseção 3.4): extração de concordâncias e *clusters*, e consulta ao material de apoio e aos especialistas.

Com relação ao critério de encaixe na árvore de domínio, Barros (2004, p. 127) adverte que

> Certas unidades terminológicas – e talvez muitas delas – podem não se encaixar no sistema preestabelecido. É preciso lembrar que um sistema nunca é definitivo e único: é o resultado de uma concepção, de uma estruturação dos elementos de acordo com certas relações de sentido que foram privilegiadas pelo terminólogo

responsável pelo projeto, portanto deve ser flexível para comportar novas relações e novos termos.

Assim, na etapa de arquitetura de nossa árvore – que foi refeita várias vezes –, algumas unidades que aparecem dentro dos intervalos selecionados e que foram confirmadas como UTs acabaram não sendo incluídas, pelo menos por ora, dadas nossas limitações de espaço e tempo. Foi o caso, por exemplo, das UTs *ativação muscular, déficit bilateral*, *massa muscular* e *desempenho muscular*.

O contrário também ocorreu: algumas UTs com baixa frequência, fora do intervalo estipulado, foram abrigadas em fichas terminológicas pelo fato de terem pertinência temática e de terem lugar no sistema de conceitos do TF representado pela árvore de domínio. É o caso, por exemplo, de *barra livre* (posição 5.1.2 na parte II da árvore), com apenas uma ocorrência (um *hápax legómenon*) no *corpus* de estudo. Além de esse termo corresponder, concretamente, a um tipo de peso livre indispensável em academias de musculação, essa UT guarda relação de hiperonímia com as UTs *barra reta* (5.1.2.1) e *barra W* (5.1.2.2).

Adicionalmente, também reconhecemos e incluímos na árvore algumas UTs que não ocorrem no *corpus* de estudo, mas que foram confirmadas como tal em consulta ao material de apoio e aos especialistas consultados. São elas a *força excêntrica máxima* (6.3.1.2) e os equipamentos de testes *tapete de salto* (8.3) e *células fotoelétricas* (8.4).

Por fim, a população de UTs incluídas na árvore a partir do termo *treinamento de força* (termo zero) somou 71, e a amostra fichada foi de 30 unidades (42,25%). Para compor essa amostra, procuramos selecionar alguns poucos termos situados em cada "galho" da árvore, de forma que cada tópico, categoria e subcategoria do sistema de conceitos do TF estivessem representados de forma equilibrada.

O último critério de seleção das UTs para inclusão na árvore de domínio e fichamento foi a existência de pelo menos um equivalente em inglês, considerando, claro, que nosso protótipo de glossário é bilíngue.

3.6 Reconhecimento de equivalentes em inglês

Os equivalentes em inglês são, num primeiro momento, verificados no nosso glossário particular preexistente. Em seguida, geramos concordâncias com os mesmos no AntConc. Numa análise qualitativa, verificamos sua cobertura semântica do conceito veiculado pela UT em português. Para tanto, recorremos aos cotextos das concordâncias, ao material de apoio (especialmente os livros-texto) e aos nossos consultores especialistas. Dessa forma, verificamos se o equivalente é, além de referencial, também funcional, ou seja, se ele expressa a mesma relação semântica e o mesmo efeito pragmático nos cotextos extraídos (cf. GÉMAR,

1998) e, assim, "funciona" em termos comunicativos de forma equivalente à UT em português nos artigos científicos sobre TF. Sempre que julgamos relevante, fazemos observações nesse sentido nas *Notas* de tradução na ficha terminológica.

Em seguida, passamos a uma análise quantitativa para verificar o equivalente com maior distribuição e, em caso de empate, com maior frequência. Esses números são informados ao lado dos equivalentes na ficha. Ponderadas essas duas análises, indicamos ao usuário o "equivalente preferencial".

Quando não encontramos um equivalente no *subcorpus* em inglês, recorremos ao material de apoio.

3.7 Reconhecimento de UFEs eventivas

Para o reconhecimento dessas unidades semifixas em português, geramos concordâncias com as UTs selecionadas (UTs-lema) no AntConc e consideramos fraseológicas aquelas com frequência/distribuição no *subcorpus* de estudo a partir de 2/2, isto é, no mínimo duas ocorrências distribuídas em, pelo menos, dois artigos científicos (ver Figura 5).

Essa decisão foi baseada na ponderação de Bevilacqua (1998, p. 125-126) sobre o critério de frequência para detecção das UFEs: "[...] consideramos que a frequência é um critério aleatório que depende de outros fatores como o tamanho e o tipo de *corpus* utilizado como fonte de coletas dessas unidades e deve, portanto, ser definido segundo as especificidades de cada trabalho". Considerando, portanto, a especificidade, o alto grau de especialidade e o tamanho do nosso *subcorpus*, o critério de frequência/ distribuição mínimas de 2/2 nos parece razoável para os fins a que se destina o glossário. Aumentando a distribuição para três artigos, quase não haveria UFEs eventivas a registrar, e o tradutor ficaria prejudicado.

É importante destacar que os equivalentes em inglês fornecidos nas fichas para as fraseologias em português nem sempre são fraseológicos, considerando que eles podem não ter um elevado grau de fixação. Conforme Reuillard e Kilian (2014, p. 475), "uma CLE [combinatória léxica especializada] pode apresentar estruturas distintas em cada língua ou até mesmo não se constituir como combinatória recorrente em determinada língua". Assim, para esses equivalentes funcionais, não nos limitamos a uma distribuição e frequência mínimas no *subcorpus* de estudo em inglês.

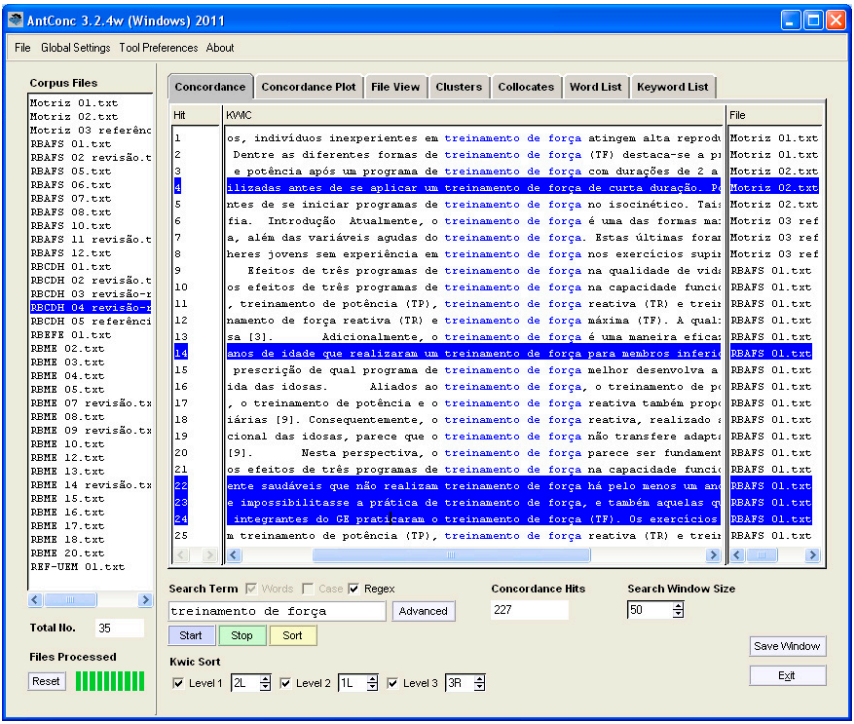


Figura 5 – Extrato de lista de concordâncias para extração de candidatas a UFEs eventivas
Fonte: Elaborado pelas autoras

3.8 O guia do usuário do glossário

Um guia de uso, para o tradutor, é indispensável. Afinal, ele precisa encontrar as informações rapidamente e saber compreender o modo de concepção da obra que lhe é oferecida. No nosso Guia do Usuário, fornecemos as seguintes informações:

- a quem se destina o glossário;
- o propósito do glossário;
- as características do glossário;
- a constituição básica do *corpus* de estudo;
- as partes do glossário: árvore de domínio, lista de termos e fichas terminológicas;
- como usar o glossário;
- listas de abreviaturas e símbolos utilizados nas fichas terminológicas;

- lista de itens microestruturais das fichas terminológicas;
- exemplar reduzido de ficha terminológica, com explicações sobre cada seção.

Ao definirmos as três partes que compõem o glossário, sugerimos ao usuário um percurso de consulta, assim representando:

Lista ⇒ Ficha ⇒ Árvore ⇒ Ficha

Ao final, está a parte que julgamos mais trabalhosa porém mais atrativa do Guia: um exemplar reduzido de ficha terminológica, com a indicação das partes que a compõem. A ideia é mostrar ao usuário tradutor, de forma simples e pontual, o papel de cada informação como facilitadora do seu processo tradutório (Figura 6).

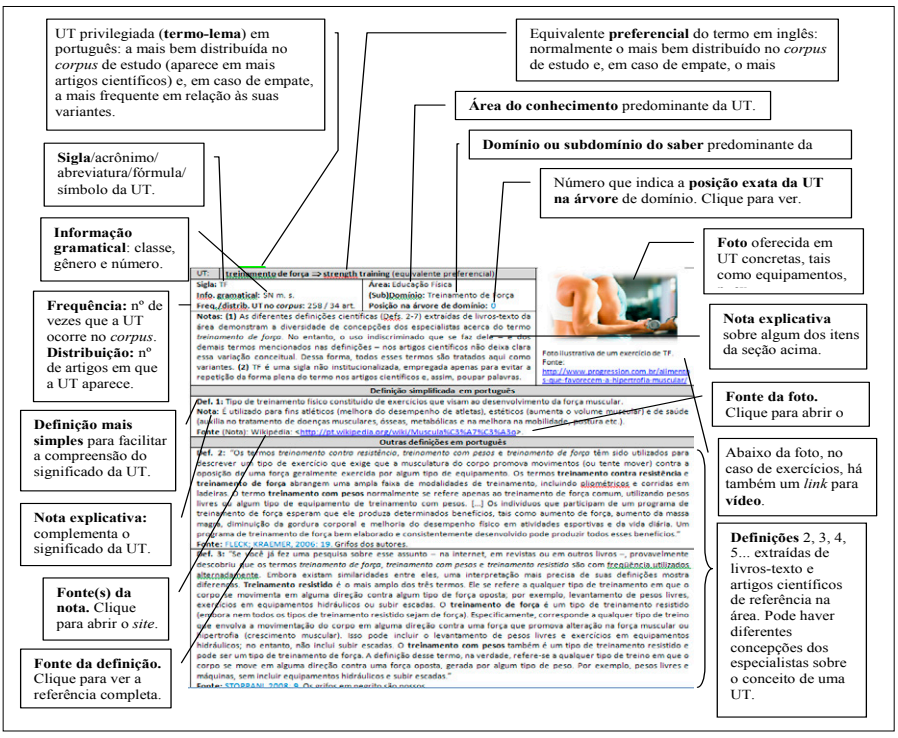


Figura 6 – Exemplar reduzido de ficha terminológica
Fonte: Elaborado pelas autoras

3.9 A lista de termos em português

A lista contempla, em ordem alfabética contínua, todas as UTs-lema e as UTs variantes apresentadas nas fichas. Pensando numa futura edição eletrônica, as UTs-lema, destacadas em azul, vêm com *hiperlink* para a respectiva ficha. Já as UTs variantes apresentam remissão, no formato *Ver*, para a UT privilegiada, que é destacada em azul e vem com *hiperlink* para a respectiva ficha. Vejamos um extrato da lista:

intervalo de recuperação
intervalo de repouso — *Ver* intervalo de recuperação
intervalo(s) de descanso — *Ver* intervalo de recuperação
musculação — *Ver* treinamento de força
período de descanso — *Ver* intervalo de recuperação
período de repouso — *Ver* intervalo de recuperação
período(s) de recuperação — *Ver* intervalo de recuperação
pesos livres
potência muscular

A lista completa registra 30 UTs-lema e 89 UTs variantes. Unidades homônimas não foram encontradas; se encontradas futuramente, com a expansão do glossário, serão identificadas com um número sobrescrito (por exemplo, UT[1], UT[2]...), já que cada uma terá uma ficha própria.

3.10 A ficha terminológica

Para a elaboração do nosso modelo de ficha, tomamos por base os estudos de Fromm (2007a) e Teixeira (2008), que investigaram quais itens o tradutor precisa que conste em um dicionário técnico. Além das propostas desses autores, valemo-nos também das de Almeida (2000), Silva e Teixeira (2010), e de dados disponíveis no já citado recurso "Dermatologia para Tradutores" do Projeto TEXTECC da UFRGS.

Nossa ficha inclui os seguintes itens microestruturais:

- UT em português;
- sigla/acrônimo/abreviatura/fórmula/símbolo (cf. o caso);
- informação gramatical da UT;
- frequência e distribuição da UT no *corpus*;
- área e (sub)domínio da UT;
- posição da UT na árvore de domínio (com *hiperlink* para a árvore);
- figura (conforme o caso);
- *hiperlink* para vídeo (por exemplo, para demonstrar um exercício);

- definição simplificada da UT em português;
- outras definições da UT em português;
- variante(s) da UT em português;
- equivalente(s) da UT em inglês;
- UFEs eventivas em português;
- equivalentes funcionais em inglês;
- exemplos de ocorrências da UT no *corpus* em português;
- exemplos de ocorrências da UT em inglês;
- UTs relacionadas em português, com remissivas (*hiperlinks*) para as fichas;
- notas explicativas e de tradução.

Para detalhes sobre cada item, com explicação dos critérios e procedimentos seguidos, bem como das dificuldades e soluções encontradas, ver Dornelles (2015).

Na Figura 7, oferecemos um exemplar de ficha, da UT *dinamômetro isocinético.*

<table>
<tr><td>UT:</td><td colspan="2">dinamômetro isocinético ⇒ isokinetic dynamometer (equivalente preferencial)</td><td rowspan="6"> 
Foto ilustrativa de um dinamômetro isocinético. Fonte: http://www.institutocohen.com.br/reabilitacaoarea_interna.php?id=4
ASSISTA AO VÍDEO: https://www.youtube.com/watch?v=ADkfNM0YRRE</td></tr>
<tr><td colspan="2">Sigla:</td><td>Área: Educação Física</td></tr>
<tr><td colspan="2">Info. gramatical: SN m. s.</td><td>(Sub)Domínio: Treinamento de Força</td></tr>
<tr><td colspan="2">Freq./distrib. UT no corpus: 12 / 06 art.</td><td>Posição na árvore de domínio: 8.1</td></tr>
<tr><td colspan="3">Nota:</td></tr>
<tr><td colspan="3">Definição simplificada em português</td></tr>
<tr><td colspan="3">Def. 1: Equipamento eletromecânico que serve para testar a produção de força em velocidade constante, esta controlada por computador.</td></tr>
<tr><td colspan="3">Nota:</td></tr>
<tr><td colspan="3">Fonte:</td></tr>
<tr><td colspan="4">Outras definições em português</td></tr>
<tr><td colspan="4">Def. 2: “Dinamômetros isocinéticos proporcionam uma avaliação acurada e confiável da força, da resistência e da potência de grupos musculares [...]. A velocidade de movimento do membro é mantida em velocidade pré-selecionada constante. Qualquer aumento na força muscular produz um aumento na resistência em vez de acelerar o segmento. Desse modo oscilações na força muscular ao longo da AM [amplitude de movimento] são combinadas por uma força contrária igual ou uma resistência adaptável. Os dinamômetros isocinéticos medem a produção de torque muscular em velocidades de 0 a 300°/s. A partir da produção registrada, o pico de torque, o trabalho total e a potência podem ser avaliados.”
Fonte: HEYWARD, 2013: 157.</td></tr>
</table>

Def. 3: "Dinamômetro isocinético. Esse equipamento computadorizado pode ser programado para movimentar-se em várias velocidades. Eles geralmente são encontrados apenas em laboratórios ou em clínicas de medicina do esporte como ferramenta para a medição da quantidade de força que um atleta pode produzir. Esse tipo de equipamento com freqüência é conectado a um computador não apenas para controlar a velocidade de movimento, mas também para medir a força aplicada. Existem diversas desvantagens nos dinamômetros isocinéticos. A primeira é o fato de serem possíveis apenas movimentos angulares. Em outras palavras, eles permitem somente movimentos de flexão e extensão de cotovelo, punho, joelho ou tornozelo. Tais equipamentos não podem ser usados em exercícios de "empurrar", como o supino, o meio desenvolvimento ou o agachamento. A outra desvantagem é que, na verdade, não existem ações musculares isocinéticas nos movimentos da vida real." **Fonte:** STOPPANI, 2008: 41.	
Variante(s) em português	
Var. 1: dinamômetro	**Freq./distrib. Var1 no *corpus*:** 07 / 05 art.
Nota: Essa é uma variante por redução. Ela tende a ocorrer nos artigos após o emprego do termo pleno *dinamômetro isocinético*, para evitar repetição e por economia linguística.	
Equivalente(s) em inglês	
Eq. 1 (preferencial):isokinetic dynamometer	**Freq./distrib. Eq1 no *corpus*:** 20 / 11 art.
Eq. 2:dynamometer	**Freq./distrib. Eq2 no *corpus*:** 30 / 11 art.
Nota: O Eq. 2 é um equivalente por redução. Ele tende a ocorrer nos artigos após o repetido emprego do termo pleno *isokinetic dynamometer*, para evitar repetição e por economia linguística.	
Fraseologia(s) em português	Equivalentes funcionais em inglês
FP1: utilizar {o/um} dinamômetro isocinético	**EFI1:** to use {an/the} isokinetic dynamometer
Nota:	Nota:
Exemplo(s) de ocorrência(s) no *corpus* em português	**Exemplo(s) de ocorrência(s) em inglês**
ExP1: "A força extensora do joelho foi mensurada, **utilizando-se o dinamômetro isocinético** Biodex System 3 (Biodex, New York, USA) antes e após o período de treinamento." (RBCDH 01)	**ExI1:** "Maximal isometric and dynamic-knee extension torques were measured in seated position **using an isokinetic dynamometer** that consisted of a computer controlled electromotor (SEW-Eurodrive, Bruchsal, Germany) instrumented with a torque transducer (Lebow®1605, accuracy level 0.05%, Troy, USA)." (EJAP 22)
Nota:	Nota:
UT relacionadas	
treinamento de força; força máxima; potência muscular; resistência muscular; taxa de produção de força	

Figura 7 – Ficha da UT *dinamômetro isocinético*
Fonte: Elaborado pelas autoras

4 Características da terminologia do Treinamento de Força nos artigos científicos

Seguindo os objetivos da pesquisa, nesta seção oferecemos uma descrição do comportamento das UTs em português e inglês, e das UFEs eventivas em português nos artigos sobre TF que compuseram nosso *corpus* de estudo.

4.1 Morfossintaxe das unidades terminológicas

As UTs em português, incluindo as 30 UTs-lema e suas 59 variantes, são, em sua grande maioria, polilexicais. Do total de 89 unidades repertoriadas no protótipo do glossário, 76 (85%) são sintagmas nominais (SN), 8 (9%) são monolexicais (substantivos) e 5 (6%) são siglas ou abreviatura. As estruturas básicas mais recorrentes dos SN são quatro:

- N + prep (+ art) + N, com 30 UTs (34%). Exs.: *treinamento de força, treinamento com pesos, variáveis do treinamento, intervalo de recuperação*.
- N + ADJ, com 27 UTs (30%). Exs.: *treinamento resistido, repetições máximas, rosca direta, pesos livres, força máxima.*
- N + N, com 5 UTs (6%). Exs.: *rosca bíceps, rosca scott, banco Scott*.
- N + ADJ + ADJ, com 5 UTs (6%). Exs.: *força máxima isométrica, força estática máxima, resistência muscular localizada*.

Os equivalentes em inglês, incluindo os 30 equivalentes preferenciais e suas 48 variantes, também são, em sua grande maioria, polilexicais. Do total de 78 unidades, 67 (86%) são *compound nouns*, 6 (8%) são monolexicais (substantivos) e 5 (6%) são *abbreviations*. Há uma variedade maior de estruturas em comparação às UTs em português. As polilexicais mais recorrentes são cinco:

- N + N, com 27 UTs (35%). Exs.: *strength training, resistance training, weight training, training variables, program variables, rest period(s)*.
- ADJ + N, com 14 UTs (18%). Exs.: *submaximal repetitions, free weights.*
- ADJ + N + N, com 6 UTs (8%). Exs.: *acute training variable(s), single-joint exercises, multiple-joint exercises, local muscle endurance*.
- N + V, com 5 UTs (6%): Exs.: *bíceps curl, arm curl, bench press* (todos exercícios).
- ADJ + ADJ + N, com 4 UTs (5%). Exs.: *maximal isometric strength, maximum isometric force*.

Percebe-se que os termos polilexicais, tanto em português como em inglês, apresentam como estruturas prototípicas (as duas primeiras em cada língua) as mesmas encontradas em sintagmas da língua geral.

4.2 UFEs eventivas

As UFEs em português não foram numerosas como esperávamos, mesmo adotando uma frequência/distribuição não muito alta, como é a 2/2. Por esse critério, foram encontradas 33 unidades no *corpus* de estudo. Os núcleos eventivos

com nominalizações são três vezes mais frequentes que com verbos e quatro vezes mais frequentes que com particípios, confirmando os achados de outras pesquisas terminológicas. Vejam-se:

- Nominalizações: 21 UFEs (64%). Exs.: *prática de treinamento de força, combinação* {*de/das*} *variáveis do treinamento, aumento* {*da/na*} *intensidade do treinamento, execução* {*de/dos*} *exercícios de força*.
- Verbos: 7 UFEs (21%). Exs.: *realizar* (*um*) *treinamento de força, determinar aintensidade do treinamento, executar* [NUM] *séries*.
- Particípio: 5 UFEs (15%). Exs.: *treinamento de força realizado, número de repetições completadas, unidade(s) motora(s) recrutada(s)*.

4.3 Variação terminológica

A variação foi um fenômeno bastante expressivo, nas duas línguas. Em português, cada UT-lema apresentou de 0 a 7 variantes. Quanto aos tipos de variação (cf. FREIXA, 2002), encontramos:

- Lexical: 31 variantes (53%), como nos pares *treinamento de força/treino de força; intensidade de treinamento/carga de treinamento; exercícios de força/exercícios resistidos; extensão de joelho/extensão de perna(s)*.
- Por redução: 12 variantes (20%), como em *rosca scott/rosca bíceps scott; barra livre/barra; hipertrofia muscular/hipertrofia*.
- Morfossintática: 7 variantes (12%), como em *extensão de joelho/extensão de joelhos/extensão do joelho/extensão dos joelhos*; *força máxima isométrica/força isométrica máxima.*
- Gráfica: 7 variantes (12%), como em *volume do treinamento/volume de treinamento; extensão de joelho/extensão de joelhos; força máxima isométrica/força isométrica máxima*.
- Complexa (lexical e redução concomitantemente): 2 variantes (3%), em *treinamento de força/musculação* e *treinamento de força/treino resistido*.

Em inglês, cada equivalente preferencial apresentou de 0 a 5 variantes. Os tipos de variação encontrados foram estes:

- Lexical: 21 variantes (43%), como nos pares *strength training/resistance training; rest period(s)/rest interval(s), movement velocity/repetition velocity*.
- Gráfica: 11 variantes (23%), como em *repetition(s)/rep(s); repetition maximum/RM; multi-joint exercise(s)/multijoint exercise(s)*; etc.
- Por redução: 8 variantes (17%), como em *training variables/training program variables; concentration curl/bíceps concentration curl.*

- Morfossintática: 8 variantes (17%), como em *rest period(s)/resting period(s); repetition maximum/repetitions maximum; muscle power/ muscular power*.

Os gráficos da Figura 8 mostram a distribuição dos tipos de variação nas amostras das UTs em inglês e português.

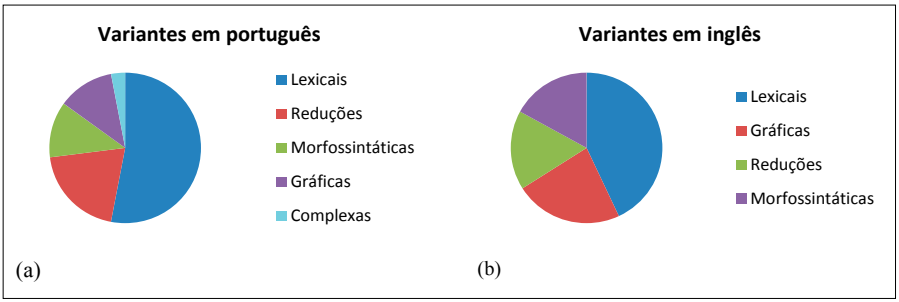


Figura 8 – Distribuição dos tipos de variação nas UTs em (a) português e (b) inglês
Fonte: Elaborado pelas autoras

5 Considerações finais

Nossa pesquisa de mestrado, aqui resumida, contemplou uma parte teórica e uma parte aplicada que se inter-relacionam e se inserem na dupla face da Terminologia, que é teórica e, ao mesmo tempo, prática. Afinal, há uma descrição de uma linguagem especializada a partir de um dado ponto de vista teórico e o desenho de um produto concreto.

Nosso protótipo de glossário inclui uma árvore de domínio em português, com uma população de 71 unidades terminológicas (UT) e uma amostra fichada de 30 UTs desse universo (42,25%). Há nele também um Guia do usuário; uma Lista de termos em português, com 30 UTs-lema e 89 UTs variantes; e 30 fichas terminológicas. São 78 termos em inglês, sendo 30 equivalentes preferenciais e 48 variantes.

Como limitação do estudo, apontamos o fato de que o protótipo de glossário não pode ser submetido à avaliação de sua "usabilidade" por parte de tradutores e de estudantes de tradução. Isso ainda deverá ser feito, para que possamos aproveitar as contribuições desses potenciais usuários para aprimorar o modelo de glossário e seguir adiante com o trabalho na sua versão editorial completa, em formato eletrônico, contando-se com recursos mais sofisticados do que pudemos ter. Por isso, vale divulgar o nosso trabalho também neste livro e para a comunidade de pesquisa de Linguística de *Corpus* do Brasil.

Agradecimentos

A primeira autora agradece à UFRGS, especialmente à Escola de Educação Física, Fisioterapia e Dança, pela oportunidade de afastamento para qualificação com este estudo. A segunda autora agradece ao PPG-Letras da UFRGS, à CAPES, ao CNPq e à FAPERGS. Ambas agradecemos a inestimável colaboração de nossos consultores especialistas, Prof. Ronei Pinto e Prof. Eduardo Cadore, da ESEFID/UFRGS.

Referências

[ABL] ACADEMIA BRASILEIRA DE LETRAS. *Vocabulário ortográfico da língua portuguesa.* 2009. Disponível em: <http://www.academia.org.br/abl/cgi/cgilua.exe/sys/start.htm?sid=23>. Acesso em: 18 out. 2017.

ALMEIDA, G. M. de B. *Teoria comunicativa da terminologia (TCT)*: uma aplicação. 2000. 2v. 290 f. Tese (doutorado em Linguística e Língua Portuguesa). Faculdade de Ciências e Letras de Araraquara, Universidade Estadual Paulista "Júlio de Mesquita Filho", Araraquara, 2000.

ANTHONY, L. *AntConc* (Version 3.2.2) [Computer software]. Tokyo: Waseda University, 2011. Disponível em: <http://www.laurenceanthony.net/software/antconc/>.

[AOLP] ANGOLA; BRASIL; CABO VERDE; GUINÉ-BISSAU; MOÇAMBIQUE; PORTUGAL. *Acordo ortográfico da língua portuguesa*. [*on-line*]. Dez. 1990. Disponível em: <http://www.portaldalinguaportuguesa.org/acordo.php?action=acordo&version=1990>. Acesso em: 18 out. 2017.

BARROS, L. A. *Curso básico de Terminologia.* São Paulo: EDUSP, 2004. 287 p.

BERBER SARDINHA, T. *Lingüística de* Corpus. Barueri: Manole, 2004. 410 p.

BEVILACQUA, C. R. Unidades fraseológicas especializadas: elementos para seu reconhecimento em *corpora* textuais. *Intercâmbio*, v. XII, p. 215-223, 2003.

_____. *Unidades fraseológicas especializadas eventivas*: descripción y reglas de formación en el ámbito de la energía solar. 2004. 242 f. Tese (doutorado em Lingüística Aplicada). Universidade Pompeu Fabra, Instituto Universitário de Lingüística Aplicada (IULA), Barcelona, 2004. Disponível em: <http://www.ufrgs.br/termisul/files/file622266.pdf>. Acesso em: 18 out. 2017.

_____. Unidades fraseológicas especializadas: novas perspectivas para sua identificação e tratamento. *Organon* – Revista do Instituto de Letras da UFRGS, Porto Alegre, v. 12, n. 26, p. 119-132, 1998. Disponível em: <http://www.seer.ufrgs.br/index.php/organon/article/view/29562/18262>. Acesso em: 18 out. 2017.

BIBER, D. Representatividade em planejamento de *corpus*. Tradução de Paula Marcolin. *Cadernos de Tradução*, Porto Alegre, n. 30, p. 11-45, jan./jun. 2012.

CABRÉ, M. T. *La terminología:* representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos. Barcelona: Institut Universitari de Linguística Aplicada (IULA)/ Universitat Pompeu Fabra, 1999a. (Sèrie Monografies, 3).

_____. Variació per tema. El discurs especialitzat o la variació funcional determinada per la temática: noves perspectives. *Caplletra:* Revista Internacional de Filología, Publicacions de l'Abadia de Montserrat, Institut de Filologia Valenciana, Valencia, n. 25, p. 173-194, 1999b.

_____. Sumario de principios que configuran la nueva propuesta teórica y consecuencias metodológicas. In: CABRÉ, M. T.; FELIU, J. (Ed.). *La terminología científico-técnica:* reconocimiento, análisis y extracción de información formal y semántica. (Informe DGES PB-96-0293). Barcelona: IULA/Universitat Pompeu Fabra, 2001a, p. 17-25.

_____. Consecuencias teóricas de la propuesta metodológica. In: CABRÉ, M. T.; FELIU, J. (Ed.). *La terminología científico-técnica*: reconocimiento, análisis y extracción de información formal y semántica. (Informe DGES PB-96-0293). Barcelona: IULA/Universitat Pompeu Fabra, 2001b, p. 27-36.

_____. Theories of terminology: their description, prescription and explanation. *Terminology*, n. 9, v. 2, p. 163-200, 2003.

_____. La teoría comunicativa de la terminología: una aproximación lingüística a los términos. *Revue Française de Linguistique Appliquée*, v. XIV-2, p. 9-15, 2009. Disponível em: <http://www.cairn.info/revue-francaise-de-linguistique-appliquee-2009-2-page-9.htm>. Acesso em: 18 out. 2017.

DAVIES, M. *COCA* – The *Corpus* of Contemporary American English [*corpus*]. Provo: Brigham Young University, 2008. Disponível em: <http://corpus.byu.edu/coca/>. Acesso em: 18 out. 2017.

DORNELLES, M. dos S. A variação no emprego da terminologia anatômica no âmbito da educação física: um estudo exploratório. *Debate Terminológico*, n. 12, p. 3-20, 2014. Disponível em: <http://seer.ufrgs.br/index.php/riterm/article/view/52587/32498>. Acesso em: 18 out. 2017.

_____. *Bases teórico-metodológicas para elaboração de um glossário bilíngue (português-inglês) de treinamento de força:* subsídios para o tradutor. 2015. 364 f. Dissertação (mestrado em Letras). Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2015. Disponível em: <http://hdl.handle.net/10183/117567>. Acesso em: 18 out. 2017.

FAULSTICH, E. Aspectos de terminologia geral e terminologia variacionista. *TradTerm*, v. 7, p. 11-40, 2001. Disponível em: <http://www.revistas.usp.br/tradterm/article/view/49140>. Acesso em: 18 out. 2017.

FINATTO, M. J. B. A definição de termos técnico-científicos no âmbito dos estudos de terminologia. *Revista de Estudos da Linguagem*, Belo Horizonte, v. 11, n. 1, p. 197-222, jan./jun. 2003. Disponível em: <http://periodicos.letras.ufmg.br/index.php/relin/article/view/2351>. Acesso em 04 fev. 2015.

_____. Orientações para a terminografia: das teorias às práticas em busca de amplitude da informação terminológica. In: ISQUERDO, A. N.; DAL CORNO, G. O. M. (Org.). *As ciências do léxico*: lexicologia, lexicografia, terminologia. Vol. VII. Campo Grande: Ed. UFMS, 2014, p. 439-457.

FREIXA, Judit. *La variació terminològica:* anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient. 2002. 397 f. Tese (doutorado) – Universitat de Barcelona, Barcelona. Disponível em: <http://www.tdx.cat/handle/10803/1677>. Acesso em 18 out. 2017.

_____. La variación denominativa en terminología: tipos y causas. *In:* ISQUERDO, A. N.; DAL CORNO, G. O. M. (Orgs.). *As ciências do léxico:* lexicologia, lexicografia, terminologia, vol. VII. Campo Grande, MS: Ed. UFMS, 2014. p. 311-329.

FROMM, Guilherme. *VoTec:* a construção de vocabulários eletrônicos para aprendizes de tradução. 2007. 215 f. Tese (doutorado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. Disponível em: <http://www.teses.usp.br/teses/disponiveis/8/8147/tde-08072008-150855/pt-br.php>. Acesso em 18 out. 2017.

GÉMAR, Jean-Claude. Les enjeux de la traduction juridique. Principes et nuances. *ASTTI Seminar :* Équivalences 1998 : Traduction de textes juridiques : problèmes et méthodes, 1998. Disponível em: <http://www.tradulex.com/Bern1998/Gemar.pdf>. Acesso em 18 out. 2017.

GOOGLE INC. *Google Acadêmico* [*site*]. 2011. Mountain View, CA, USA. Disponível em: <http://scholar.google.com.br/>. Acesso em 18 out. 2017.

HURTADO ALBIR, Amparo. *Traducción y traductología:* introducción a la traductología. 4. ed. Madrid: Cátedra, 2008. 695 p.

[ISO 1087] _____. *ISO 1087 :* Terminologie – Vocabulaire. Genebra, ISO, 1990. Disponível em: <http://www.iso.org/iso/home.html>.

KRIEGER, Maria da Graça; FINATTO, Maria José Bocorny. *Introdução à terminologia:*teoria & prática. São Paulo: Contexto, 2004. 223 p.

MACIEL, Anna Maria Becker. Pertinência pragmática e nomenclatura de um dicionário terminológico. *In:* KRIEGER, M. G.; MACIEL, A. M. B. (Orgs.) *Temas de terminologia.* Porto Alegre/São Paulo: Ed. Universidade/UFRGS/Humanitas/USP, 2001. p. 275-284.

_____. Terminologia e *Corpus*. *In:* TAGNIN, S.; BEVILACQUA, C. (Orgs.) Corpora *na Terminologia.* São Paulo: HUB Editorial, 2013. p. 29-45.

MICROSOFT CORPORATION. *Microsoft Office Word 2007* [Programa computacional]. 2006.

NILC – Núcleo Interinstitucional de Linguística Computacional; IME – Instituto de Matemática e Estatística, Universidade de São Paulo; FFLCH – Faculdade de Filosofia, Letras e Ciências Humanas, USP. *Lácio-Ref* [*corpus*]. 2004. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/index.htm>.

PACTE group. Building a translation competence model. Results of the validation of the PACTE translation competence model: translation problems and translation competence. In: *Methods and strategies of process research:* integrative approaches in Translation Studies. Amsterdam: John Benjamins, 2011. Disponível em: <http://grupsderecerca.uab.cat/pacte/sites/grupsderecerca.uab.cat.pacte/files/PACTE%202011_%20Validation%20TC%20Model.pdf>. Acesso em 18 out. 2017.

PEARSON, J. Como ter acesso a elementos definitórios nos textos especializados. Tradução de Carolina Huang e Sandra Dias Loguercio. *Cadernos de Tradução.* Instituto de Letras da UFRGS, Porto Alegre, RS, n. 17, p. 51-66, out/dez 2004.

REUILLARD, P. C. R.; KILIAN, C. K. Combinatórias léxicas especializadas de direito ambiental em uma base de dados para tradutores. In: ISQUERDO, A. N.; DAL CORNO, G. O. M. (Org.). *As ciências do léxico:* lexicologia, lexicografia, terminologia. Vol. VII. Campo Grande: Ed. UFMS, 2014, p. 473-485

[SBA] SOCIEDADE BRASILEIRA DE ANATOMIA. Federative Committee on Anatomical Terminology (FCAT) / Comissão Federativa da Terminologia Anatômica (CFTA). *Terminologia anatômica:* Terminologia anatômica internacional. 1. ed. (brasileira). São Paulo: Manole, 2001.

SILVA E TEIXEIRA, R. de B. *Termos de (onco)mastologia:* uma abordagem mediada por *corpus*. 2010. 365 f. Dissertação (mestrado em Letras). Faculdade de Filosofia, Comunicação, Letras e Artes, Pontifícia Universidade Católica de São Paulo, São Paulo, 2010. Disponível em: <https://tede.pucsp.br/bitstream/handle/13496/1/Rosana%20de%20Barros%20Silva%20e%20Teixeira.pdf>. Acesso em: 18 out. 2017.

TEIXEIRA, E. D. *A Lingüística de* Corpus *a serviço do tradutor*: proposta de um dicionário de Culinária voltado para a produção textual. 2008. 439 f. Tese (doutorado em Letras). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2008. Disponível em: <http://www.teses.usp.br/teses/disponiveis/8/8147/tde-16022009-141747/pt-br.php>. Acesso em: 18 out. 2017.

Dicionário Olímpico: a semântica de *frames* encontra a lexicografia eletrônica

Olympic Dictionary: frames semantics meets electronic lexicography

Rove Chishman
Larissa Moreira Brangel
Diego Spader de Souza
Aline Nardes dos Santos
Bruna da Silva
Sandra de Oliveira

Rove Chishman – Professora no Programa de Pós-Graduação em Linguística Aplicada da Universidade do Vale do Rio dos Sinos (UNISINOS), bolsista de produtividade no CNPq – rove@unisinos.br.
Larissa Moreira Brangel – Pós-doutoranda no Programa de Pós-Graduação em Linguística Aplicada da Universidade do Vale dos Sinos (UNISINOS), bolsista PNPD/CAPES – larissabrangel@gmail.com.
Diego Spader de Souza – Doutorando do Programa de Pós-graduação em Linguística Aplicada da Universidade do Vale do Rio dos Sinos (UNISINOS), bolsista CAPES (PROSUC) – dspader-souza@gmail.com.
Aline Nardes dos Santos – Doutoranda do Programa de Pós-graduação em Linguística Aplicada da Universidade do Vale do Rio dos Sinos (UNISINOS), bolsista CAPES (PROSUC) – aline.nardes@gmail.com.
Bruna da Silva – Doutoranda do Programa de Pós-graduação em Linguística Aplicada da Universidade do Vale do Rio dos Sinos (UNISINOS), bolsista CAPES (PROSUC) – bbrunas@outlook.com.
Sandra de Oliveira – Mestranda do Programa de Pós-graduação em Linguística Aplicada da Universidade do Vale do Rio dos Sinos (UNISINOS), bolsista CAPES (PROSUC) – sandra_san05@hotmail.com.

Resumo: O presente trabalho discute desdobramentos teóricos e práticos da compilação do *Dicionário Olímpico*, elaborado pelo grupo SemanTec na ocasião dos jogos olímpicos de 2016. No seu viés teórico, o trabalho procura elucidar a possibilidade de interface entre a Lexicografia, a Semântica de *Frames* e a Linguística de *Corpus*, de modo a demonstrar como essas três esferas do conhecimento podem se articular e fornecer um arcabouço teórico consistente para a elaboração de um dicionário eletrônico sobre modalidades esportivas. No seu viés prático, o trabalho relata alguns desafios enfrentados pelos compiladores durante o período de elaboração do *Dicionário Olímpico* – desafios, esses, que ajudaram a aprimorar a reflexão sobre o fazer lexicográfico e, consequentemente, os futuros projetos a serem desenvolvidos pelo grupo SemanTec.

Palavras-chave: Semântica de *Frames*. Lexicografia eletrônica. Linguística de *Corpus*. Esportes olímpicos.

Abstract: The present work discusses theoretical and practical unfolding of the development of *Dicionário Olímpico*, created by the SemanTec research group in the occasion of the 2016 Olympic Games. In its theoretical bias, this work aims at elucidating the possibility of the interface between Lexicography, Frame Semantics and *Corpus* Linguistics, thus demonstrating how these three fields of knowledge can articulate and provide a consistent theoretical framework to the development of an electronic dictionary of sports. In its practical bias, the work narrates some of challenges faced by the developers during the process of compilation of *Dicionário Olímpico*, considering that such challenges have helped enrich the process of reflecting about the lexicographic work and consequently the future projects that are yet to be developed by the SemanTec group.

Keywords: Frame Semantics. Electronic Lexicography. *Corpus* Linguistics. Olympic Sports.

1 Introdução

O presente trabalho tem como objetivo discorrer sobre alguns desafios enfrentados pelo grupo de pesquisa SemanTec durante o processo de compilação do *Dicionário Olímpico* (CHISHMAN et al., 2016), um recurso lexicográfico eletrônico que apresenta o léxico das modalidades olímpicas, utilizando a noção de *frame* (ou cenário) como princípio organizador. O *Dicionário Olímpico*, assim como o *Field: Dicionário de Expressões do Futebol*, é fruto de um projeto mais amplo, cujo propósito é demonstrar as possibilidades de convergência entre Semântica de *Frames* e Lexicografia. Conforme será apresentado nas páginas que seguem, essa convergência imprime na obra uma série de características atípicas de serem encontradas nos produtos lexicográficos em geral, o que confere ao *Dicionário Olímpico* um caráter diferenciado no que diz respeito a sua estruturação e fundamentação teórica.

Além da interface estabelecida com a Semântica de *Frames*, o projeto também buscou contemplar a teoria lexicográfica e a Linguística de *Corpus* dentre os recursos teóricos e metodológicos do dicionário. Sobre esse aspecto, é importante ressaltar que, ainda que as três áreas do conhecimento ora mencionadas possam

ter assumido frentes em comum, combinando-se para delinear determinados aspectos da obra, cada uma acabou por atuar de maneira mais específica em determinados pontos da pesquisa. Assim, por exemplo, a teoria lexicográfica auxiliou a consolidar e a sistematizar a forma de apresentação das informações oferecidas pela obra, servindo de suporte para a escolha e para a organização do conteúdo do dicionário no âmbito macro, médio e microestrutural. A Linguística de *Corpus*, por sua vez, elucidou, por intermédio de um acervo textual exaustivo, a gama de termos e de conceitos essenciais do Domínio Olímpico, auxiliando a equipe a entender e a explanar o conhecimento que subjaz cada um dos esportes lematizados. A Semântica de *Frames*, por fim, apresentou-se como o marco teórico sobre o qual foi desenvolvida a estrutura do dicionário, o que resultou em uma forma de apresentação e de disposição das informações bastante particular, conforme já mencionado. Com vistas a compartilhar com o público um pouco da *expertise* adquirida durante a compilação do *Dicionário Olímpico*, apresentamos e discutimos, nas páginas que seguem, esses e outros aspectos do processo de elaboração da obra lexicográfica.

Assim, iniciamos apresentando a Semântica de *Frames* e a sua relevância como aporte teórico escolhido como princípio organizador do *Dicionário Olímpico*. Na seção 3, discorremos acerca da aplicação da teoria fillmoriana em um recurso lexicográfico eletrônico. A seção 4 se volta em particular para os recursos computacionais e métodos empregados no desenvolvimento do dicionário. Coube à seção 5 apresentar o processo de seleção de unidades lexicais e seus principais desafios. Antes de apresentar as considerações finais, tecemos, na seção 6, um par de considerações sobre o projeto de compilação de nosso novo instrumento lexicográfico, o Dicionário Paraolímpico.

2 Semântica de *Frames* e alguns desdobramentos

Como foi dito na introdução, o *Dicionário Olímpico* segue o arcabouço teórico-metodológico da Semântica de *Frames* como princípio organizador da estrutura do dicionário. Nesta seção, o objetivo é apresentar a Semântica de *Frames* como modelo teórico pertencente ao quadro da Linguística Cognitiva e discutir a relevância dessa teoria para o desenvolvimento de recursos lexicográficos como o *Dicionário Olímpico*.

A Semântica de *Frames* é uma teoria desenvolvida pelo linguista norte-americano Charles J. Fillmore durante os anos 1970 e 1980. Hoje, é uma das teorias que integram a Linguística Cognitiva (e, mais especificamente, a Semântica Cognitiva), sendo considerada uma de suas principais hipóteses para a descrição do significado (SALOMÃO et al., 2013).

O movimento conhecido como Linguística Cognitiva surge no fim da década de 1970 a partir do descontentamento de um grupo de pesquisadores (dentre os quais destacam-se nomes como George Lakoff, Gilles Fauconnier e Ron Langacker, entre outros) com os paradigmas linguísticos vigentes à época, a saber a Semântica Formal e a Linguística Gerativa. A Linguística Cognitiva parte do princípio de que a linguagem é uma capacidade cognitiva humana e funciona em conjunto com as demais habilidades cognitivas, como a percepção, a memória, a visão etc. Além disso, defende a hipótese da corporeidade ou cognição corporificada (cf. JOHNSON, 1987; LAKOFF; JOHNSON, 1999). A tese da corporeidade diz respeito a uma corrente de pensamento que estabelece que a cognição está sempre em uma relação intrínseca e de mão dupla com a nossa experiência mundana. Isto é, a forma como agimos no mundo afeta a cognição e vice-versa. As nossas experiências e o ambiente em que vivemos moldam o pensamento e a forma como nos expressamos, inclusive através da linguagem. Nesse sentido, a semântica e a pragmática, antes negligenciadas, acabam se tornando um dos pontos centrais das pesquisas em Linguística Cognitiva.

A Semântica de *Frames* se encaixa nessa perspectiva, uma vez que entende o significado linguístico como algo construído a partir do uso da linguagem e da visão que os falantes têm do mundo que os rodeia. Para Fillmore (1982), isso se traduz no conceito de *frame* semântico. Segundo o linguista (1982), um *frame* designa um sistema de conceitos inter-relacionados, de modo que a compreensão de um desses conceitos pressupõe o entendimento do sistema como um todo. Quando nos deparamos com uma determinada forma linguística (seja no nível lexical, morfológico ou sintático), a compreensão dessa forma se dá através do acesso a uma estrutura de conhecimento, isto é, a um *frame*. Nesse sentido, *frames* são como "porções" de conhecimento de mundo que organizam informação referente a uma determinada situação, instituição ou evento social. Para Croft e Cruse (2004), Fillmore introduz os *frames* na teoria semântica não só como uma estratégia para organizar conceitos, mas também como uma forma de repensar fundamentalmente os propósitos da semântica linguística. Fillmore, no texto *Frames and the semantics of understanding*, de 1985, categoriza a Semântica de *Frames* como sendo uma "semântica da compreensão", em oposição a teorias formalistas, que formam as "semânticas da verdade". O objetivo da Semântica de *Frames*, segundo Petruck (1992), é justamente evidenciar as continuidades entre o significado linguístico e a experiência humana, descobrindo e analisando a forma como nos compreendemos através da linguagem e como compreendemos o mundo ao nosso redor. Vale ressaltar que, nos primeiros estágios da teoria, durante os anos 1970, Fillmore apostava na distinção entre o que seria o *frame* e o que seria uma *cena*. Segundo o linguista (cf. FILLMORE, 1975, 1977), a cena daria conta do nível conceptual, sendo o sistema de conceitos em si, isto é, a porção de conhecimento. O *frame*, por sua vez, se limitaria a um sistema de escolhas linguísticas, ou seja, o conjunto

de formas que evocam a estrutura conceptual. Anos mais tarde, durante os anos 1980, Fillmore abandonou essa dicotomia, e o conceito de *frame* passou a integrar tanto o nível conceptual ou cognitivo quanto o nível linguístico. A partir disso, o *frame* é, portanto, uma esquematização de determinada experiência que é acessada (ou evocada) por certos itens da língua. Recorrendo a um exemplo clássico, usado por Croft e Cruse (2004), para entender a palavra *garçom*, por exemplo, é preciso retomar uma estrutura maior do que a palavra em si, que, nesse caso, seria um *frame* de restaurante ou lanchonete. "Garçom" faz parte de um domínio em que se relaciona com outros elementos, como *cardápio*, *conta*, *cliente*, *chef* etc. Nesse sentido, o *frame* é como uma foto ou uma cena; ele captura determinado evento ou ação. Entendemos *garçom*, portanto, porque entendemos o seu papel no *frame* em que está inserido, e, dessa forma, entender o *frame* pressupõe compreender todos os outros elementos ao redor de *garçom*.

Outro exemplo interessante que nos auxilia a entender o conceito de *frame* é o da palavra *ponto* do português brasileiro. Considerem-se as seguintes sentenças:

(i) O ônibus passa no **ponto** às 15h30min.
(ii) Não quero falar sobre isso e **ponto**!
(iii) Entendi o seu **ponto**, porém discordo.
(iv) **Pontos** colineares pertencem à mesma reta.

A noção de *frame* nos permite compreender como, nos quatro exemplos listados acima, a palavra *ponto* assume significados diferentes. Em (i), **ponto** se refere a uma localização geográfica, a um espaço específico; nesse sentido, a palavra é entendida nos termos de um *frame* de espaço. Em (ii), contudo, **ponto** serve para especificar o fim de uma discussão, representando, possivelmente, um *frame* de ortografia, uma vez que destaca a noção de ponto final. Na sentença (iii), contudo, tem-se outra ideia: **ponto** significa ideia ou argumento, sendo assim um *frame* de argumentação. Por último, em (iv), **ponto** assume um significado especializado, pois está sendo utilizado como uma noção pertencente à matemática.

Os *frames* são, *grosso modo*, apenas um termo para se referir à estrutura conceptual. As formas linguísticas são pontos de acesso para estruturas de conhecimento armazenadas na mente. Tais estruturas são, portanto, *frames*. Por que, ao interpretarmos a frase *Joana assoprou as velinhas*, automaticamente imaginamos uma festa de aniversário? Porque, por mais que a palavra *aniversário* sequer apareça, a frase na sua totalidade descreve uma ação que nos remete a um evento específico, neste caso a festa de aniversário.

Para um recurso como o *Dicionário Olímpico*, cujo objetivo é descrever o léxico e o funcionamento de modalidades esportivas, a Semântica de *Frames* oferece a possibilidade de abordar os significados das palavras e expressões a partir dos *frames*, que funcionam, de fato, como cenas dos eventos e ações que compõem

cada um desses esportes. A relevância da Semântica de *Frames* para a Lexicografia e para o *Dicionário Olímpico* está na forma como ela explica o significado das palavras a partir do contexto em que tais palavras ocorrem.

Importa ressaltar, contudo, que a relação entre a Semântica de *Frames* e a Lexicografia não iniciou aqui. Assim, nos próximos parágrafos, apresentamos brevemente a plataforma *FrameNet*, projeto de aplicação da Semântica de *Frames* à Linguística Computacional e à Lexicografia.

Inspirados pela Semântica de *Frames*, Fillmore e pesquisadores associados iniciaram um projeto de aplicação da teoria à computação e à Lexicografia. Surge assim, em 1997, a plataforma *FrameNet* (https://framenet.icsi.berkeley.edu/fndrupal/). Desenvolvida no International Computer Science Institute (ICSI) da Universidade da Califórnia em Berkeley, a *FrameNet* é uma base de dados lexical cujo objetivo é prover informação sintático-semântica acerca do léxico da língua inglesa com base em *frames* semânticos. O trabalho da *FrameNet* consiste, basicamente, na descrição de unidades lexicais (doravante ULs) a partir da identificação e descrição de seus *frames*. Uma UL[1], para a *FrameNet*, designa o pareamento de uma forma linguística e um *frame*. A descrição do *frame*, por sua vez, inclui outros conceitos importantes, como *frame element* (elemento de *frame*) e *frame to frame relations* (relações entre *frames*).

Consideremos a imagem a seguir:

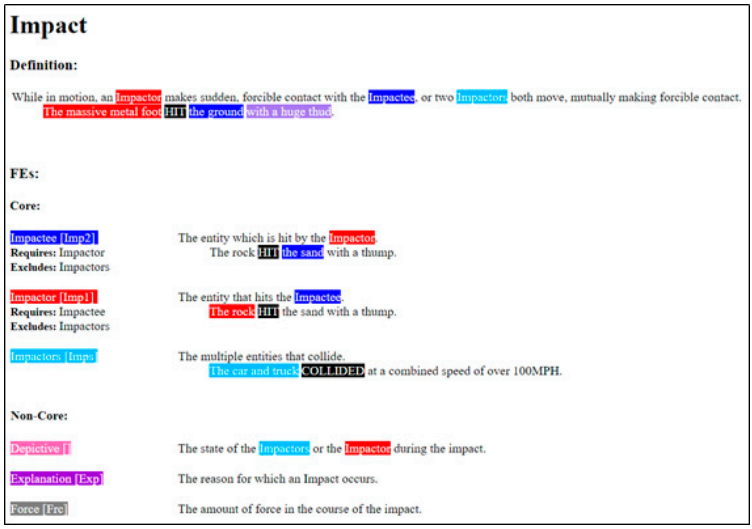
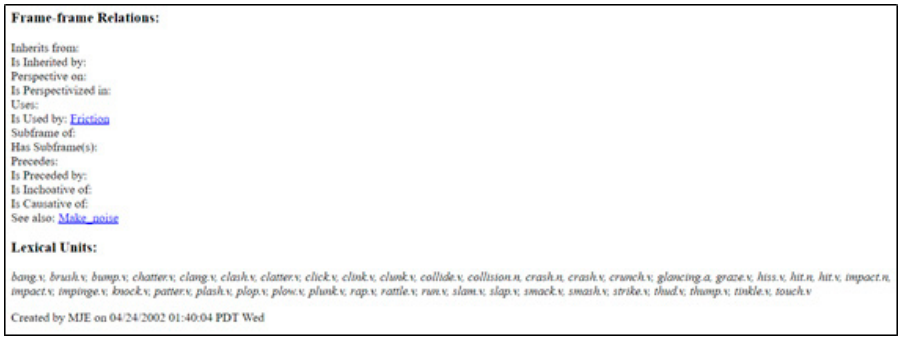


Figura 1 – *Frame* Impact
Fonte: *FrameNet*

[1] Neste trabalho, também se utiliza "palavra" como sinônimo de "unidade lexical".

A Figura 1, que retrata o *frame* Impact da *FrameNet*, contém, como podemos ver, uma definição (ou glosa), e os elementos de *frame*. A primeira característica que podemos apontar é a separação dos EFs em duas categorias distintas: *core* (centrais) e *non-core* (periféricos). Os elementos de *frame* centrais são, segundo Ruppenhofer et al. (2010), os elementos obrigatórios para a realização de um determinado *frame*. Para Impact, eles são: algo que sofre um impacto (*empactee*), e algo que impacta ou colide (*impactor* e *impactors*). Os elementos periféricos, por sua vez, designam os que não são obrigatórios, mas que costumam aparecer ao redor do *frame*, sendo eles: o estado dos elementos centrais no momento do impacto (*depictive*), a explicação para o evento (*explanation*) e a força do impacto (*force*). Os elementos de *frame* estão, de certa forma, na interface entre a semântica e a sintaxe. De um lado, são a versão da *FrameNet* para os papéis semânticos; de outro, também podem assumir formas similares aos papéis temáticos. Os elementos de *frame* assumem posições sintáticas específicas e, por isso, podem ser utilizados em estudos sobre predicação e valência verbal. Outro ponto a ser mencionado aqui é que a glosa ou definição é composta utilizando os elementos de *frame* e os marcando com suas respectivas cores. Vejamos agora a imagem abaixo:


Frame-frame Relations:

Inherits from:
Is Inherited by:
Perspective on:
Is Perspectivized in:
Uses:
Is Used by: Friction
Subframe of:
Has Subframe(s):
Precedes:
Is Preceded by:
Is Inchoative of:
Is Causative of:
See also: Make_noise

Lexical Units:

bang.v, brush.v, bump.v, chatter.v, clang.v, clash.v, clatter.v, click.v, clink.v, clunk.v, collide.v, collision.n, crash.n, crash.v, crunch.v, glancing.a, graze.v, hiss.v, hit.n, hit.v, impact.n, impact.v, impinge.v, knock.v, patter.v, plash.v, plop.v, plow.v, plunk.v, rap.v, rattle.v, run.v, slam.v, slap.v, smack.v, smash.v, strike.v, thud.v, thump.v, tinkle.v, touch.v

Created by MJE on 04/24/2002 01:40:04 PDT Wed

Figura 2 – Relações do *frame* Friction e do Make_noise
Fonte: *FrameNet*

No que concerne às relações entre *frame*, para este cenário específico, a *FrameNet* postula duas relações: uma relação de uso com o *frame* Friction e uma relação do tipo "veja também", com o *frame* Make_noise. As relações entre os *frames* formam mapeamentos entre as diferentes estruturas. Esses elementos, tanto em relação aos nomes a elas atribuídos quanto à própria dinâmica das relações em si, apresentam certo nível de complexidade que é compatível com o conhecimento que o público-alvo da *FrameNet* detém. As relações entre *frames*, que também podem ser visualizadas a partir de uma ferramenta visual denominada FrameGrapher, acabam determinando, na realidade, relações ontológicas entre as estruturas conceptuais (SCHEFFCZYK; BAKER; NARAYANAN, 2010). Um

exemplo disso é a relação de herança, que estabelece *frames* filhos, que herdam características de *frames* pais (Arriving é *frame* filho do *frame* pai Motion, por exemplo). Uma discussão muito mais aprofundada sobre as relações entre *frames* pode ser encontrada em Fillmore e Baker (2010).

A partir do que foi exposto, podemos perceber que os dados disponibilizados pela *FrameNet* e a forma como são apresentados evidenciam o interesse em atender um público de linguistas. O *Dicionário Olímpico*, por outro lado, é concebido com a finalidade de atender um tipo diferente de usuário. A plataforma *FrameNet*, que inspirou também a criação de projetos similares para línguas como o espanhol, o alemão, o português brasileiro, entre outras, serve aos propósitos de auxiliar a pesquisa linguística e linguístico-computacional (oferece uma linguagem legível por humanos e uma legível por máquinas) e de enriquecer a prática lexicográfica. Não pode, contudo, servir como modelo de dicionário baseado em *frames*.

Nesse sentido, visando à aplicação da Semântica de *Frames* a recursos lexicográficos voltados para o usuário que não está familiarizado com teorias linguísticas, a próxima seção discute a adaptação do modelo desenvolvido por Fillmore à Lexicografia.

3 Aplicação da Semântica de *Frames* no *Dicionário Olímpico*

Enquanto, no projeto *FrameNet*, a aplicação da Semântica de *Frames* se dá na interface com a Lexicografia Computacional, no *Dicionário Olímpico*, a interface é feita com a Lexicografia tradicional[2]. Por essa razão, muitas das escolhas feitas no desenvolvimento da *FrameNet* não poderiam ser simplesmente replicadas em um dicionário de natureza distinta, como é o caso do *Dicionário Olímpico*. O objetivo desta seção, portanto, é o de tratar das adaptações exigidas pela aplicação da Semântica de *Frames* a um dicionário tradicional, enfatizando a importância que os aspectos relacionados ao desenvolvimento de um dicionário do tipo tradicional e eletrônico exerceram nesse processo de tomada de decisões.

Segundo Atkins e Rundell (2008), há, no processo de desenvolvimento de um dicionário, um estágio denominado "pré-Lexicografia", que corresponde ao período em que todas as decisões que determinam a identidade do produto lexicográfico são tomadas. Nessa fase pré-lexicográfica, é a identidade do público que o dicionário deseja atender que embasa muitas das escolhas relacionadas à

[2] Uma vez que o termo "tradicional" pode entrar em conflito com a aplicação de uma teoria semântica cognitiva, esclarecemos que nosso objetivo, ao utilizá-lo, é o de contrapor a prática de desenvolvimento de dicionários que se volta, predominantemente, para a descrição dos significados das palavras ao desenvolvimento de "sistemas de computação utilizados para a compreensão e geração de linguagem natural" (FABER BENÍTEZ et al., 1998, p. 2), de que se ocupa a Lexicografia Computacional.

seleção de informações que irão compor o dicionário e a forma como serão apresentadas. Afinal, a atividade de consulta deve constituir um ato de comunicação bem-sucedido entre lexicógrafo e usuário (ATKINS, 2008).

Por esse motivo, há diferenças entre o grau de complexidade que a *FrameNet* e o *Dicionário Olímpico* exibem. Este último constitui um tipo de recurso lexicográfico que, ao mesmo tempo em que mantém uma relação íntima com a Semântica de *Frames*, se insere numa proposta voltada para um público não especializado e bastante heterogêneo.

Sendo assim, a preocupação em desenvolver um recurso *user-friendly* fez com que o processo de desenvolvimento do *Dicionário Olímpico* levasse em conta a necessidade de (i) adaptar informações que aparecem na *FrameNet*, de modo a se tornarem mais facilmente compreendidas pelo consulente leigo e (ii) suprimir informações que não seriam relevantes para esse tipo de usuário. Vale destacar que muitas dessas adaptações aparecem, ainda que de modo distinto, no *Field: Dicionário de Expressões do Futebol*, que foi desenvolvido na mesma interface em que se insere o *Dicionário Olímpico*.

Uma primeira adaptação que se mostrou central diz respeito ao modo de se referir às estruturas que organizam o dicionário. Conceitos como os de *frame* e de unidade lexical são utilizados no projeto *FrameNet* com estatuto de metalinguagem, porque o conteúdo que veiculam é facilmente recuperado pelos usuários linguistas. Porém, em um dicionário tradicional, esses conceitos poderiam, em alguma medida, intimidar o consulente leigo.

Por essa razão, o *Dicionário Olímpico* mantém a decisão tomada no processo de desenvolvimento do *Field* de substituir os conceitos acima pelas noções de cenário[3] e de palavra (respectivamente), que fazem parte do vocabulário do consulente comum e, por isso, são mais facilmente compreendidas por esse público. Assim, ao acessar o dicionário, o usuário pode optar por pesquisar por um cenário ou palavra com a ferramenta de busca ou acessar uma das modalidades e, assim, visualizar as listas de cenários e de palavras da modalidade em questão.

É importante destacar que o significado da palavra *cenário* (no sentido de *situação*) está relacionado, em alguma medida, com a noção de *frame* e, por isso, a noção de *cenário* foi considerada adequada no contexto do *Dicionário Olímpico*. A noção de *palavra*, por sua vez, foi adotada por estar intrinsecamente relacionada à experiência de usuários de dicionários tradicionais.

Mais importante do que manter a nomenclatura teórica é fazer com que o consulente entenda a proposta de que os significados das palavras são determinados de acordo com um conhecimento que lhes serve de fundo, um "cenário"

[3] A escolha pelo termo "cenário" para se referir aos *frames* do *Dicionário Olímpico* não representa uma tentativa de recuperar a distinção feita por Fillmore (1975), apresentada na seção 2.

ou um roteiro subjacente. Assim, a escolha foi feita com o intuito de popularizar as noções teóricas.

Uma segunda adaptação diz respeito ao modo como são expressas as relações entre cenários no *Dicionário Olímpico*. Tendo em vista que o modo como a *FrameNet* exibe as relações entre *frames* pressupõe certa familiaridade com o arcabouço teórico da Semântica de *Frames*, o *Dicionário Olímpico* apresenta essa informação de modo a dar destaque a outras dimensões das relações entre cenários, como a organização e a classificação dos eventos, por exemplo.

A título de comparação, a imagem abaixo, que apresenta as relações do *frame* Commercial_transaction, mostra como são exibidas as relações entre *frames* na *FrameNet*.


Frame-frame Relations:

Inherits from: Reciprocality
Is Inherited by:
Perspective on:
Is Perspectivized in:
Uses:
Is Used by:
Subframe of: Commerce_scenario
Has Subframe(s): Commerce_goods-transfer, Commerce_money-transfer
Precedes:
Is Preceded by:
Is Inchoative of:
Is Causative of:
See also:

Figura 3 – Relações do *frame* Commercial_transaction
Fonte: *FrameNet*

As relações do *Dicionário Olímpico* nada têm a ver com noções teóricas subjacentes. A palavra que expressa o tipo de relação estabelecida entre os cenários carrega o significado do senso comum. A imagem a seguir mostra as relações apresentadas para o cenário Arbitragem da modalidade voleibol.

Figura 4 – Cenários relacionados do cenário Arbitragem (voleibol)
Fonte: CHISHMAN et al., 2016

Os nomes das relações entre cenários, no *Dicionário Olímpico*, não foram predefinidos. Ou seja, não há, como na *FrameNet*, uma lista de relações de acordo com a qual as relações entre *frames* são classificadas. Um dos motivos que reforçou a decisão pela adaptação na apresentação das relações foi o *feedback* dos colaboradores (especialistas nas modalidades). Por não estarem familiarizados com os conceitos teóricos, relataram dificuldade em compreender as relações quando estabelecidas de acordo com os critérios da *FrameNet*.

Em terceiro lugar, foram necessárias adaptações no que se refere aos elementos de *frame*. Como vimos na seção anterior, a *FrameNet* classifica um elemento de *frame* como sendo central, periférico ou extratemático. A importância de fornecer essa informação para o usuário se dá porque o recurso, além de exibir informações semânticas das palavras da língua inglesa, assim como o *Dicionário Olímpico* faz para o léxico de cada umas das modalidades olímpicas, também exibe informações sintáticas, como padrões de coocorrência de palavras, informações essas que mantêm estreita relação com o fato de a *FrameNet* constituir uma ferramenta também voltada para Processamento da Linguagem Natural (PLN).

O *Dicionário Olímpico*, por sua vez, não teve a intenção de apresentar ao seu público-alvo informações sobre valência verbal ou aspectos sintáticos que não lhes seriam úteis. Por essa razão, os elementos que compõem os cenários do *Dicionário Olímpico* não equivalem aos elementos de *frame* da *FrameNet*. Ao invés disso, no *Dicionário Olímpico*, esses elementos podem ser vistos como itens que compõem os cenários ou como participantes que atuam nesses cenários.

Como mostra a imagem abaixo, a glosa de cada cenário do *Dicionário Olímpico* traz em destaque (negrito) itens e participantes do cenário, palavras consideradas centrais para a compreensão do cenário.

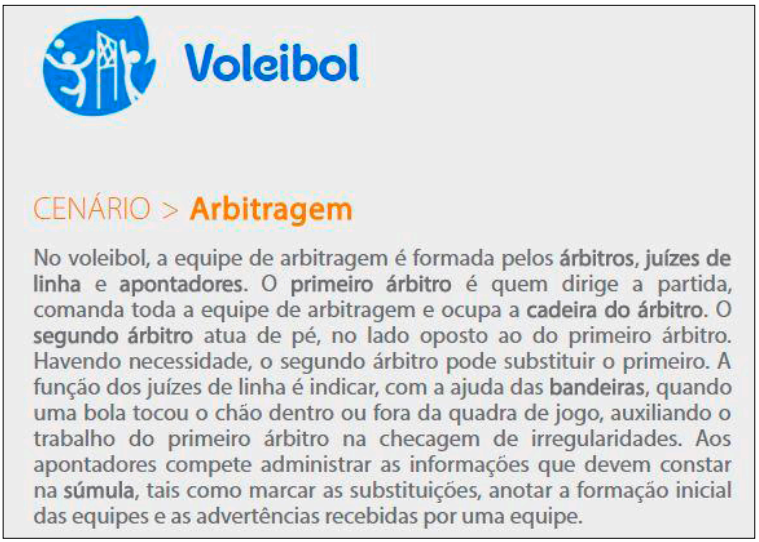


Figura 5 – Glosa do cenário Arbitragem
Fonte: CHISHMAN et al., 2016

Um último ponto a ser mencionado se refere aos traços que o *Dicionário Olímpico* apresenta que estão relacionados à aplicação da Semântica de *Frames* ao desenvolvimento de um dicionário digital. Se, por um lado, o projeto *FrameNet* coloca em evidência a possibilidade de relação entre Semântica de *Frames* e Lexicografia, por outro, há também o destaque para o quão bem a Lexicografia Eletrônica parece combinar-se com a proposta de descrição lexical baseada em *frames*. Fillmore (2003), ao estabelecer um comparativo entre a aplicação da noção de *frame* a um dicionário digital e a um dicionário impresso, demonstra estar ciente dos aspectos que aproximam a Semântica de *Frames* da Lexicografia Eletrônica e dos aspectos que a distanciam da Lexicografia impressa.

No desenvolvimento do *Dicionário Olímpico*, acabamos por nos envolver, mesmo que de modo indireto, com questões relacionadas ao desenvolvimento de dicionários digitais. Uma análise rápida poderia revelar que esse envolvimento se deu de forma ingênua e despretensiosa. Porém, um olhar mais atento revela que os aspectos do desenvolvimento do *Dicionário Olímpico* que demonstram, em alguma medida, preocupação com as questões centrais para a Lexicografia Eletrônica estão intimamente relacionados aos pressupostos teóricos da Semântica de *Frames*. Ou seja, os recursos utilizados que podem ser interpretados como uma forma de atender às preocupações da Lexicografia Eletrônica são, na verdade, aspectos que buscam evidenciar a forma como se dá a construção do significado na Semântica de *Frames* e na Linguística Cognitiva como um todo.

Fillmore (2003) destacou importantes afinidades entre a Semântica de *Frames* e a Lexicografia Eletrônica, tais como a possibilidade de inserção de *hiperlinks* e a consequência disso, que torna desnecessária a repetição de informações, em comparação aos dicionários impressos. A bibliografia da Lexicografia Eletrônica, por sua vez, revela mais algumas afinidades entre os campos, dentre as quais se destaca a utilização de *corpus* como fonte de evidência empírica para a geração de listas de palavras e extração de sentenças-exemplo e outras possibilidades.

Na análise do *Dicionário Olímpico*, um aspecto que se destacou foi a afinidade entre a ideia de conhecimento enciclopédico, que está na base da Semântica Cognitiva e, de modo especial, na base da Semântica de *Frames* (EVANS; GREEN, 2006), e a possibilidade de utilização de recursos multimodais oferecida pelo meio digital. Nesse sentido, o *Dicionário Olímpico* apresenta fotos e mapas conceituais, cujas inserções se deram no intuito de incluir aspectos relacionados às experiências de base corporal, tais como memória visual, e a fenômenos de organização do conhecimento, como a categorização (LAKOFF, 1987). As imagens a seguir exemplificam o uso das fotos e dos mapas no *Dicionário Olímpico*.

Figura 6 – Foto que ilustra o cenário Bloqueio do voleibol
Fonte: CHISHMAN et al., 2016. Foto: Gaspar Nóbrega

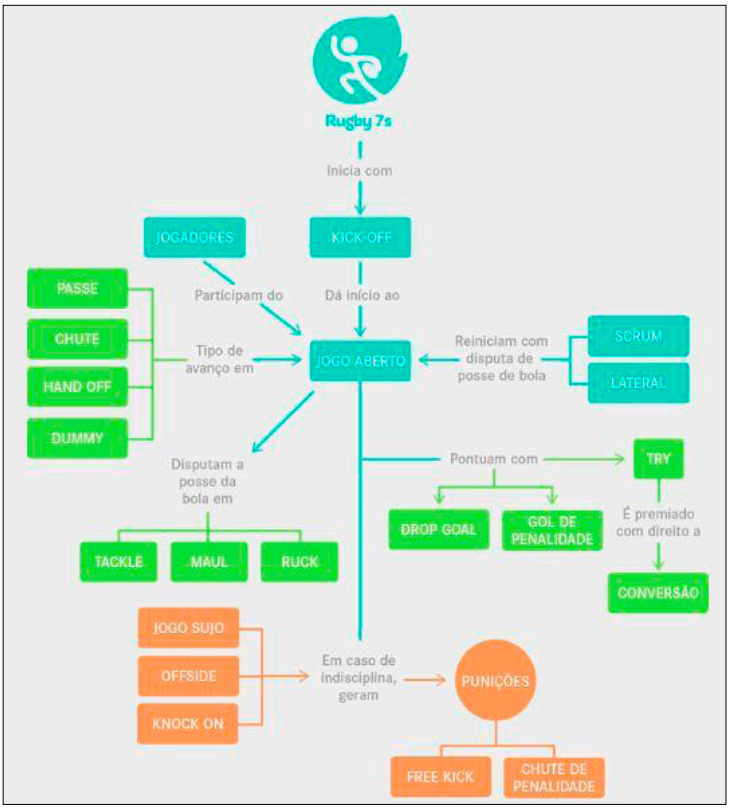


Figura 7 – Mapa da modalidade rugby 7s
Fonte: CHISHMAN et al., 2016

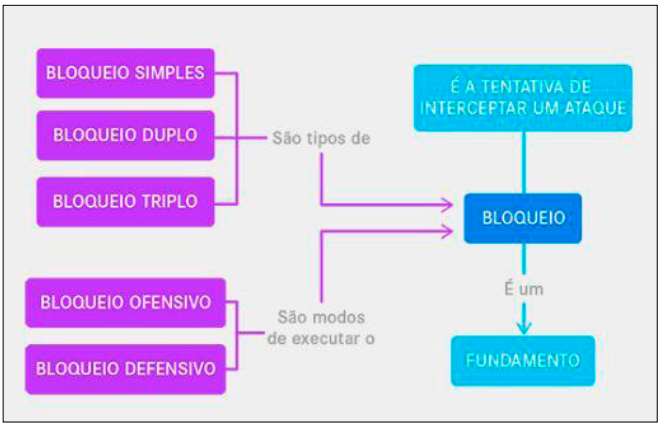


Figura 8 – Minimapa do cenário Bloqueio do voleibol
Fonte: CHISHMAN et al., 2016

Além desses recursos, o dicionário apresenta uma seção intitulada "Você sabia?", como mostra a imagem abaixo, que traz curiosidades sobre o universo olímpico.

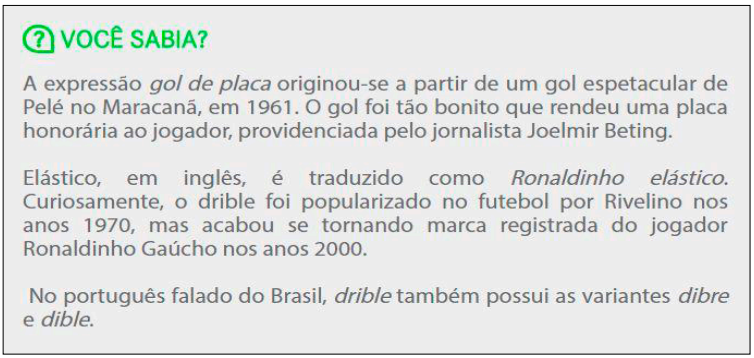


Figura 9 – Seção "Você sabia?" do futebol
Fonte: CHISHMAN et al., 2016

A descoberta dessas afinidades abriu caminho para uma investigação mais profunda acerca da interface entre Semântica de *Frames* e Lexicografia Eletrônica, com o objetivo de explorar de que forma o potencial para a descrição do significado lexical da Semântica de *Frames* e os preceitos da Lexicografia Eletrônica, especialmente a tradição em pesquisa sobre o uso de dicionários, podem se combinar de modo a desenvolver um produto que atenda aos interesses de ambas as áreas.

As adaptações relativas ao meio digital distinguem ainda mais o *Dicionário Olímpico* da *FrameNet*, deixando bastante clara a influência que os interesses do

público-alvo exercem sobre a tarefa de delinear a identidade do produto lexicográfico. Com base no que foi exposto, foi possível fornecer um panorama dos pontos que caracterizam o *Dicionário Olímpico* de modo amplo, que atendeu desde as adaptações nos nomes das estruturas que organizam o dicionário até os recursos que auxiliam na compreensão dos cenários.

A próxima seção trata dos recursos e métodos utilizados na compilação do *Dicionário Olímpico*, descrevendo os procedimentos realizados na etapa monolíngue.

4 Recursos e métodos

Seguindo pressupostos da Linguística de *Corpus* e da Lexicografia contemporânea, os recursos lexicográficos construídos no âmbito do grupo SemanTec são elaborados a partir de dados autênticos, efetuando-se, conforme recomendam autores como Atkins e Rundell (2008), as adaptações necessárias – consoante o seu público-alvo e o propósito de construir *frames* semânticos relativos às ações que caracterizam cada esporte. Nesta seção, são abordados alguns desdobramentos relativos à coleta de *corpora*, ao uso de ferramentas para processamento e à definição dos métodos, com foco nas compilações e descrições monolíngues, abordando também a necessidade de adaptação de alguns procedimentos metodológicos a partir dos desafios enfrentados ao longo do trabalho.

No contexto de pesquisa do grupo SemanTec, o projeto que culminou na construção do *Dicionário Field* permitiu o estabelecimento de alguns critérios gerais para coleta de *corpora* e que serviram como diretriz inicial para o planejamento do *Dicionário Olímpico*, quais sejam: (a) desenho do "projeto de *corpus*" (ALUÍSIO; ALMEIDA, 2006, p. 160), começando pela busca de materiais, na modalidade escrita, que fossem representativos da temática esportiva a ser explorada lexicograficamente; (b) a partir dessa busca inicial, definição dos *sites* e dos gêneros mais tematicamente representativos do esporte, criando-se um arquivo de texto para gerenciamento coletivo de todas as fontes selecionadas; (c) coleta, nomeação, limpeza e formatação dos arquivos.

No caso do *Dicionário Olímpico*, a meta inicial consistiu na compilação de *subcorpora* com uma média de 250 mil palavras – tamanho estipulado como consideravelmente viável no que se refere ao tempo de coleta, considerando-se a necessidade de finalização dos *subcorpora* relativos aos 40 esportes conforme o cronograma estipulado pelo grupo; e, ao mesmo tempo, como suficientemente grande para a exploração em ferramentas de processamento *corpora*. Tais critérios, é claro, sempre estariam sujeitos a modificações, visto que os domínios esportivos que compõem as modalidades olímpicas são bastante heterogêneos.

No entanto, as primeiras explorações relativas à coleta de *corpora* para a construção do *Dicionário Olímpico* indicaram que apenas os esportes populares no Brasil, os quais, assim como o futebol, são consideravelmente midiatizados – a exemplo do vôlei e do basquete –, permitiam que essas etapas relativas ao desenho de cada projeto de *corpus* e a sua compilação fossem seguidas. Para esses casos, além das notícias que detalhavam as partidas – também conhecidas como *match reports* –, foi possível coletar documentos oficiais sistematizando as regras publicadas pelas respectivas confederações esportivas.

No caso de esportes com pouca ou nenhuma midiatização, e/ou nas situações em que as respectivas notícias focavam apenas nos resultados das partidas, sem trazer detalhes das ações que caracterizam a competição e que permitem a descrição de *frames* semânticos, o único documento cuja coleta foi recorrente e sistemática concerniu às regras de cada esporte. Dessa forma, em virtude da escassez de textos escritos, para esses casos, foi necessário recorrer a recursos audiovisuais que servissem como suporte adicional à descrição desses esportes, principalmente vídeos contendo a narração das partidas, ou detalhando didaticamente os movimentos característicos da modalidade. Visto que esse material multimodal não foi processado, mas sim usado como recurso de referência para fins de comparação e complementação dos poucos dados linguísticos encontrados para esses esportes, não se trata de um *corpus* de estudo, mas sim de um *corpus* de apoio – terminologia também usada por Cruz (2017) para definir os recursos impressos que foram usados em sua investigação, de modo a complementar sua pesquisa linguística em *corpus* eletrônico.

A principal ferramenta utilizada para manipulação do *corpus* de estudo foi o Sketch Engine (KILGARRIFF et al., 2004). Conforme já explanado em Chishman et al. (2014, 2016), o principal motivo para a escolha desse recurso pelo grupo, desde a construção do *Dicionário Field*, foi a integração entre as ferramentas Word Sketch e Concordance, de modo a permitir a exploração de elementos de *frame*, unidades polissêmicas e colocações, dentre outros fenômenos. Nesse contexto, destaca-se o uso da Word Sketch para explorar todas as possíveis combinatórias de um item lexical no *corpus*. A imagem a seguir ilustra a Word Sketch para *ataque* no *corpus* do futebol:

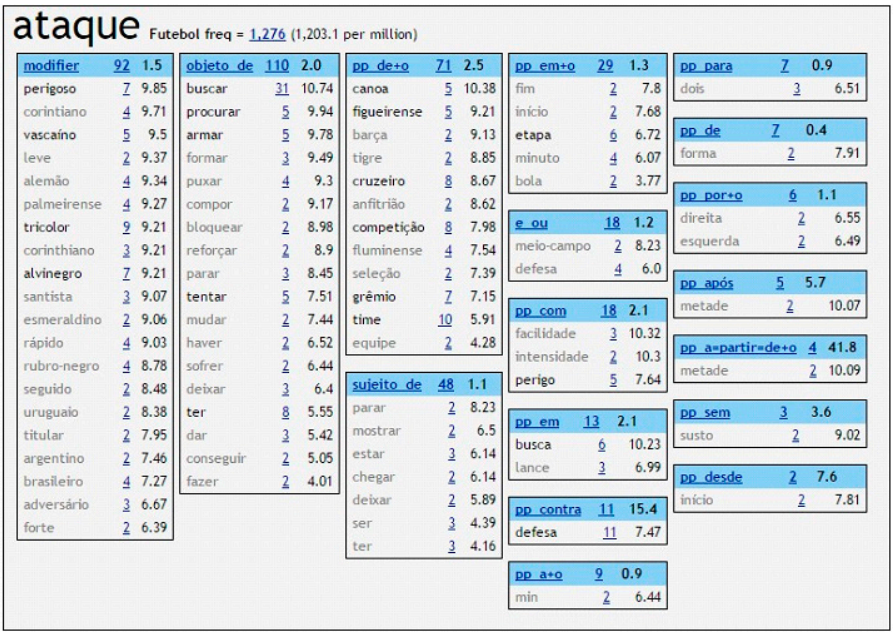


Figura 10 – Word Sketch de *ataque*
Fonte: CHISHMAN et al., 2015, p. 784

Ainda quanto ao uso do Sketch Engine na manipulação do *corpus* de estudo, é importante observar que não se trata de um *software* gratuito. Desse modo, apesar de os membros do grupo, ao longo de todo o projeto, terem usado múltiplas contas vinculadas a uma licença paga, o número de perfis não era suficiente para que todos pudessem trabalhar simultaneamente. Assim, em virtude de tal limitação, o AntConc (ANTHONY, 2014) foi utilizado em alguns momentos, por se tratar de um programa gratuito cuja versão portátil dispensa instalação, possibilitando o armazenamento do aplicativo em dispositivos móveis e em plataformas virtuais de uso coletivo.

No que se refere aos procedimentos metodológicos, as descrições lexicográficas realizadas no âmbito do grupo SemanTec têm se iniciado pelo direcionamento *top-down* – elaboram-se os *frames* semânticos para então se elencarem as unidades lexicais correspondentes, a partir dos *corpora* compilados. Além disso, as unidades extraídas dos *corpora* são analisadas qualitativamente no que concerne à evocação dos *frames* construídos, estabelecendo-se uma segunda fase de análise, cuja abordagem é *bottom-up*. Dessa forma, o conjunto das etapas metodológicas pode ser categorizado como tendo direcionamento *middle-out* (MÜLLER, 2015), pois as análises resultam da convergência entre as abordagens *top-down* e *bottom-up*.

O procedimento de descrição do *frame* realizado pelo grupo vai ao encontro das etapas iniciais estabelecidas pela *FrameNet Berkeley*, que parte de uma "caracterização informal do tipo de entidade ou situação representada pelo *frame*" (SANTOS; CHISHMAN, 2015, p. 441); ao mesmo tempo em que se diferencia de abordagens como a de Schmidt (2009), as quais têm, como único ponto de partida inicial, o léxico. Tal procedimento escolhido por Schmidt não seria possível no caso do *Dicionário Olímpico*, principalmente pelo fato de que a descrição dos esportes menos midiatizados ancorou-se consideravelmente no *corpus* de apoio, como forma de suprir as deficiências dos *corpora* de estudo.

Essa contraparte multimodal também foi crucial na primeira etapa, relativa à caracterização introspectiva do *frame*. Segundo a metodologia estabelecida por Fillmore e Baker (2009), a definição inicial do *frame* é feita a partir do conhecimento enciclopédico que o analista tem quanto aos meios linguísticos usados para descrever tal cenário. No entanto, quando se trata de um domínio especializado, principalmente no caso de esportes menos conhecidos, não era possível esboçar minimamente os *frames* correspondentes, em virtude da falta de conhecimento do grupo em relação à modalidade. Dessa forma, foi necessário recorrer ao procedimento de reconhecimento do domínio – aspecto bastante valorizado nas pesquisas em Terminologia, visto que promove uma aproximação inicial dos analistas com a área de conhecimento a ser descrita (KRIEGER; FINATTO, 2004). A organização desse estudo do domínio foi feita a partir da construção de mapas conceituais[4], que também serviram para organizar e inter-relacionar os *frames* de cada esporte.

A partir dessa etapa inicial de reconhecimento das estruturas conceptuais subjacentes aos domínios, foi elaborada uma breve definição ou glosa de cada *frame*. Em um segundo momento, as listas de palavras e de candidatas a unidades complexas de cada *subcorpora* passaram a ser exploradas e debatidas em grande grupo, contando-se também com o auxílio de especialistas – principalmente atletas do esporte e representantes de entidades oficiais – para a realização de mudanças nos mapas conceituais e nas listas de unidades lexicais. No caso das modalidades cujos *corpora* de estudo eram pouco representativos, trabalhou-se também com listas de unidades extraídas manualmente do *corpus* de apoio, as quais foram também debatidas coletivamente e avaliadas por especialistas. Esse material multimodal também foi aproveitado quando os *corpora* de estudo dos esportes menos populares se mostraram insuficientes para extração e adaptação dos exemplos relativos a cada unidade lexical. O esquema a seguir, no formato de fluxograma de processos, sintetiza esse fluxo metodológico da contraparte monolíngue, em que as três principais frentes de trabalho se inter-relacionam – definição dos *corpora*, estudo do domínio para descrição dos *frames* e seleção das unidades lexicais. Para

[4] As ferramentas utilizadas para a construção dos mapas conceituais foram o CMap Tools, disponível em <https://cmap.ihmc.us/>, e o Bubbl.us, disponível em <https://bubbl.us/>.

sua elaboração, seguindo as orientações de Dhunna e Dixit (2010), foi adotada a simbologia comumente empregada em processos de gestão e programação, estipulada pela ANSI (*American National Standards Institute*):

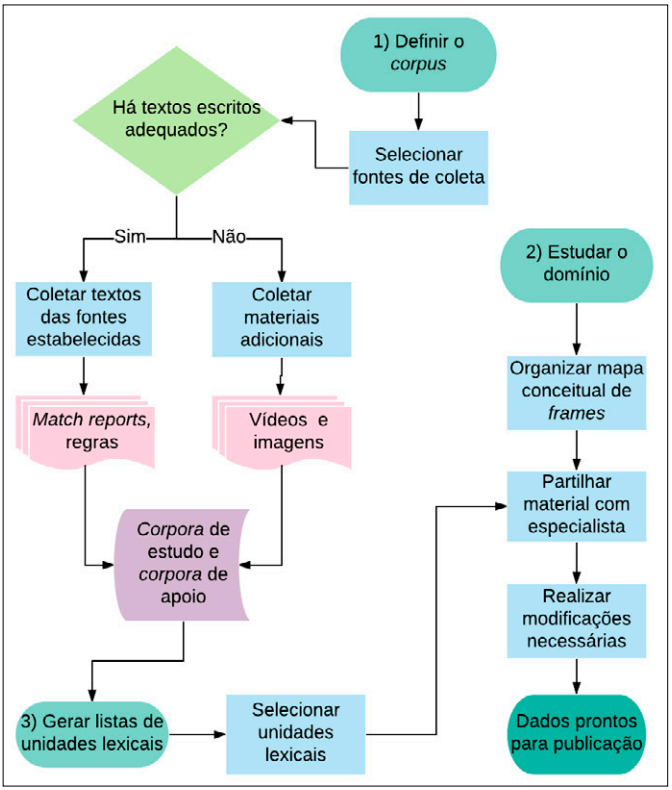


Figura 11 – Fluxo metodológico da etapa monolíngue do projeto
Fonte: Elaborada pelo Grupo SemanTec

Finalmente, é importante observar que o uso de *corpus* de apoio na construção dos *frames* também acabou diferenciando a metodologia estabelecida pelo grupo em relação aos procedimentos adotados pela *FrameNet*, visto que, "Dada a necessidade da *FrameNet* em verificar empiricamente *frames* para fins de descrição das combinatórias sintáticas e semânticas, o escopo do *frame* se reduz àquilo que suas instanciações linguísticas reproduzem" (SANTOS, 2016, p. 54). Assim, tendo em vista a escassez de material linguístico para alguns esportes, a construção do *Dicionário Olímpico* também se apoiou em instanciações audiovisuais dos *frames*, em consonância com a concepção enciclopédica de significado preconizada por Fillmore (1982).

Na próxima seção, serão abordados os fatores que influenciaram as decisões acerca da composição das listas de palavras e os aspectos a elas relacionados.

5 A seleção de unidades lexicais e seus desafios

Dando continuidade ao exposto acerca dos recursos e, em especial, da metodologia, reservamos esta seção para tratar, de forma mais detida, do trabalho concernente às unidades lexicais. Para tal, importa iniciar lembrando que, em um dicionário organizado com base nos preceitos da Semântica de *Frames*, uma unidade lexical, conforme Fillmore (1982), é o pareamento de sua forma com seu *frame*, o que pressupõe que saber o significado de uma palavra implica associá-la às estruturas de experiência ou de conhecimento subjacentes. Seguindo esse princípio, as palavras não são tratadas de forma isolada, à medida em que há referência não apenas ao *frame* (ou cenário), como também a outras palavras que remetem ao mesmo domínio.

As consequências em se assumir tal perspectiva semântica podem ser evidenciadas pela própria composição das entradas lexicais, que apresentam, entre outras informações, o cenário evocado e as unidades lexicais pertencentes ao mesmo cenário. Por estarmos em um contexto de Lexicografia Eletrônica, conforme vimos na seção 3, tanto o cenário como as palavras relacionadas são *links* para o cenário indicado ou para as entradas lexicais das palavras que fazem parte do mesmo cenário. A Figura 12, reproduzindo a tela do verbete *ala*, da modalidade basquetebol, nos mostra tal configuração.

Figura 12 – Verbete *ala* – basquetebol
Fonte: CHISHMAN et al., 2016

Outro traço do *Dicionário Olímpico* decorrente de tal compromisso fica evidente na própria organização das listas de unidades lexicais. Como exemplo, consideremos a modalidade do futebol, que traz em sua lista as unidades lexicais

ala direita e *ala esquerda* duplicadas. A justificativa, conforme se percebe ao acessar as respectivas entradas, é o fato de tais unidades lexicais evocarem cenários distintos: tanto *ala direita* como *ala esquerda* remetem aos cenários ontológicos Jogadores e Campo.

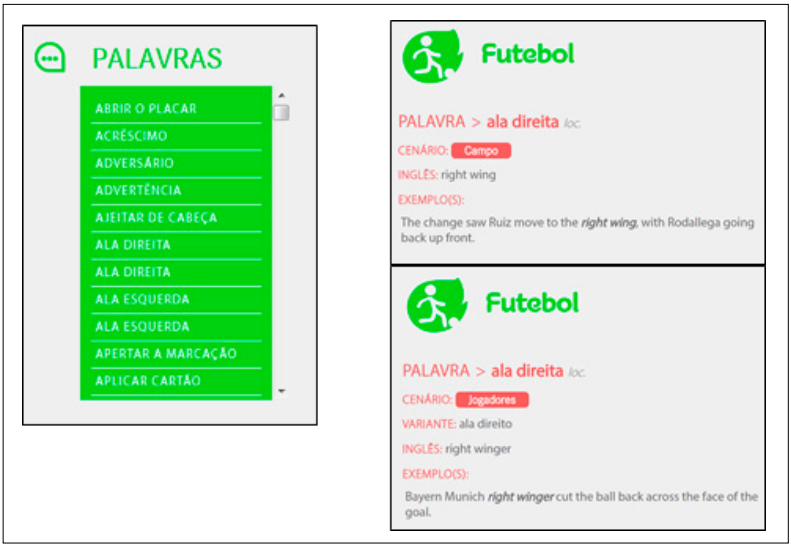


Figura 13 – Unidades lexicais duplicadas – *ala direita*
Fonte: CHISHMAN et al., 2016

Fillmore e Atkins (2010) ressaltam que é graças à assunção de que as unidades lexicais evocam *frames* que a Semântica de *Frames* se torna uma perspectiva pertinente para tratar do fenômeno da polissemia em um dicionário, como é o caso supracitado. Caso similar, também do domínio do futebol, ocorre com o verbo *marcar*, que remete aos cenários de Marcação, como em "O time da casa marcava forte", e Gol, como em "O atacante uruguaio marcou o segundo gol".

Feitas tais considerações relacionadas à adoção da noção de *frame* como princípio organizador, convém também fazer um registro sobre a desafiadora tarefa de decidir sobre quais unidades devem ou não compor a lista. Relacionada à tarefa de seleção de unidades lexicais, há uma outra igualmente crítica: escolher quais unidades lexicais devem ter estatuto de cabeças de verbete (ou *headwords*).

Conforme já exposto, um dos requisitos para fazer parte da lista é tratar-se de unidade lexical evocadora de *frame*. Importa ressaltar que, para Fillmore et al. (2013), são evocadoras de *frames* as palavras com estatuto de predicadores, tais como verbos, nomes e adjetivos. Ainda que não tenhamos tratado dos aspectos valenciais de predicação nos moldes da *FrameNet*, a seleção de nomes e verbos segue tal princípio, assim como a exclusão de categorias predicadoras menos

típicas, como é o caso de adjetivos e advérbios, e de categorias gramaticais, como preposições, artigos e conjunções.

Sobre o levantamento das unidades lexicais, os seguintes aspectos tiveram importância: (i) a construção de *corpus* comparável e (ii) a colaboração do especialista para validar casos duvidosos, assim como sugerir novos itens. Houve, no entanto, um outro desafio, este de natureza qualitativa: a dificuldade em lidar com as expressões multivocabulares, haja vista o caráter fluido de suas fronteiras.

Atkins e Rundell (2008) apresentam a seguinte classificação para as expressões multivocabulares: (i) sintagmas fixos e semifixos; (ii) expressões idiomáticas; (iii) compostos, (iv) *phrasal verbs* e (v) verbos suporte ou verbos leve. Dada a sua elevada incidência nos *corpora* do projeto, enfatizamos aqui os compostos e as construções com verbos suporte.

No caso dos compostos, os que apresentam teor idiomático e figurativo, em diferentes graus, foram mais fáceis de serem reconhecidos. É o caso de *cama de gato* (futebol), *ciclone parafuso*, *peixe espada* e *rabo de peixe* (nado sincronizado), com um alto grau de figuratividade, e *bandeja reversa*, *zona morta* (basquetebol), *ataque aéreo* (canoagem slalom) e *funil de chegada* (maratona aquática), considerados pelos autores como compostos semifigurativos. Como exemplos com verbos, podemos citar *queimar a largada* (natação) e *furar a água* (ginástica artística).

São, sem dúvida, os compostos não idiomáticos os mais difíceis, já que são semanticamente transparentes e podem ser considerados como simples usos produtivos de duas palavras. Expressões nominais como *cartão amarelo* (futebol), *partida falsa* (pentatlo), *posse de bola* (handebol) e *erro forçado* (tênis) estão entre os compostos não idiomáticos considerados de fácil reconhecimento pela equipe. *Marcar falta* e *fazer gol* (futebol), assim como *zerar o percurso* (hipismo), estes ilustrando compostos com verbos, também foram encaixados nesta categoria.

O mesmo não se pode dizer de expressões como *raia central* e *raia livre* (tiro com arco) ou *conduta incorreta*, *conduta ofensiva* e *conduta rude* (vôlei de praia). Nesses casos, a consulta aos *corpora*, em especial, às listas de frequência, foi determinante, assim como a familiaridade com a modalidade esportiva em estudo.

Como exemplos de verbos suporte inclusos na lista de unidades lexicais, temos as expressões do futebol que contêm o verbo *dar*, como *dar um carrinho*, *dar um chapéu* e *dar vantagem* (futebol).

Tal como as expressões multivocabulares, outra atividade igualmente desafiadora foi a de agregar traduções às unidades lexicais em português, considerado, no contexto do *Dicionário Olímpico*, como a língua-fonte, sendo o inglês a língua-alvo.

Uma rápida apreciação das listas de unidades lexicais em português das 40 modalidades já nos mostra um dado importante: a presença pervasiva de estrangeirismos e neologismos. Ainda que não tenhamos feito uma avaliação de cada uma dessas realidades esportivas e, por que não?, linguísticas, fica evidente que

a incidência de palavras estrangeiras é mais alta em esportes de baixa popularidade no Brasil ou em esportes claramente vinculados a determinadas culturas estrangeiras. Merecem destaque modalidades como golfe, rugby 7s, hóquei sobre grama, tênis e as ginásticas, todas contendo muitas palavras em inglês, algumas já empregadas como palavras da língua portuguesa, com modificações fonológicas ou morfológicas.

A seguir, apresentamos uma sistematização das principais situações enfrentadas ao proceder com a busca pelas traduções.

1) Unidade lexical em português na língua-fonte com unidade lexical equivalente em inglês na língua-alvo: para esportes com escassez de material no gênero *match report*, foram utilizados *corpora* comparáveis para a extração dos equivalentes de tradução. Casos de termos como *bola anulada* (rugby 7s), *corda de arco* (tiro com arco), sendo o tiro com arco um esporte de baixíssima popularidade, levaram à utilização da *web* como *corpus* (cf. GATTO, 2012) para busca das traduções.

2) Unidade lexical em português na língua-fonte sem unidade lexical em inglês: esta é uma situação bem específica de esportes muito populares no Brasil, como o futebol. O fato de ser popular, faz com que sua linguagem seja diversificada, apresentando expressões metafóricas, como *entrar de sola*, e casos especiais de termos como *gol de placa*, *gol de primeira*, os quais não possuem equivalência na língua-alvo. Para esses casos, foi mantida a unidade lexical no dicionário sem respectivo equivalente de tradução.

3) Estrangeirismo na língua-fonte: alguns esportes, ainda considerados relativamente menos populares, utilizam unidades lexicais da língua-alvo na língua-fonte. *Drive* (badminton), *birdie* (golfe), *drag flick* (hóquei sobre grama) são alguns exemplos. Para esses casos, o estrangeirismo foi mantido como unidade lexical da língua-fonte, e o significado da palavra foi esclarecido através de uma nota explicativa. No caso do *birdie*, a nota diz: "Expressão, sem equivalente em português, que indica a pontuação de um competidor que emboca a bola no buraco com uma tacada abaixo do par (número médio de tacadas definido para cada buraco).".

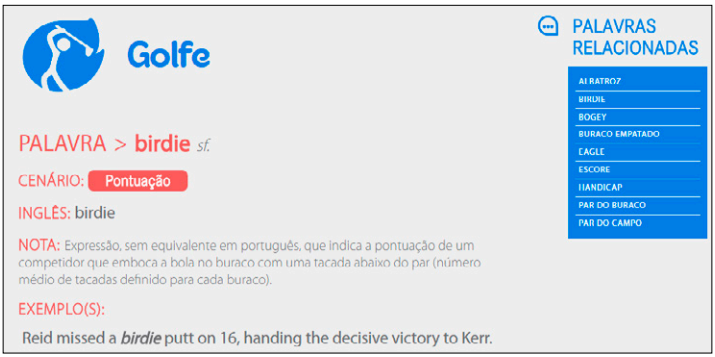


Figura 14 – Verbete *birdie* – Golfe
Fonte: CHISHMAN et al., 2016

4) Unidade lexical em português e em inglês na língua-fonte: em alguns esportes que tem forte tradição em países da língua inglesa, encontram-se diferentes situações: embora haja unidades lexicais em português e em inglês, como a em português não é usada, optou-se pela unidade lexical em inglês na língua-fonte. Termos em português como *ensaio*, *chute-livre* (rugby 7s), que correspondem respectivamente a *try* e *free-kick*, ainda que incluídos em glossários, não são usados. Isso explica porque foram mantidos os estrangeirismos na lista de unidades lexicais em português. Além dessa, outra situação problemática foi casos em que termos em português e em inglês concorrem no uso, gerando a dificuldade de escolha da unidade lexical que encabeçaria a entrada, como *lateral* e *line-out* (rugby 7s). Nesses casos, foi imprescindível a colaboração do especialista. Dessa forma, a palavra *lateral* foi incluída como cabeça do verbete, com a seguinte nota: "*Lateral* e *line-out* concorrem no uso, porém o termo em português sobrepõe-se ao termo em inglês.".

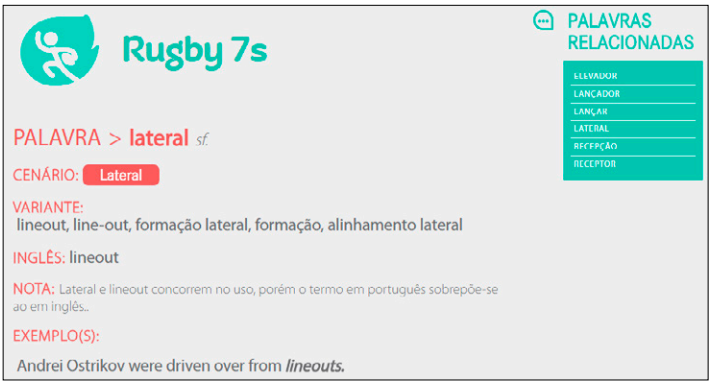


Figura 15 – Verbete *lateral* – rugby 7s
Fonte: CHISHMAN et al., 2016

5) Neologismo na língua-fonte (palavra originária da língua inglesa formando uma nova palavra em português): em alguns esportes, foram identificadas algumas palavras na língua-fonte que sofreram adaptação fonológica como *kipe* (ginástica artística) que corresponde a *kip* na língua-alvo, *jabe* (boxe) que corresponde a *jab* na língua-alvo. Outras sofreram adaptação tanto fonológica como morfológica, que é o caso de *taclear* (rugby 7s), verbo esse gerado a partir do substantivo *tackle*. São palavras em inglês aportuguesadas e adaptadas. Outro caso identificado de neologismo são as expressões mescladas, formadas parte pela língua-fonte e parte pela língua-alvo, como *pivot cossaco* (ginástica rítmica), *giro bourrée* (ginástica rítmica), *duplo-skiff* (remo), *lob curto* (tênis), *giro do scrum* (rugby 7s).

Com essas considerações, encerramos nossa exposição sobre as unidades lexicais no *Dicionário Olímpico* e nos encaminhamos para algumas reflexões finais sobre a experiência relatada no presente trabalho. Antes, no entanto, traçaremos um par de considerações sobre o próximo projeto do grupo, o *Dicionário Paraolímpico*.

6 *Dicionário Paraolímpico*: desafios e perspectivas

Tendo em vista que, no ano de 2017, o grupo SemanTec deu início ao projeto de compilação de um novo instrumento lexicográfico, o *Dicionário Paraolímpico*, consideramos pertinente encerrar as considerações do presente trabalho apresentando para o público os primeiros passos da nova pesquisa e discutindo alguns desafios observados já na fase pré-compilatória do dicionário.

O projeto *Dicionário Paraolímpico*, que se encontra alinhado com as pesquisas anteriores do grupo e respaldado pela expertise adquirida nas experiências

passadas, buscará oferecer um novo instrumento lexicográfico a apreciadores e interessados em conhecer um pouco mais sobre os esportes paraolímpicos. Alicerçado nos mesmos fundamentos teórico-metodológicos do *Dicionário Olímpico* e do *Dicionário Field*, o *Dicionário Paraolímpico* também primará pela intersecção entre, pelo menos, três campos de estudo distintos, a saber, a teoria lexicográfica, a Linguística de *Corpus* e a Semântica de *Frames*.

Assim, seguindo a mesma linha do *Dicionário Olímpico*, presume-se que, no *Dicionário Paraolímpico*, as três áreas do conhecimento supracitadas influenciem de maneira mais específica em determinados aspectos da pesquisa. É esperado, por exemplo, que a teoria lexicográfica auxilie o grupo na sistematização das informações a serem oferecidas pela obra; que Linguística de *Corpus* ofereça, através da manipulação de conceitos e de termos do domínio, o conhecimento de fundo necessário para a caracterização dos esportes paraolímpicos e das paraolimpíadas; e que a Semântica de *Frames* forneça os princípios norteadores que orientarão o desenvolvimento da estrutura do dicionário. No que toca essa última associação (Semântica de *Frames* e Lexicografia), é importante ressaltar que a noção de estrutura de conhecimento engendrado pela Semântica de *Frames* assumirá um papel central na elaboração do *Dicionário Paraolímpico*. Por isso, nas páginas que seguem, explanaremos mais detalhadamente os reflexos dessa associação.

Em primeiro lugar, é necessário ressaltar que a adoção da Semântica de *Frames* como modelo norteador da obra lexicográfica ora proposta se repercutirá de duas maneiras distintas no produto final. A primeira influência, discutida exaustivamente nas seções anteriores do presente trabalho, está relacionada à forma de apresentação da obra e na maneira (diferenciada) que um dicionário baseado em *frames* pode apresentar informações ao consulente. Assim, conforme já demonstrado na exposição do *Dicionário Olímpico*, o *Dicionário Paraolímpico* também contará com informações consideradas atípicas em obras lexicográficas, tais como ilustrações de mapas conceituais, organização das informações sob a forma de cenários e listagem das unidades lexicais (ULs) características de cada esporte. Tais aspectos consistem em uma influência direta da Semântica de *Frames*, que, nesses casos, assume um caráter organizacional na obra.

Outra influência, que assume um aspecto mais conceitual, está relacionada à transposição do conhecimento sobre o *frame* paraolímpico para as páginas do dicionário, de modo a oferecer aos consulentes da obra a complexa gama de informações que permeia essa estrutura de conhecimento. É sobre esse segundo ponto que traçaremos um par de considerações a partir de agora.

Conforme apresentado anteriormente, a organização do *Dicionário Olímpico*, que será replicada no *Dicionário Paraolímpico*, parte da caracterização de cada esporte olímpico como uma estrutura de conhecimento, ou seja, um *frame* (cf. seção 3). Juntos, esses *frames* relacionam-se uns aos outros e compartilham informações do domínio dos esportes olímpicos, totalizando uma estrutura

maior e mais complexa de conhecimento, que pode ser concebida como o *frame* olímpico. O *Dicionário Olímpico*, em sua totalidade, consiste em uma tentativa de descrever o *frame* Olímpico de maneira acessível ao público e, ao mesmo tempo, fiel às normas e aos princípios estipulados pelos comitês olímpicos. Da mesma forma, no *Dicionário Paraolímpico*, a totalização das informações de cada esporte paraolímpico organizado e relacionado com os demais esportes resultará no que poderemos chamar de descrição do *frame* Paraolímpico.

As atividades de identificação e descrição desse novo *frame* representam, para o grupo idealizador do dicionário, um desafio ainda maior que o enfrentado durante a elaboração do *Dicionário Olímpico*, haja vista algumas peculiaridades do domínio paraolímpico. Em pesquisas iniciais direcionadas à familiarização do grupo com conceitos basilares do âmbito paraolímpico, foi possível observar que estávamos diante de uma estrutura de conhecimento diferenciada e que provavelmente ativa conhecimentos enciclopédicos extras, oriundos de outras áreas que não apenas o esporte. Assim, a primeira importante conclusão do grupo é que tais conhecimentos devem ser incorporados às informações do dicionário se quisermos produzir uma obra que reflita de fato o cenário paraolímpico.

Diante de tais constatações, a primeira importante dúvida repousa no modo como esses conhecimentos se farão presentes nos componentes canônicos e/ou segmentos informativos do dicionário. Em outras palavras, é necessário que pensemos em estratégias de inserção de tais informações, de modo que apareçam de maneira calculada e harmônica na obra lexicográfica, refletindo fielmente aspectos sobre os atletas, os jogos, as normas e os preceitos do domínio das paraolimpíadas. Obviamente, muitos desafios a esse respeito surgirão concomitantemente à compilação do dicionário, o que torna impossível uma previsão precisa sobre como lidaremos com cada uma das adversidades que parecem se apresentar já na fase pré-compilatória. Alguns prognósticos, no entanto, já podem ser feitos nessa primeira fase da pesquisa, e serão, alguns deles, apresentados a partir de agora. Para tanto, dividiremos as informações em conformidade com os componentes canônicos das obras lexicográficas, quais sejam: textos externos, macroestrutura, medioestrutura e microestrutura (HAUSMANN; WIEGAND, 1989). Para os propósitos da presente discussão, nos restringiremos aos âmbitos macro e microestrutural.

A macroestrutura, que pode ser definida, em termos gerais, como a lista ordenada das entradas ou a progressão vertical da obra lexicográfica, constitui o elemento central do dicionário, que torna possível a localização da informação tanto pelo compilador como pelo usuário (HARTMANN; JAMES, 2001, s.v. *macrostructure*). No caso do *Dicionário Paraolímpico*, esse segmento canônico aparecerá em, pelo menos, três alocações distintas da obra, a saber, a lista das entradas dos esportes paraolímpicos, a lista de unidades lexicais caracterizadoras de cada esporte e a lista dos cenários também caracterizadores de cada esporte.

Em comparação com o *Dicionário Olímpico*, é esperado que o *Dicionário Paraolímpico* apresente um aumento na densidade macroestrutural no que diz respeito à lista de unidades lexicais e à lista de cenários dos esportes. Isso ocorrerá devido à provável inserção de unidades lexicais que perpassam o âmbito dos esportes, cujos termos e conceitos parecem muito frequentes na literatura da área. Uma rápida análise de anais referentes a congressos paradesportivos, por exemplo, já revela uma constante intersecção entre conhecimento sobre esportes e conhecimentos sobre outras áreas. Uma das consequências diretas desse fenômeno é que o domínio paraolímpico acaba por absorver palavras e conceitos não recorrentes no cenário olímpico, a exemplo de termos como órtese/prótese (MOREIRA et al., 2012), deficiência (OLIVEIRA et al., 2011), inclusão social/ressocialização (BRANCATTI et al., 2011), esporte adaptado (CARMONA et al., 2014), dentre outros. Nesse sentido, é imprescindível observar que tais intersecções exigirão dos compiladores do dicionário a manipulação de novos conhecimentos, que se refletirão diretamente nas informações oferecidas pela obra. Por essa razão, no caso restrito da macroestrutura, prevemos a necessidade de lematização de muitas dessas palavras e expressões, seja na lista de unidades lexicais, seja na lista de cenários.

Ainda no âmbito da macroestrutura, é também importante pontuar que, como bem coloca Bugueño Miranda (2007), um dicionário deve ser sempre formulado em consonância com o seu público-alvo e com seus objetivos. Dessa forma, é necessário que o lexicógrafo projete a macroestrutura da obra lexicográfica de acordo com as necessidades do consulente. A partir desse parâmetro, tem-se o insumo inicial para delinear algumas características sobre a densidade da macroestrutura do *Dicionário Paraolímpico*. Sendo o *Dicionário Paraolímpico* uma obra que atenderá um público difuso, ou seja, uma obra que não contempla um público específico de consulentes, sendo-nos, por isso, impossível traçar o perfil de seus usuários, é importante estipular, desde já, que a macroestrutura do *Dicionário Paraolímpico* acolherá uma proposta exaustiva. A macroestrutura deverá estar em consonância com o objetivo da obra, que é fornecer ao público geral (leigo ou especialista em esportes paraolímpicos) um instrumento que reflita em linguagem clara o conhecimento subjacente ao *frame* paraolímpico. Nesse sentido, é esperado que a macroestrutura da obra seja densa e detalhada, de modo a oferecer ao consulente as palavras caracterizadoras dos esportes paraolímpicos.

O outro componente canônico que nos dispomos a discutir, a microestrutura, é definido por Hartmann e James (2001, *s.v. microstructure*) como "o desenho interno de uma unidade de referência" e por Hausmann e Wiegand (1989, p.328) como "a estrutura de informações dentro do artigo lexicográfico". A microestrutura corresponde, portanto, às informações circunscritas ao verbete do dicionário. As informações microestruturais podem ser divididas em dois grupos, que recebem o nome de *comentário de forma* e *comentário semântico*. Ao comentário de forma, correspondem informações como ortografia, gramática e pronúncia, ou seja,

informações relativas ao signo linguístico como significante, ao passo que ao comentário semântico correspondem informações como definição, etimologia e marcas de uso, ou seja, informações relativas ao signo linguístico como significado (HARTMANN; JAMES, 2001, *s.v. comment*, BUGUEÑO MIRANDA, 2009). Nessa fase de pré-compilação dos verbetes do *Dicionário Paraolímpico*, procuramos averiguar quais informações poderão sofrer influência das especificidades do domínio paraolímpico e como isso poderá se refletir nos verbetes do dicionário.

Assim, no aspecto microestrutural, é esperado que o *Dicionário Paraolímpico* apresente glosas permeadas por novos tipos de conhecimento, ou seja, informações que ultrapassam o âmbito do esporte e que abarcam outras áreas de conhecimento, conforme previamente discutido nas considerações sobre a macroestrutura. Palavras como *inclusão*, *acessibilidade*, *reabilitação*, *preconceito*, *superação*, dentre outras, parecem ativar o *frame* Paraolímpico, haja vista a sua incidência na literatura da área (cf. CPB/CPI, 2011, 2012, CPB, 2014), fornecendo, por isso, conhecimentos imprescindíveis para a caracterização dos esportes. Muitas dessas palavras, portanto, não devem se restringir ao componente macroestrutural do *Dicionário Paraolímpico*, devendo também permear a microestrutura. O modo como essas palavras integrarão a microestrutura, no entanto, é uma pergunta que ainda se encontra em aberto em razão da fase inicial na qual o projeto se encontra. Nossas hipóteses são que elas apareçam com maior incidência nos textos das glosas, seguidas pelos exemplos e, talvez, também na forma de rubrica de algumas unidades lexicais ou cenários.

Outro importante aspecto a ser observado no componente microestrutural da obra diz respeito à forma de apresentação e de descrição das informações no dicionário, que devem estar em consonância com as exigências da comunidade das pessoas com deficiência, respeitando a dignidade e primando pelo respeito aos atletas. Na fase pré-exploratória do domínio paraolímpico, já foi possível observar a existência de termos e palavras que, embora recorrentes na linguagem do dia a dia, imprimem cargas negativas à comunidade das pessoas com deficiência, devendo, por isso, ser evitadas e corrigidas quando possível. É o caso, por exemplo, da palavra *deficiente*, que, apesar de aparecer com frequência na linguagem falada e escrita, é considerada pejorativa em alguns manuais sobre linguagem inclusiva, sendo aconselhável a sua substituição por *pessoa com deficiência*, termo utilizado pelas Organizações das Nações Unidas (ONU). O mesmo vale para os termos *pessoas portadoras de deficiência* e *pessoas com necessidades especiais*, e, em um grau ainda mais alto na escala pejorativa, *aleijado*, *defeituoso*, *incapacitado* e *inválido*[5].

[5] Recomendações retiradas do manual de comunicação da Secom, disponível na página oficial do senado federal, no endereço <https://www12.senado.leg.br/manualdecomunicacao/redacao-e-estilo/estilo/linguagem-inclusiva>, Acesso em 02 out. 2017.

Além dos aspectos ressaltados até agora, cabe também salientar que as dificuldades e limitações enfrentadas nos projetos passados serviram de base para reflexões sobre a elaboração de heurísticas que auxiliassem a compilação de obras lexicográficas futuras. Assim, uma das grandes expectativas para a compilação do *Dicionário Paraolímpico* é que o grupo consiga lidar com algumas dificuldades intrínsecas ao fazer lexicográfico de maneira mais estratégica e menos intuitiva[6], aperfeiçoando cada vez mais as obras que disponibiliza ao grande público e contribuindo para uma interseção cada vez mais harmônica entre a Lexicografia, a Linguística de *Corpus* e a Semântica de *Frames*.

7 Considerações finais

As considerações apresentadas no presente trabalho originaram-se da necessidade de se relatar a experiência de compilação *Dicionário Olímpico*. Pretendemos, através dessa exposição, compartilhar com o meio acadêmico um pouco do conhecimento construído a partir dos desafios plantados pela elaboração do dicionário e suscitar alguns debates importantes no âmbito da teoria e da prática lexicográfica.

No que concerne o seu viés teórico, a presente exposição procurou demonstrar de que maneira algumas vertentes teóricas podem ser úteis na compilação de uma obra lexicográfica, com especial atenção à Semântica de *Frames*, à Linguística de *Corpus* e à Lexicografia teórica (também chamada de Metalexicografia). Em seu aspecto prático, o trabalho buscou relatar algumas adversidades enfrentadas pelo grupo durante a compilação da obra, bem como as estratégias utilizadas para contornar alguns desses obstáculos. O objetivo maior da aproximação dessas duas esferas é que se estabeleçam cada vez mais pontos de contato entre a teoria e a prática no âmbito da Lexicografia, de modo a se buscar na literatura de diferentes áreas alicerces que permitam a edificação de heurísticas visando à prática lexicográfica.

Ao término da presente discussão, é importante observar também que, ainda que inúmeras considerações tenham sido levantadas, explorando os mais diversos aspectos do *Dicionário Olímpico*, de maneira alguma tivemos a pretensão de oferecer uma análise completa e exaustiva do processo de compilação do *Dicionário Olímpico*, haja vista os muitos outros tópicos que acabaram ficando de fora da discussão. Além disso, em relação aos tópicos que foram abordados, é também importante ressaltar que, de modo algum, representam discussões que possam se dar por encerradas juntamente com essas últimas páginas. Nesse sentido, as experiências, as limitações, as soluções e os desafios ora relatados representam apenas

[6] Nessa nova etapa, procuraremos, por exemplo, elaborar um *template* para a redação das glosas, que consiga sumarizar as informações essenciais que esse segmento informativo deve apresentar na explanação do esporte, além de fornecer um modelo sintático que o lexicógrafo possa seguir durante a redação das glosas.

os primeiros passos de um longo trajeto em direção a uma prática lexicográfica mais orientada e que busca sempre se aprimorar.

Referências

ALUÍSIO, S. M.; ALMEIDA, G. M. de B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. *Calidoscópio*, São Leopoldo, v. 4, n. 3, set./dez. 2006. Disponível em: <http://revistas.unisinos.br/index.php/calidoscopio/article/view/6002/3178>. Acesso em: 06 out. 2017.

ATKINS, B. T. S.; RUNDELL, M. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press, 2008.

ANTHONY, L. *AntConc* (Version 3.4.3) [Computer Software]. Tokyo: Waseda University, 2014. Disponível em: <http://www.laurenceanthony.net>. Acesso em: 07 out. 2017.

BUGUEÑO MIRANDA, F. V. O que é macroestrutura no dicionário de língua? In: ALVES, I. M. A.; ISQUERDO, A. N. (Org.). *As ciências do Léxico*: Lexicologia, Lexicografia e terminologia. Campo Grande: Humanitas, 2007, p. 261-272.

_____. Sobre a microestrutura em dicionários semasiológicos do alemão. *Contigentia*, Porto Alegre, v. 4, p. 60-72, 2009.

BRANCATTI, P. R. et al. Basquetebol sobre rodas de Presidente Prudente: da

iniciação, das conquistas e das vitórias sociais. In: CONGRESSO PARALÍMPICO

BRASILEIRO, 2., e CONGRESSO PARADESPORTIVO INTERNACIONAL, 1., 2011,

Uberlândia. *Anais*... Uberlândia: UFU, 2011, p. 80-81.

CARMONA, E. K. et al. Cenários da produção do conhecimento sobre o esporte adaptado no Brasil. In: CONGRESSO PARADESPORTIVO INTERNACIONAL, 4., 2014, Florianópolis. *Anais*... Florianópolis: UFSC, 2014, p. 46-50.

CHISHMAN, R. L. de O. et al. *Dicionário Olímpico*. São Leopoldo: Unisinos, 2016. Disponível em: <http://www.dicionarioolimpico.com.br>. Acesso em: 15 set. 2017.

_____. Field – Dicionário de Expressões do Futebol: um recurso lexicográfico baseado no aporte teórico-metodológico da Semântica de Frames. *Signo*, Santa Cruz do Sul, v. 39, n. 67, p. 25-35, 2014. Disponível em: <https://online.unisc.br/seer/index.php/signo/article/view/5128/3819>. Acesso em: 09 out. 2017.

_____. The relevance of the Sketch Engine software to build Field – Football Expressions Dictionary. *RELIN*, Belo Horizonte, v. 23, Edição Especial, p. 769-796, 2015. Disponível em: <http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/8918/8796>. Acesso em: 09 out. 2017.

CPB/CPI 2011. CONGRESSO PARALÍMPICO BRASILEIRO, 2., e CONGRESSO PARADESPORTIVO INTERNACIONAL, 1., 2011, Uberlândia. *Anais*... Uberlândia: UFU, 2011.

CPB/CPI 2012. CONGRESSO PARALÍMPICO BRASILEIRO, 3., e CONGRESSO PARADESPORTIVO INTERNACIONAL, 1., 2012, Natal. *Anais*... Natal: UFRN, 2012.

CPI 2014. CONGRESSO PARADESPORTIVO INTERNACIONAL, 4., 2014, Florianópolis. *Anais*... Florianópolis: UFSC, 2014, p. 46-50.

CRUZ, J. C. *Uso das preposições* a *e* para *em espanhol*: análise baseada em *corpus* de aprendizes de espanhol como língua estrangeira. 2017. 157 f. Dissertação (mestrado em Linguística e Língua Portuguesa). Faculdade de Ciências e Letras, Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP), Araraquara, 2017.

DHUNNA, M.; DIXIT, J. B. *Information Technology in Business Management*. New Delhi: University Science Press, 2010.

EVANS, V.; GREEN, M. *Cognitive Linguistics:* an introduction. Edimburgo: Edinburgh University Press, 2006.

FILLMORE, C. J. Frame semantics. In: THE LINGUISTIC SOCIETY OF KOREA (Ed.). *Linguistics in the Morning Calm*. Seoul: Hanshin, 1982.

FILLMORE, C. J.; BAKER, C. A frames approach to semantic analysis. In: HEINE, B.; NARROG, H. (Ed.). *The Oxford Handbook of Linguistic Analysis*. New York: Oxford University Press, 2010, p. 313-339.

FILLMORE, C. J.; JOHNSON, C.; PETRUCK, M. R. L. Background to *FrameNet*. *International Journal of Lexicography*, Oxford, v. 16, n. 3, p. 235-250, 2003.

FRAMENET. Berkeley, [2013]. Disponível em <https://framenet.icsi.berkeley.edu/>. Acesso em: 15 out. 2017.

GATTO, M. *The Web as a* Corpus: Theory and Practice. London: Bloomsbury Academic, 2014.

HARTMANN, R. R. K.; JAMES, G. *Dictionary of lexicography*. London: Routledge, 2001.

HAUSMANN, F. J.; WIEGAND, H. E. Component parts and structures of general monolingual dictionaries: A survey. In: HAUSMANN, F. J. et al. (Hrsgn.). *Wörterbücher, dictionaries*, dictionnaires. Ein internationales Handbuch zur Lexikographie. Band 1. Berlin/New York: Walter de Gruyter, 1989, p. 328-360.

KILGARRIFF, A. et al. The Sketch Engine. Lorient: Euralex, 2004. Disponível em: <http://www.sketchengine.co.uk/>. Acesso em: 09 out. 2017.

KRIEGER, M. da G.; FINATTO, M. J. B. *Introdução à Terminologia*. Teoria & Prática. São Paulo: Contexto, 2004.

LAKOFF, G. *Woman, Fire and Dangerous Things* – What Categories Reveal about the Mind. Chicago: University of Chicago Press, 1987.

MOREIRA, L. de M. et al. Legislação referente ao uso de órteses na modalidade paralímpica de bocha. In: CONGRESSO PARALÍMPICO BRASILEIRO, 3., e CONGRESSO PARADESPORTIVO INTERNACIONAL, 1., 2012, Natal. *Anais*... Natal: UFRN, 2012, p. 290-291.

MÜLLER, C. *Princípios metodológicos para a construção de uma ontologia baseada na Semântica de Frames*. 2015. 173 p. Tese (doutorado em Linguística Aplicada). Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, 2015.

OLIVEIRA, R. B. de; CANDELORI, M. H.; BERTONI, S. A inserção da pessoa com deficiência no esporte de alto rendimento: revisando a literatura. In: CONGRESSO PARALÍMPICO BRASILEIRO, 2., e CONGRESSO PARADESPORTIVO INTERNACIONAL, 1., 2011, Uberlândia. *Anais*... Uberlândia: UFU, 2011, p. 42-43.

PETRUCK, M. R. L. *Frame semantics*. Berkeley: University of California, 1996.

SANTOS, A. N.; CHISHMAN, R. L. de O. O papel da Semântica de Frames na construção de um recurso dicionarístico: a organização lexicográfica do Field – Dicionário de Expressões do Futebol. *Revista da ABRALIN*, [S.l.], v. 14, n. 3, p. 433-468, 2015.

SANTOS, A. N. *Direito, aborto e anencefalia no Brasil:* uma análise semântico-cognitiva do processo da ADPF-54. 2016. 161 f. Dissertação (mestrado em Linguística Aplicada). Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, 2016. Disponível em: <http://www.repositorio.jesuita.org.br/handle/UNISINOS/5203>. Acesso em: 09 out. 2017.

SCHMIDT, T. The Kicktionary – A multilingual lexical resource of football language. In: BOAS, H. (Ed.) *Multilingual FrameNets in computational lexicography*: Methods and applications. Berlin: Mouton de Gruyter, 2009, p. 102-132.

SVENSÉN, B. *A Handbook of Lexicography*. The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press, 2009.

Colocações especializadas na área do Direito Comercial Internacional e proposta de glossário trilíngue[1]

Specialized collocations from the field of international trade law and a proposal of a trilingual glossary

Jean Michel Pimentel Rocha
Adriane Orenha-Ottaiano

[1] O estudo ora apresentado sintetiza as discussões desenvolvidas na dissertação de mestrado intitulada "*Fraseologia jurídico-comercial e proposta de um glossário de colocações especializadas trilíngue baseado em corpus*", defendida no Programa de Pós-Graduação em Estudos Linguísticos, da UNESP, Câmpus de São José do Rio Preto.
A íntegra da pesquisa encontra-se disponível no repositório institucional da UNESP, no seguinte endereço: https://repositorio.unesp.br/handle/11449/149766].

Jean Michel Pimentel Rocha – Doutorando do Programa de Pós-Graduação em Estudos Linguísticos, do Instituto de Biociências, Letras e Ciências Exatas, da UNESP, Câmpus de São José do Rio Preto, mestre em Estudos Linguísticos pela UNESP – jeanpimenttel@gmail.com.
Adriane Orenha-Ottaiano – Professora assistente doutora do Departamento de Letras Modernas, do Instituto de Biociências, Letras e Ciências Exatas, da UNESP, Câmpus de São José do Rio Preto, doutora em Estudos Linguísticos pela UNESP – adriane@ibilce.unesp.br.

Resumo: Neste trabalho, com base no referencial teórico da Linguística de *Corpus* e da Fraseologia, apresentamos e discutimos os resultados de um estudo teórico-metodológico acerca dos procedimentos necessários ao levantamento e à análise sintático-morfológica, léxico-semântica e tradutológica das colocações especializadas extraídas do *corpus* paralelo, em inglês e em espanhol, constituído pelos anuários (1968-2010) da UNCITRAL (Comissão das Nações Unidas para o Direito do Comércio Internacional); e de dois *corpora* comparáveis em português: um compilado pela ferramenta BootCat Front End (ZANCHETTA; BARONI; BERNARDINI, 2011) e outro coletado a partir de textos *on-line* da área do Direito Comercial Internacional. A partir desse estudo, determinamos as colocações funcionalmente equivalentes (TOGNINI-BONELLI; MANCA, 2004) em português e elaboramos uma proposta de glossário trilíngue nas direções tradutórias inglês→espanhol e espanhol→português.

Palavras-chave: Linguística de *Corpus*. Fraseologia. Colocações especializadas. Glossário trilíngue.

Abstract: In this paper, based on the theoretical background of *Corpus* Linguistics and Phraseology, we present and discuss the results of a theoretical and methodological study on the necessary procedures for the extraction, syntactic-morphological, lexical-semantic and translational analysis of specialized collocations from a parallel *corpus* in English and Spanish, consisting of the UNCITRAL (United Nations Commission on International Trade Law) yearbooks (1968 – 2010); and from two comparable *corpora* in Portuguese: one of them compiled using the BootCat Front End tools (ZANCHETTA; BARONI; BERNARDINI, 2011) and the other collected from on-line texts from the field of International Trade Law. On the basis of that, we determined the functionally equivalents collocations (TOGNINI-BONELLI; MANCA, 2004) in Portuguese and elaborated a proposal for a trilingual glossary in the translation directions English→Spanish and Spanish→Portuguese.

Keywords: *Corpus* Linguistics. Phraseology. Specialized collocations. Trilingual glossary.

1 Introdução

A Comissão das Nações Unidas para o Direito Comercial Internacional – UNCITRAL (*The United Nations Commission on International Trade Law*) é o principal órgão jurídico das Nações Unidas no âmbito do Direito Comercial Internacional. Objetiva a modernização e harmonização das regras para as transações comerciais internacionais e, para tanto, lança mão de textos em diferentes categorias, tais como textos legislativos (convenções, leis-modelo, guias legislativos), textos contratuais (regras e cláusulas, como as de Arbitragem) e explicativos (recomendações e guias legais), os quais oferecem soluções legais para países de diferentes tradições jurídicas e estágios de desenvolvimento econômico (UNCITRAL, 2013).

O Brasil, em sua prática comercial, norteia-se por muitos dos textos propostos pela UNCITRAL – a exemplo da *Convenção de Viena para a Compra e Venda Internacional de Mercadorias e da Lei de Arbitragem*. No entanto, a documentação legal regulada e produzida por esse órgão é publicada em todas as línguas oficiais

das Nações Unidas (árabe, chinês, inglês, francês, russo e espanhol), e como pode ser notado, o português não é uma das línguas oficiais. Nesse contexto, a redação de documentos em língua portuguesa pode se tornar problemática, já que as chances de equívocos na tradução podem aumentar consideravelmente, acarretando riscos e custos nas transações comerciais, além de problemas legislativos.

Atentando-se, então, para a inexistência dessa documentação em língua portuguesa, e diante da necessidade de padronização das estruturas fraseológicas utilizadas em transações comerciais internacionais brasileiras, desenvolvemos um estudo acerca dos aspectos teórico-metodológicos necessários para extrair e analisar as colocações especializadas mais frequentes e comumente empregadas na área do Direito Comercial Internacional, baseado em *corpora*, nas línguas inglesa, espanhola e portuguesa. A partir dele, elaboramos uma proposta de glossário de colocações especializadas trilíngue na área de Direito Comercial Internacional, nas direções tradutórias inglês→espanhol e espanhol→português, que atenda às necessidades do tradutor, aprendizes de tradução, bem como de profissionais da área do Direito.

2 Linguística de *Corpus* e Fraseologia

A pesquisa que ora apresentamos fundamenta-se no referencial teórico da Linguística de *Corpus* (SINCLAIR, 1996; BIBER; CONRAD; REPPEN, 1998; HUNSTON; FRANCIS, 2000; TOGNINI-BONELLI, 2001; BERBER SARDINHA, 2004) em sua interface com a Fraseologia (ZULUAGA, 1980; CORPAS PASTOR, 1996; RUIZ GURILLO, 1997; GRANGER; PAQUOT, 2008; ORENHA-OTTAIANO, 2004, 2009, 2012, 2015).

A exemplo de Berber Sardinha (2004), concebemos a Linguística de *Corpus* (LC) como uma abordagem empírica no estudo da língua, a qual é vista como um sistema probabilístico, cujos fenômenos lexicais, morfossintáticos, fonológicos etc. não são aleatórios, mas regulares. Assim, qualquer tipo de aleatoriedade nos estudos da língua é, como defende Sinclair (1996), uma possibilidade remota, visto que nela tudo é altamente determinado, planejado e organizado.

Atentar-se para a regularidade dos fenômenos linguísticos é uma tarefa possível graças à utilização de um *corpus*. No contexto desta pesquisa, o emprego de *corpora* mostra-se fundamental, o que nos leva à adoção dos procedimentos teórico-metodológicos da LC, os quais nos permitem a identificação e a análise de padrões associativos, nem sempre possíveis pela introspecção (BIBER; CONRAD; REPPEN, 1998).

Por padrões de determinado vocábulo, entendemos, pautando-nos em Hunston e Francis (2000), as estruturas que regularmente acompanham certa unidade lexical e que contribuem para seu sentido. Para identificar padrões em *corpora*, é preciso, segundo esses autores, uma conexão entre uma teoria, um

método e uma técnica. Por essa razão, similarmente à Hunston e Francis (2000) e Tognini-Bonelli (2001), posicionamo-nos em favor da adoção de uma teoria que privilegie o significado em contexto e não por palavras isoladas, desprovidas de contexto. Essa posição requer, ainda, a escolha de uma metodologia em que a repetição e a coocorrência sejam prioritárias, como a possibilitada pelo uso das linhas de concordância (KWIC – *Keywords in Context*).

Frente a tais questões, estamos fundamentados na LC em sua estreita relação com a Fraseologia, especificamente nos estudos fraseológicos baseados na frequência (GRANGER; PAQUOT, 2008), os quais decorrem dos trabalhos lexicográficos levados a cabo por Sinclair (1987), e que adotam, com auxílio de programas de análise lexical, uma abordagem *bottom-up* na extração e identificação de coocorrências lexicais, sem que haja, necessariamente, uma categorização linguística previamente estabelecida. Esse novo viés, possibilitado pela LC para o tratamento das associações lexicais, revolucionou os estudos da Fraseologia, permitindo a identificação e a descrição de um grande número de unidades fraseológicas.

A rede de unidades fraseológicas existente é bastante extensa, abarcando, conforme Granger e Paquot (2008), as unidades que possuem função pragmática ou comunicativa (provérbios, *slogans*, fórmulas de rotina, fórmulas conversacionais); as que apresentam função textual (conjunções e preposições complexas, conectivos); e as que apresentam função referencial, a exemplo das expressões idiomáticas, dos verbos frasais, dos binômios e das colocações. Diante dessa gama de fraselogismos, elegemos como objeto de estudo as colocações, especificamente as colocações especializadas e as colocações especializadas estendidas.

Para esta pesquisa, delimitamos as colocações como unidades fraseológicas recorrentes e convencionalizadas, estruturadas em um sintagma (substantivo + substantivo, adjetivo + substantivo, verbo + substantivo, advérbio + adjetivo etc.). Dessa forma, em consonância com Corpas Pastor (1996, p. 53), partindo de uma concepção coseriana nos estudos da língua, concebemo-las como "unidades fraseológicas que, do ponto de vista do sistema da língua, são sintagmas completamente livres, gerados a partir de regras, mas que, ao mesmo tempo, apresentam certo grau de restrição combinatória determinada pelo uso"[1].

Entretanto, distintamente das colocações da língua geral, comumente empregadas pelos falantes nas mais diversas situações da vida cotidiana, a partir do léxico geral que compartilham entre si, debruçamo-nos sobre as *colocações especializadas*, as quais caracterizam uma área de especialidade e podem apresentar, em sua constituição, uma unidade lexical especializada ou um termo, e, podem, ainda, estruturar-se em formações sintagmáticas mais extensas, conhecidas

[1] *"unidades fraseológicas que, desde el punto de vista del sistema de la lengua, son sintagmas completamente libres, generados a partir de reglas, pero que, al mismo tiempo, presentan cierto grado de restricción combinatoria determinada por el uso."*

como *colocações especializadas estendidas*, conforme nomenclatura proposta por Orenha-Ottaiano (2009).

A propósito de ilustração, apresentamos abaixo, exemplos dos tipos colocacionais mencionados:

- **Colocações da língua geral**: *rio caudaloso*, *ledo engano*, *dúvida cruel, mentira cabeluda*; *aventar uma hipótese*, *arquear as sobrancelhas*, *aproveitar o ensejo*; *redondamente enganado*.
- **Colocações especializadas**: *contrato de transporte*, *violação de contrato*, *contrato formal*, *contrato definitivo*, *contrato verbal*, *impetrar um mandado*, *celebrar um contrato*, *violar um contrato*.
- **Colocações especializadas estendidas**: *contrato de compra e venda de mercadorias*, *contrato de compra e venda internacional de mercadorias*, *contrato chave na mão por preço global*, *violação essencial de um contrato*.

De modo geral, a partir das características elencadas por alguns teóricos (CORPAS PASTOR, 1996; TAGNIN, 1999, 2013) e revisitadas por nós, entendemos que as colocações, sejam elas da língua geral ou especializada, em maior ou menor grau, caracterizam-se por apresentar:

i. **Frequência relativa**: a combinação precisa apresentar certa regularidade de ocorrências, evidenciando-se tratar-se de uma associação não aleatória e que apresente sentido.
ii. **Fixação no nível da norma**: são combinações linguísticas de duas ou mais palavras convencionalizadas, sancionadas pela norma, isto é, referendadas e habitualmente compartilhadas pela comunidade de falantes nas mais diversas situações comunicativas.
iii. **Contexto sociocomunicativo**: há um contexto social e comunicativo que requer a utilização da combinação.
iv. **Estruturação sintagmática**: seus constituintes, em geral, estão em relação sintagmática, com destaque para as associações *adjetivo* + *substantivo*; *susbstantivo* + *substantivo*; *substantivo* + *preposição* + *substantivo*; *verbo* + *advérbio*; *verbo* + *substantivo* etc., não necessariamente adjacentes.
v. **Composicionalidade**: seus elementos possuem uma semântica mais transparente, o que não os impedem de apresentar certo grau de metaforicidade e idiomaticidade.
vi. **Idiomaticidade relativa**: embora caracterizadas pela composicionalidade, seus elementos podem adquirir uma dimensão metafórica que as qualificam como idiomáticas.

Vale ressaltar que o tratamento por nós dado às colocações especializadas e às colocações especializadas estendidas pauta-se, em grande medida, por autores que desenvolveram trabalhos no escopo da Fraseologia da língua geral. Isso se deve ao fato de, a nosso ver, as unidades fraseológicas, tanto nesse âmbito quanto no âmbito das línguas de especialidade, compartilharem características similares. O que se observa, na verdade, é uma ampla terminologia que, muitas vezes, referenciam o mesmo fenômeno. A variação terminológica das unidades fraseológicas pode ser observada, por exemplo, em Bevilacqua (2016, no prelo), ao mapear a diversidade de nomenclaturas existentes nas pesquisas acerca da Fraseologia Especializada no Brasil. De acordo com a autora, há combinações que se aproximam mais das colocações, constituindo-se por uma base e um colocado, como as Combinatórias Léxicas Especializadas (CLEs) e as Unidades Fraseológicas Eventivas com pivô terminológico (UFEs); e há combinações que delas se distanciam estruturalmente, mas aproximam-se das colocações especializadas estendidas como as Unidades Fraseológicas Eventivas sem pivô terminológico e as Combinatórias Léxicas Especializadas Jurídicas (CLEs jurídicas). Dessa maneira, pelas questões elencadas e por considerar que a Fraseologia da língua geral fornece subsídios teóricos que dão conta das colocações especializadas e estendidas, não nos alinhamos diretamente a autores que se identificam com a chamada Fraseologia Especializada.

3 A compilação do *corpus* de estudo e a extração das colocações

Nosso *corpus* de estudo, conforme se observa na Figura 1, a seguir, está organizado em quatro *subcorpora*, a saber: (i) o *corpus* em inglês (15.724.787 *tokens*), constituído pelos anuários da UNCITRAL, nesta língua; (ii) o *corpus* em espanhol (16.006.493 *tokens*), também composto pelos anuários da UNCITRAL, traduzidos da versão em inglês; (iii) um *corpus* comparável em português (8.652.659 *tokens*), compilado a partir de documentos oficiais do governo brasileiro (decretos, leis, atos internacionais disponibilizados na *web,* além de outros documentos jurídicos e artigos da área coletados via *web;* e (iv) e outro *corpus* comparável em português, compilado pela ferramenta BootCat (8.652.659 *tokens*).

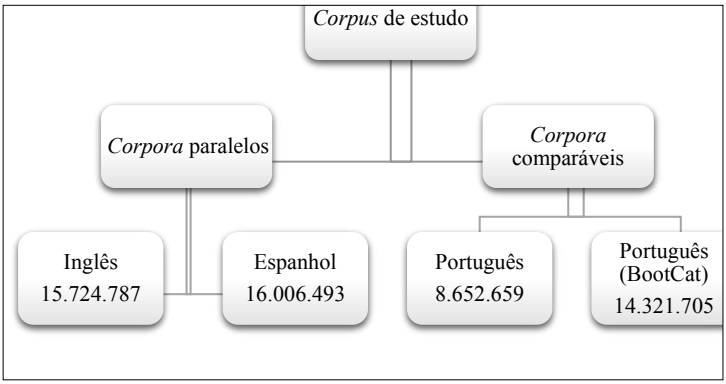


Figura 1 – Organização do *corpus* de estudo
Fonte: Elaborado pelos autores

Nosso *corpus* de estudo apresenta as seguintes características:

i. **Língua escrita**: textos constituídos pelos anuários da UNCITRAL em inglês e espanhol; e por documentos em matéria de Direito Comercial Internacional, em língua portuguesa, coletados da *web*.
ii. **Não-etiquetados**: não possuem etiquetagem morfossintática, trabalho inviável dado o tamanho do *corpus* e o tempo de desenvolvimento da pesquisa.
iii. **Trilíngue**: textos nas línguas inglesa, espanhola e portuguesa.
iv. **Paralelo**: o *corpus* em inglês é paralelo ao espanhol, visto que o segundo é tradução do primeiro.
v. **Comparável**: o *corpus* em língua portuguesa, em relação aos *corpora* de estudo em inglês e em espanhol, pode ser comparável, dada a similaridade dos textos que o compõe; os textos que formam o *corpus* em português também são comparáveis entre si.
vi. **Diacrônico**: se levarmos em conta a recolha dos anuários (1968-2010), podemos afirmar que a distribuição dos textos no *corpus* paralelo se dá diacronicamente.
vii. **Especializado**: são especializados por possuírem textos comuns à área do Direito Comercial Internacional.

Escolhemos os anuários – disponíveis no *site* da UNCITRAL[2], publicados de 1968 (primeiro ano da publicação) a 2010 (último ano disponibilizado à época da

[2] *Links* para os anuários em inglês e em espanhol, respectivamente:
<http://www.uncitral.org/uncitral/en/publications/yearbook.html>.
<http://www.uncitral.org/uncitral/es/publications/yearbook.html>.

compilação do *corpus*) – em razão de reunirem a documentação essencial regulada pela UNCITRAL, não havendo, assim, necessidade de recolher documento por documento, o que demandaria mais tempo. A escolha dos anuários em espanhol também foi estratégica. Cientes da inexistência dos anuários em língua portuguesa, antecipamos possíveis problemas na identificação de equivalentes tradutórios, haja vista que o *corpus* em português poderia não alcançar a representatividade suficiente para a extração de colocações especializadas equivalentes. Assim sendo, dada a proximidade linguística entre o português e o espanhol – línguas românicas –, principalmente em relação aos aspectos lexicais e sintático-morfológicos, hipotetizamos que o *corpus* em língua espanhola pudesse oferecer pistas para a identificação das colocações funcionalmente equivalentes em nosso idioma. Uma vez concluída a recolha, renomeamos os textos conforme o ano de publicação, armazenando-os, em formato *.pdf e em formato *.txt, em distintas pastas.

Em relação à compilação do *corpus* comparável, nossa inquietação estava em torno da seguinte questão: como compilar um *corpus* que fosse representativo de determinado conjunto de textos, em sua grande maioria, inexistentes em nossa língua? Para alcançar esse propósito, valemo-nos de dois procedimentos. Optamos pela compilação automática de um *corpus* comparável, empregando o conjunto de ferramentas BootCaT – Bootstrap *Corpora* and Terms from the web (BARONI; BERNARDINI, 2004), especificamente, a versão 0.71, BootCaT Front End[3] (ZANCHETTA; BARONI; BERNARDINI, 2011).

O conjunto de ferramentas BootCaT Front End (ZANCHETTA; BARONI; BERNARDINI, 2011), em termos gerais, permite a compilação automática de *corpora* via *web*, a partir da combinação de uma lista de sementes (*seeds*), ou palavras-chave, de determinada área de especialidade ou do léxico da língua geral. Utilizamos como sementes palavras recorrentes nos textos da UNCITRAL, como: *direito do comércio internacional*, *regras de arbitragem*, *transporte de mercadorias*, *convenções*, *leis-modelo* etc. Listadas e combinadas as sementes e ajustados alguns parâmetros de pesquisa, o programa faz uma varredura em *sites,* por meio da utilização de motores de busca como Bing e Google, à procura dos textos em que elas ocorrem. Encerrada a varredura, as páginas selecionadas são baixadas e os códigos *HTML* são removidos. Após tais procedimentos, tem-se um *corpus* compilado, salvo em um arquivo em formato *.txt.

Para a compilação do segundo *corpus* comparável, realizamos um procedimento similar ao do BootCat: recolhemos textos que versavam sobre questões diversas em matéria de Direito Comercial Internacional, especialmente relacionadas à UNCITRAL, no entanto, fizemos uma busca manual.

Ambos os *corpora* comparáveis apresentaram resultados bastante satisfatórios, evidenciados pelas listas de palavras-chave (*Keywords*), com muitas

[3] Disponível em: <http://bootcat.sslmit.unibo.it/?section=download>.

correspondências entre as palavras dos *corpora*. A própria busca pelas colocações funcionalmente equivalentes também atestou a qualidade dos *corpora*, já que a grande maioria dos exemplos do glossário foi retirada deles.

O levantamento das colocações foi possível pela utilização das ferramentas básicas – Concord, Keywords e Wordlist – do programa Word Smith Tools 6.0 (SCOTT, 2012). Valemo-nos, para tanto, do método N-Gram/cluster analysis, o qual possibilita a extração de sequências de duas ou mais palavras, conhecidas na literatura como *n-grams* (*bigrams*, *trigrams*), *clusters, bundles* etc.; e do método *cooccurrence analysis*, em que prevalece a coocorrência de uma palavra com outra, estatisticamente determinada (GRANGER; PAQUOT, 2008). Destacamos, nesse processo, o uso dos *corpora* de referência. Com a finalidade de criarmos a listagem de palavras-chave do *corpus* de estudo, utilizamos três *corpora* de referência, a saber: o BNC (British National *Corpus*), o *Corpus* da Folha, e o CREA (*Corpus* de Referencia del Español Actual).

No que diz respeito à identificação das colocações funcionalmente equivalentes nas línguas espanholas e portuguesa, orientamo-nos pela metodologia proposta por Tognini-Bonelli e Manca (2004). Essa metodologia consiste, basicamente, em três passos: primeiramente, estuda-se o perfil léxico-gramatical do nódulo de busca e determina-se os seus colocados mais frequentes. Em seguida, uma tradução *prima-facie* é atribuída e a investigação de seus cotextos é conduzida, de modo a levantar possíveis padrões colocacionais e coligacionais. A partir da análise dos padrões levantados, uma tradução adequada, funcionalmente determinada, tendo-se em vista o contexto em que ocorrem, pode ser identificada (TOGNINI-BONELLI; MANCA, 2004).

4 Discussão dos resultados

Extraímos cerca de 200 bases candidatas a integrarem o glossário, porém, não fizemos um estudo exaustivo de cada uma delas. Elegemos, assim, como modelo de análise a ser aplicada às demais, a base *contract* – dada sua alta frequência e chavicidade no *corpus* de estudo em inglês – e as colocações que dela se desdobram. Em relação a essa base, identificamos em torno de 180 colocações, em cada uma das línguas. Classificamo-las, com base em Cowie (1978), Hausmann (1985) e Benson, Benson e Ilson (2009), conforme sua estrutura sintagmática, as quais subdividimos em três padrões colocacionais: nominais, adjetivais e verbais, exemplificados no quadro abaixo:

Quadro 1 – Quadro resumitivo dos padrões sintagmáticos para o nódulo *contract*

Inglês	Espanhol	Português
Colocações nominais		
Noun + noun *Framework contract*	*contrato/acuerdo marco*	*contrato/acordo quadro*
Noun + noun *Contract award*	*adjudicación de un contrato*	*adjudicação de um contrato*
Noun + prep. + noun *breach of a contract*	*incumplimiento/transgresión de un contrato*	*violação/descumprimento de um contrato*
Noun + prep. + noun + prep. + noun *contract of sale of goods*	*contrato de compraventa internacional de mercaderías*	*contrato de compra e venda internacional de mercadorias*
Colocações adjetivais		
Adjective + noun *commercial contract*	*contrato comercial*	*contrato comercial*
original contract	*contrato originario/de origen/ original*	*contrato original/inicial*
written contract	*contrato escrito*	*contrato escrito*
Colocações verbais		
verb + det. + noun *avoid a contract*	*resolver un contrato*	*resolver um contrato*
noun + verb *a contract contains/provides*	*un contrato contiene/estipula*	*um contrato contém/estipula*

Fonte: Elaborado pelos autores

As colocações nominais foram as mais produtivas em nossos *corpora*, distribuindo-se, como sintetizado na tabela, ora pela união de dois substantivos ora pela estruturação em um sintagma preposicionado, principalmente nas línguas latinas. As colocações adjetivais aparecem na estrutura *adj. + noun* (*contract*), na língua inglesa e também nas línguas espanhola e portuguesa, diferindo apenas em relação à ordem, visto que, na maioria das colocações adjetivais nessas línguas, o adjetivo aparece posposto ao nome. Quanto às colocações estruturadas em um sintagma verbal, tem-se que, na maior parte das ocorrências, *contract* aparece na condição de objeto, sendo raras as ocorrências em que aparece na condição de sujeito (*a contract contains* / *un contrato contiene* / *um contrato contém*).

Ao enfatizar os aspectos léxico-semânticos e tradutológicos das colocações, destacamos as variações colocacionais. Além das variações no nível morfossintático (*contract of assignment* / *assignment contract*, *contract of barter* / *barter contract*, *contract of carriage* / *carriage contract; negociación contractual* / *negociación de un*

contrato; negociação contratual / *negociação de um contrato*, estruturando-se em sintagmas distintos, as colocações variaram também no nível lexical, como se observa no quadro abaixo:

Quadro 2 – Variação das colocações no nível lexical

Inglês	Espanhol	Português
framework agreement *framework contract*	*contrato marco* *acordo marco*	*acordo quadro* *contrato quadro*
prior contract *pre-existing contract* *preliminary contract*	*contrato previo*	*contrato preliminar* *pré-contrato* *contrato prévio* *contrato predeterminado*
original contract *initial contract*	*contrato de origen* *contrato originário* *contrato original* *contrato inicial*	*contrato original* *contrato originário* *contrato inicial*
contract of carriage *transport contract*	*contrato de transporte*	*contrato de transporte*
kind of contract *type of contract*	*tipo de contato* *clase de contrato*	*tipo de contrato*
subject-matter of the contract *object of the contract* *purpose of the contract*	*objeto del contrato* *finalidad del contrato*	*objeto do/de um contrato* *finalidade do contrato*

Fonte: Elaborado pelos autores

A nosso ver, as variantes lexicais estabelecem entre si uma relação parassinonímica e compartilham, dessa maneira, significados comuns, no entanto, não podem ser consideradas perfeitamente sinônimas, já que não podemos afirmar categoricamente que uma pode substituir a outra em qualquer contexto. Além disso, adotamos a noção de equivalência funcional entre as colocações, pois o fato de estarmos lidando com uma língua de especialidade que envolve o léxico de sistemas jurídicos diferentes que se refletem, por exemplo, no significado das colocações, não nos permite afirmar que tais colocações sejam equivalentes totais. Em muitos casos, os próprios excertos dos documentos dos *corpora* apontam para essa relação. *Contract of affreightment*, por exemplo, a depender do sistema jurídico, poderia ser entendido como *"volume contract", "tonnage contract"* e *"quantity contract"*, segundo pode ser observado em excerto de um dos documentos analisados, o *yearbook* da UNCITRAL (2003b, p. 420):

> Consequently, clear definitions should be provided in the draft instrument in order to circumscribe the exact scope of any exclusion. It was pointed out that a "volume" contract, also referred to as an "ocean transportation contract" or "OTC", had

> few distinctive characteristics when compared to a **carriage** contract. Expressions such as "contract of affreightment", "volume contract", "tonnage contract" and "quantity contract", were also used and, depending on the legal system, appeared to be treated as synonymous.

Outro excerto, extraído do *yearbook* (2006, p. 895), em inglês, mostra que *contract of affreightment* pode ser entendido como *bills of landing* e *charter parties*.

> The terminology is, however, in some replies found to be problematic concerning the meaning of "volume contracts". The term "**contract of affreightment**" is in one of the replies understood to be synonymous to "volume contracts". A "contract of affreightment" is also understood to refer to bills of lading and/or to charter parties.

Diante dessas questões, não podemos incorrer no erro de afirmar que as colocações apresentam uma relação de sinonímia perfeita. Assim sendo, acreditamos que possam compartilhar semas comuns, mas uma não recobre totalmente o significado de outra, por isso, preferimos a noção de parassinonímia.

Para exemplificar de modo mais detalhado as variações nos níveis lexicais e morfológicos, bem como a relação de parassinonímia e equivalência funcional entre as colocações, tomamos por modelo a colocação *breach of a contract*.

Trata-se de uma colocação bastante recorrente no *corpus* em língua inglesa, com cerca de 822 ocorrências, apresentando variações tanto em espanhol como em português, evidenciadas, em um primeiro momento, pela discrepância de frequência nos *corpora* paralelos. Na língua espanhola, como se observa pelos exemplos, *breach of a contract* traduz-se por *incumplimiento de un contrato* (363) e *incumplimiento contratual* (8) – variações no nível morfossintático. Mas há também as variações no nível lexical, como em: *transgresión de un contrato* (66), *violación de un contrato* (51) e *ruptura de un contrato* (5), que podem ser entendidas como colocações parassinônimas:

> Damages for **breach of contract** by one party consist of a sum equal to the loss, including loss of profit, suffered by the other party as a consequence of the breach. Such damages may not exceed the loss which the party in breach foresaw or ought to have foreseen at the time of the conclusion of the contract, in the light of the facts and matters of which he then knew or ought to have known, as a possible consequence of the **breach of contract**.

> La indemnización de daños y perjuicios por el **incumplimiento del contrato** en que haya incurrido una de las partes comprenderá el valor de la pérdida sufrida y el de la ganancia dejada de obtener por la otra parte como consecuencia del incumplimiento. Esa indemnización no podrá exceder de la pérdida que la parte que haya incurrido en incumplimiento hubiera previsto o debiera haber previsto en el momento de la celebración del contrato, tomando en consideración los hechos de

que tuvo o debió haber tenido conocimiento en ese momento, como consecuencia posible del **incumplimiento del contrato.**

In this case, the buyer filed an action to enforce a contract for the sale of real property and to recover damages arising out of an alleged **breach of** that **contract** against the sellers.

En esta causa, el comprador entabló juicio contra el vendedor para forzar el cumplimiento de un contrato de compraventa de bienes raíces y solicitar indemnización por los daños y perjuicios sufridos por el supuesto **incumplimiento contractual**.

Damages for **breach of contract** by one party consist of a sum equal to the loss, including loss of profit, suffered by the other party as a consequence of the breach. Such damages cannot exceed the loss which the party in breach foresaw or ought to have foreseen at the time of the conclusion of the contract, in the light of the facts and matters which he then knew or ought to have known, as a possible consequence of the **breach of contract**.

Los daños y perjuicios causados por una **violación del contrato** cometida por una de las partes consisten en una suma igual a la pérdida, incluido el lucro cesante, sufrida por la otra parte como consecuencia de la violación. Dichos daños y perjuicios no pueden exceder de la pérdida que la parte transgresora haya previsto o debió haber previsto al tiempo de la celebración del contrato, tomando en consideración los hechos y elementos que conocía o debía haber conocido entonces, como consecuencia posible de la **transgresión del contrato**.

"**Breach of contract**" means the failure of a party to perform the contract or any performance not in conformity with the contract;

Por « **violación del contrato** » se entenderá toda inejecución de las obligaciones de una parte o cualquier cumplimiento que no fuere conforme al contrato;

Any third party to which the shipper or the consignee has assigned its rights, depending on which of the above parties suffered the loss or damage in consequence of a **breach of the contract** of carriage.

Cualquier tercera parte a la que el cargador o el consignatario haya asignado sus derechos, según cuál de las partes antes mencionadas haya sufrido la pérdida o el daño como consecuencia de uma **ruptura del contrato** de transporte.

Na língua portuguesa, presumimos que a tradução *prima facie* seria similar às traduções em espanhol. Porém, decidimos também averiguar a tradução de *breach* em dicionário especializado (MELLO, 2012). As traduções sugeridas foram *infração*, *violação*, *ruptura* ou *quebra*. Dessas, apenas a unidade lexical *infração* não encabeçou a colocação no *corpus*. Em língua portuguesa, encontramos no

corpus as seguintes colocações parassinônimas, as quais variaram no nível lexical: *descumprimento de um contrato* (34), *quebra de um contrato* (24) e *incumprimento de um contrato* (35), *ruptura de um contrato* (2); e no nível morfossintático: *violação de um contrato* (60), *violação contratual* (18), sendo que a colocação *violação de um contrato* mostrou-se mais recorrente que as demais:

> As perdas e danos decorrentes de **violação do contrato** por uma das partes consistirão no valor equivalente ao prejuízo sofrido, inclusive lucros cessantes, sofrido pela outra parte em consequência do descumprimento. Esta indenização não pode exceder à perda que a parte inadimplente tinha ou devesse ter previsto no momento da conclusão do contrato, levando em conta os fatos dos quais tinha ou devesse ter tido conhecimento naquele momento, como consequência possível do descumprimento do contrato.
>
> Se o vendedor cometeu uma violação fundamental do contrato, as disposições dos artigos 67, 68 e 69 não prejudicam o recurso aos meios de que o comprador dispõe em virtude daquela **violação contratual**.
>
> Salvo se tiver recebido a comunicação do vendedor de que não cumprirá suas obrigações no prazo fixado conforme o parágrafo anterior, o comprador não poderá exercer qualquer ação por **descumprimento do contrato**, durante o prazo suplementar. Todavia, o comprador não perderá, por este fato, o direito de exigir indenização das perdas e danos decorrentes do atraso no cumprimento do contrato.
>
> Nenhuma disposição da Convenção ou do presente Protocolo prejudica a responsabilidade de um credor no caso de **quebra de contrato** conforme a lei aplicável, na medida em que o referido contrato diga respeito a um bem aeronáutico.
>
> No Direito Brasileiro o **incumprimento do contrato** por parte de um dos contraentes (devedor) pode o contratante pontual, em vez da atitude passiva da defesa, adotar um comportamento ativo na preservação de seus direitos de fato, se o incumprimento resulta de culpa de um dos contraentes, a lei concede ao outro uma alternativa, com efeito, pode ele: a) exigir do outro contraente o cumprimento da obrigação; ou b) pedir judicialmente a resolução do contrato.

Incumprimento de um contrato (35) é mais comum no português europeu, já que é bastante recorrente em documentos da Comunidade Europeia, cerca de 25 ocorrências no *corpus*. Essa colocação não se encontra nos dicionários de apoio à pesquisa e o item lexical *incumprimento* também não está dicionarizado (HOUAISS, 2009). No entanto, como aponta o último exemplo em português, o vocábulo existe no português brasileiro, atestado pelo *corpus* comparável, além de ser reconhecido pelo Vocabulário Ortográfico da Língua Portuguesa (VOLP)[4].

[4] Disponível em: <http://www.academia.org.br/nossa-lingua/busca-no-vocabulario>.

Além dessas consultas, ao buscar pela colocação *incumprimento de um contrato*, utilizando o método de pesquisa avançada do Google, configurando a busca por páginas brasileiras em português associada à UNCITRAL, obtivemos 363 ocorrências, o que atesta o uso da colocação em nossa língua, mesmo com frequência baixa.

No que diz respeito às colocações especializadas estendidas, extraímos, na língua inglesa, a colocação especializada estendida *fundamental breach of a contract.* Nas línguas latinas, observaram-se variações no nível lexical entre as colocações. Assim, em espanhol, há as colocações especializadas estendidas *transgresión* e *incumplimiento esencial de un contrato*; e, em português, as colocações *violação fundamental* e *violação essencial de um contrato*.

Não são só as colocações especializadas estendidas que auxiliam na definição das colocações equivalentes funcionalmente determinadas, outros colocados em comum auxiliam nesse processo, a exemplo de padrões como *breach of the contract by the seller*, *by the buyer*, *by one party*, *by the assignor* e *by the contractor*; *incumplimiento por parte del cedente*, *por el vendedor*, *por el comprador*, *por las partes*, *por el deudor*, *por el porteador*; *violação/incumprimento/quebra do contrato por parte do vendedor*, *pelo comprador*, *por uma das partes*; *recursos em caso de violação*, *a parte que invoca a violação.*

Vê-se, então, que não apenas a determinação das colocações estendidas auxilia na busca por uma tradução nas línguas latinas, mas os itens lexicais que ocorrem nas adjacências da colocação estudada também ajudam, o que atesta que a metodologia empregada é viável para identificá-las, principalmente quando a frequência das colocações é alta.

Além dos aspectos acima observados, vale ressaltar que encontramos, em nossos *corpora*, casos de empréstimos, tanto na língua espanhola (*contrato de factoring*) quanto na língua portuguesa (*contrato de engineering*, *contrato de leasing*, *contrato turn-key*) em pelo menos um dos constituintes da colocação. Destacamos, ainda, embora em número reduzido, os casos de variação ortográfica em português (*contrato de fretamento* /*afretamento*). Por fim, notamos que, nas colocações especializadas extraídas, além da equivalência funcional atestada contextualmente, há uma similaridade ortográfica entre os vocábulos que as constituem (*cancel a contract*, *cancelar un contrato*, *cancelar um contrato* etc.) nas línguas analisadas, a começar pela própria base – *contract*, de origem latina (HOUAISS, 2009), o que indica que tais possuem etimologia comum, sendo, portanto, palavras cognatas.

5 Delimitação da macro e microestrutura do glossário

Extraídas e identificadas as colocações, registramo-las em *fichas fraseológicas* (Figura 2), as quais contemplam um conjunto de informações linguísticas e extralinguísticas acerca de determinada base que comporá o glossário. Tais fichas foram

criadas com a ajuda do programa Microsoft Access. A partir delas, delimitamos a macro e micro estrutura do glossário.

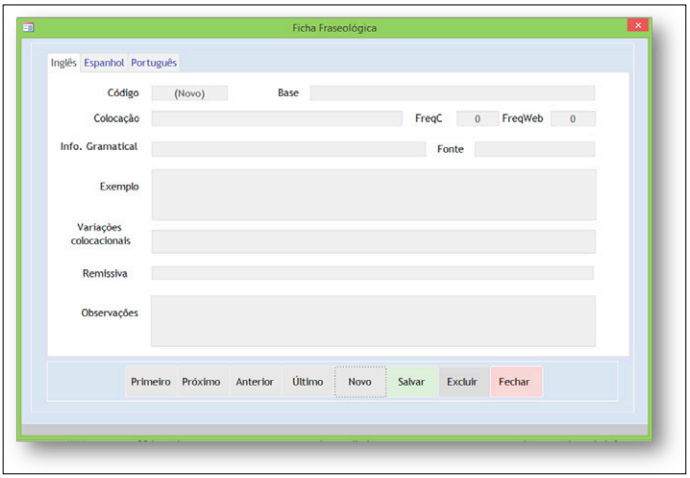


Figura 2 – Ficha fraseológica do glossário
Fonte: Elaborado pelos autores

Para cada língua, inserimos campos para:

1. **bases das colocações**: BaseIng, BasePort, BaseEsp.
2. **colocações**: ColocaçãoIng, ColocaçãoEsp, ColocaçãoPort.
3. **informações gramaticais**: InfoGramaticIng, InfoGramaticEsp, InfoGramaticPort.
4. **exemplos**: ExemploIng, ExemploEsp, ExemploPort.
5. **frequências nos *corpora* de estudo**: FreqIng, FreqEsp e FreqPort.
6. **frequências na *web***: FreqWebIng, FreqWebEsp, FreqWebPort.
7. **variações colocacionais**: VarColIng, VarColEsp, VarColEsp.
8. **fontes dos exemplos**: FonteIng, FonteEsp, FontePort.
9. **remissivas**: RemIng, RemEsp e RemPort.
10. e para **observações gerais**: ObsIng, ObsEsp, ObsPort.

Em se tratando de uma obra de cunho léxico-fraseográfico, para desenvolver a micro e a macroestrutura do nosso glossário, fundamentamo-nos em autores que se dedicaram à compilação de dicionários e glossários de colocações, guiamo-nos, principalmente, pelos trabalhos de Orenha-Ottaiano (2004, 2009, 2015) e de autores clássicos nos estudos das colocações (COWIE, 1978; HAUSMANN, 1985; BENSON; BENSON; ILSON, 2009). Além dessas referências, contamos também com o apoio da Lexicografia (BARBOSA, 1999) e, em parte, da Terminografia (BARROS, 2004).

As informações gerais do nosso glossário, as quais fazem parte de sua macroestrutura, abarcam (i) **um texto introdutório**, com informações sobre a obra (objetivos, público-alvo, procedimentos metodológicos e a descrição de sua própria organização interna); (ii) **abreviações** das classes gramaticais nas diferentes línguas trabalhadas: a exceção do substantivo em inglês, cuja grafia mantivemos (*noun*), abreviamos *article / artículo / artigo (art.)*; *adjective / adjetivo / adjetivo (adj.); substantivo (subst.); substantivo (sust.); preposition / preposición / prepeposição (prep.); e verb/verbo/verbo (verb.)*. Por fim, prevemos um (iii) índice com as entradas do glossário, em ordem alfabética, que direciona o consulente à localização do verbete, encabeçado por uma base.

Partindo das informações constantes na *ficha fraseológica*, organizamos a microestrutura do glossário, conforme dados do verbete representados na figura a seguir:

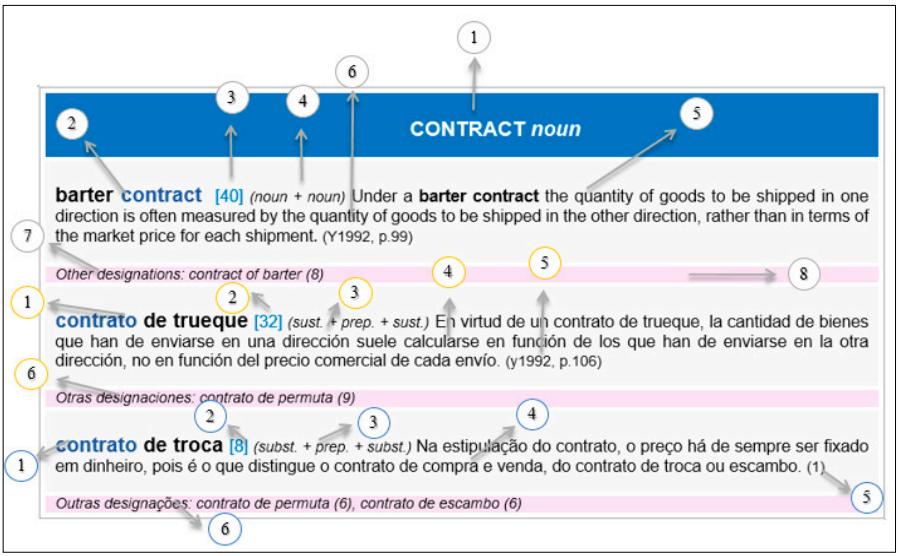


Figura 3 – Elementos da microestrutura
Fonte: autoria própria

A partir da língua inglesa, as informações da microestrutura estão dispostas pela:

1. Entrada: base da colocação que encabeça o verbete na língua de partida (LP), seguida por sua categoria gramatical (*noun*, *adjective*, *verb*). Os vocábulos da entrada estão grafados em letra maiúscula e dispostos na forma singular/masculino, quando substantivos e adjetivos; e na forma infinitiva, quando verbos.

2. Colocação especializada na LP: grafada em letra minúscula e em negrito, com a base na cor azul. A colocação está disposta na forma singular, a não ser

que seja um plural lexicalizado ou que apresente ocorrências muito altas no plural. No caso das colocações especializadas estendidas, inserem-se no mesmo quadro das colocações especializadas que as origina, funcionando como um hipônimo da colocação. Ressaltamos que só compõem o glossário colocações estendidas mais restritas, e não as que apresentam itens lexicais que frequentemente coocorrem com outras colocações especializadas.

3. Frequência da colocação no *corpus* de estudo: destacada entre colchetes, na cor azul.

4. Estrutura sintagmática da colocação na LP: escrita em itálico e entre parênteses. Apenas no caso das colocações especializadas estendidas, não apresentamos essa informação.

5. Exemplo na LP: retirado do *corpus* de estudo em inglês. No corpo do texto, a colocação está destacada em negrito.

6. Fonte do exemplo na LP: menção ao ano e à página do anuário de onde o exemplo fora retirado.

7. *Other designations*: campo destinado à inserção das colocações parassinônimas, que possuem subentradas próprias apenas no caso de variação lexical. As variações no nível morfossintático estão inseridas nesse campo, seguidas da sua frequência no *corpus*.

8. Remissiva *see*: considerando que o colocado em uma dada colocação poderá ser a base, ou seja, a entrada de outra colocação, a remissiva *see* remete a uma entrada do glossário, possibilitando o consulente conhecer as demais colocações que acompanham tal base. Pode apresentar ainda as colocações parassinôminas, remetendo o consulente a determinada subentrada de uma mesma base. Quando não houver informações para preenchimento, tanto o campo *outras designações* quanto o campo *remissivo* serão omitidos.

Na língua espanhola:

1. Colocações especializadas equivalentes em espanhol, em negrito e com a base na cor azul.

2. Frequência da colocação no *corpus* em espanhol, apresentada entre colchetes.

3. Estrutura sintagmática da colocação em espanhol.

4. Exemplo da colocação em espanhol, retirado do *corpus* de estudo em espanhol, com a colocação em destaque no texto.

5. Fonte do exemplo em espanhol, onde também se observa o ano do anuário e a página da qual o exemplo foi retirado.

6. *Otras designaciones*: assim como em inglês, este campo destina-se à inserção das colocações parassinônimas, com variações no nível morfossintático e lexical.

E na língua portuguesa:

1. Colocações especializadas equivalentes em português, em negrito e com a base na cor azul.

2. Frequência da colocação no *corpus* em português, ou na *web*, quando não foi encontrada no *corpus*. Quando o número de ocorrências for verificado na *web*, logo após a sua indicação aparece a palavra *web.*

3. Estrutura sintagmática da colocação em português.

4. Exemplo da colocação em português, retirado do *corpus* comparável.

5. Fonte do exemplo em português, marcação numérica que sinaliza um *link* para o *site* de onde os exemplos foram retirados.

6. Outras designações em português, campo reservado para inserção das colocações parassinônimas, quando houver.

Ao todo, elegemos cerca de 200 bases para integrar o glossário. A fim de ilustrá-lo, reproduzimos algumas entradas que se desdobram da base *contract*. Escolhemos dois exemplos de colocações nominais, dois de colocações adjetivais e dois de colocações verbais:

CONTRACT *noun*
barter contract [40] *(noun + noun)* Under a **barter contract** the quantity of goods to be shipped in one direction is often measured by the quantity of goods to be shipped in the other direction, rather than in terms of the market price for each shipment. (Y1992, p.99)
Other designations: contract of barter (8), contract of exchange (3)
contrato de trueque [32] *(sust. + prep. + sust.)* En virtud de un contrato de trueque, la cantidad de bienes que han de enviarse en una dirección suele calcularse en función de los que han de enviarse en la otra dirección, no en función del precio comercial de cada envío. (y1992, p.106)
Otras designaciones: contrato de permuta (9)
contrato de troca [8] *(subst. + prep. + subst.)* Na estipulação do contrato, o preço há de sempre ser fixado em dinheiro, pois é o que distingue o contrato de compra e venda, do contrato de troca ou escambo. (1)
Outras designações: contrato de permuta (6), contrato de escambo (6)

contract of carriage [3046] *(noun + noun)* "**Contract of carriage**" means a contract in which a carrier, against the payment of freight, undertakes to carry goods from one place to another. The contract shall provide for carriage by sea and may provide for carriage by other modes of transport in addition to the sea carriage. (y2008, p. 86)	
Other designations: carriage contract (9)	*see transport contract*
contrato de transporte [2816] (*sust. + prep + sust.)* Por "**contrato de transporte**" se entenderá todo contrato en virtud del cual un porteador se comprometa, a cambio del pago de un flete, a transportar mercancías de un lugar a otro. Dicho contrato deberá prever el transporte marítimo de las mercancías y podrá prever, además, su transporte por otros modos. (y2008, p. 102).	
contrato de transporte [1114] *(subst. + prep. + subst.)* **Contrato de transporte** significa um contrato no qual o transportador, mediante pagamento de frete, responsabiliza-se pelo transporte de cargas de um lugar para outro. O contrato deverá proporcionar o transporte marítimo e deve fornecer outros meios de transporte além deste. (62)	

commercial contract [266] *(adj. + noun)* The aim of this work is to give guidance to banks and other guarantors called on to issue guarantees payable on the simple or first demand of the beneficiary without proof of loss or of default in the underlying **commercial contract**. (y1983, p. 161)

contrato comercial [180] (*sust. + adj.)* El propósito de este trabajo es establecer principios rectores para uso de bancos y otros garantes a los que se pide que extiendan garantías pagaderas a la demanda simple o primera del beneficiario sin prueba de pérdida o de incumplimiento en el **contrato comercial** principal. (y1983, p. 169)

contrato comercial [299] *(subst. + adj.)* Os documentos de que trata o caput compreendem os documentos de instrução das declarações aduaneiras, a correspondência comercial, incluídos os documentos de negociação e cotação de preços, os instrumentos de **contrato comercial**, financeiro e cambial, de transporte e seguro das mercadorias, os registros contábeis e os correspondentes documentos fiscais, bem como outros que a Secretaria da Receita Federal do Brasil venha a exigir em ato normativo. (102)

formal contract [19] *(adj. + noun)* In discussing the requirements of a **formal contract** it is useful in the first place to distinguish between rules requiring the writing as the only formality and other rules requiring a second formal step, especially registration. (y1977, p. 180)

contrato solemne [13] (*sust. + adj.)* Al analizar los requisitos de un **contrato solemne** es útil distinguir en primer lugar entre las normas que requieren la forma escrita como la única solemnidad y las demás normas que requieren un segundo requisito de solemnidad, especialmente la inscripción. (y1997, p. 201-202)

Otras designaciones: contrato formal (5)

contrato formal [34] *(subst. +adj.)* **Contratos formais** ou solenes. São aqueles contratos em que não basta o mero acordo de vontades para sua formação, mas ao invés, depende de uma formalidade exigida em lei. Ou seja, só se aperfeiçoam quando o consentimento é expresso pela forma exigida em lei. (109)

Outras designações: contrato solene (11)

conclude a contract [183] *(verb + art.+ noun)* A proposal to **conclude a contract** made through one or more data messages which is not addressed to one or more specific persons, but is generally accessible to persons making use of information systems is to be regarded merely as an invitation to make offers, unless it indicates the intention of the person making the proposal to be bound in case of acceptance. (183, y2004, p. 807)

celebrar un contrato [186] (verb. + art. + *sust.)* Toda propuesta de **celebrar un contrato** presentada por medio de uno o más mensajes de datos que no vaya dirigida a una o a varias personas determinadas, sino que sea generalmente accesible para toda persona que haga uso de un sistema de información, se tendrá por una mera invitación para presentar ofertas, salvo que en ella se indique la intención de la persona que presenta la propuesta de quedar obligada en caso de aceptación. (y2004, p. 917)

celebrar um contrato [53] *(verb. + art. + subst.)* Contido no artigo 1.1 dos Princípios do UNIDROIT, a liberdade contratual determina que as partes são livres para celebrar um contrato e determinar o seu conteúdo. (141)

perform a contract [203] *(verb + art. + noun)* An abnormally low tender is one that involves a risk that "the tenderer would be unlikely to be able to **perform the contract** at [the tender price] or could do so using only substandard workmanship or materials by suffering a loss it could also indicate collusion between the tenderers". (203, y2006, p. 690)

Other designations: fulfil a contract (15) | *see fulfil a contract*

cumplir un contrato [233] (verb. + art. + *sust.)* Por oferta anormalmente baja se entiende toda oferta que suponga un riesgo de que el ofertante no pueda **cumplir el contrato** [al precio ofrecido] o que haya de hacerlo recurriendo a mano de obra o materiales de calidad inferior o incurriendo en pérdidas esta oferta puede ser además indicio de colusión entre los ofertantes (2006, p. 803)

Otras designaciones: ejecutar un contrato (115)

cumprir um contrato [30] *(verb. + art.+ subst.)* A parte que não executar as suas obrigações deve comunicar à outra parte o impedimento e os efeitos deste sobre a sua capacidade de **cumprir o contrato**. (146)

Outras designações: executar um contrato (25)

Partindo da base *contract*, o glossário apresenta cerca de 150 colocações especializadas para cada língua. O trabalho de análise das demais bases – muitas das quais também apresentam grande quantidade de colocações – prossegue, de modo que elas também possam ser inseridas no glossário.

6 Considerações finais

Discorremos, neste capítulo, sobre os procedimentos teóricos e metodológicos que possibilitaram o levantamento e a análise das colocações especializadas e estendidas dos *corpora* paralelo, em espanhol e em português, constituídos pelos anuários da UNCITRAL; e de dois *corpora* comparáveis em língua portuguesa, compilados a partir de documentos da área do Direito Comercial Internacional. Desse estudo, tivemos como resultado prático uma proposta de glossário trilíngue, nas direções tradutórias inglês→espanhol e espanhol→português. Para a proposta, debruçamo-nos sobre as colocações que se desdobraram da base *contract*, a partir da qual pudemos extrair e analisar cerca de 180 colocações para cada uma das línguas.

Dando continuidade à pesquisa, almejamos, como trabalho futuro, a análise das demais bases; estudamos ainda a possibilidade de inserção do glossário em uma plataforma *on-line*, como parte do projeto "A compilação de materiais didáticos e glossários especializados baseados em *corpora* e sua contribuição para uma Pedagogia do Léxico e da Tradução", sob responsabilidade da profª. drª. Adriane Orenha-Ottaiano. Ademais, a alocação do glossário nessa plataforma ampliaria sua divulgação, o que ajudaria também na divulgação de obras de cunho léxico-fraseográfico baseadas em *corpus*, especialmente se considerarmos seu potencial para o ensino e aprendizagem de LE, especificamente para fins específicos, e para a tradução.

Referências

BARBOSA, M. A. Estrutura, funções e processos de produção de dicionários terminológicos multilíngues. *Revista do GELNE*, Fortaleza, p. 41-44, 1999.

BARONI, M.; BERNARDINI, S. BootCaT: Bootstrapping *corpora* and terms from the web. In: LREC 2004. INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 4. *Anais*... Lisboa: Elda, 2004, p. 1313-1316.

BARROS, L. A. *Curso Básico de Terminologia*. São Paulo: Editora da Universidade de São Paulo, 2004.

BENSON, M.; BENSON, E.; ILSON, R. *The BBI Combinatory Dictionary of English*. Your Guide to collocations and grammar. Amsterdam: John Benjamins Publishing, 2009.

BERBER SARDINHA, T. *Linguística de* Corpus. Barueri: Manole, 2004.

BEVILACQUA, C. R. Fraseologia especializada: panorama das pesquisas realizadas no Brasil. In: SILVA, S. (Org.). *Fraseologia & cia:* entabulando diálogos reflexivos. São Paulo: Pontes Editora, 2016, p. 87-111. [No prelo].

BIBER, D.; CONRAD, S.; REPPEN, R. Corpus *linguistics*: investigating language, structure and use. New York: Cambridge University Press, 1998.

CETENFOLHA (*Corpus* de Extractos de Textos Electrónicos NILC/*Folha de S. Paulo*). Corpus *de cerca de 24 milhões de palavras em português brasileiro retirados do jornal* Folha de S. Paulo - os textos podem ser baixados via FTP / HTTP ou consultados no Projeto AC/DC. Disponível em: <http://www.linguateca.pt/cetenfolha/index_info.html>. Acesso em: 5 ago. 2015.

CORPAS PASTOR, G. *Manual de fraseología española*. Madrid: Gredos, 1996.

COWIE, A. P. The place of illustrative material and collocations in the design of a learner's dictionary. In: STREVENS, P. *In Honour of A. S. Hornby*. Oxford: Oxford University Press, 1978, p. 127 -139.

REAL ACADEMIA ESPAÑOLA. Corpus *de referencia del español actual:* banco de datos (CREA). Disponível em: <http://www.rae.es.> Acesso em: 15 out. 2015.

GRANGER, S.; PAQUOT, M. Disentangling the phraseological web. In: _____. (Ed.). *Phraseology*: an interdisciplinary perspective. Amsterdam: John Benjamins, 2008, p. 27-50.

MELLO, M. C. de. *Dicionário jurídico português-inglês/inglês-português.* São Paulo: Método, 2012.

HAUSMANN, F. J. Kollokationen im deutschen wörterbuch: win beitrag zur theorie des lexikographischen beispiels. In: BERGENHOLTZ, H.; MUGDAN, J. (Org.). *Lexikographie und Grammatik*. Tübingen: Niemeyer, 1985, p. 118-129.

HUNSTON, S.; FRANCIS, G. *Pattern Grammar*: A *corpus*-driven approach to the lexical grammar of English. Amsterdam: John Benjamins, 2000.

HOUAISS, A. *Dicionário eletrônico Houaiss da língua portuguesa*. Versão 3.0. Rio de Janeiro: Objetiva, 2009. 1CD.

ORENHA-OTTAIANO, A. *A compilação de um glossário bilíngue de colocações, na área de jornalismo de Negócios, baseado em* corpus *comparável.* 2004. 246 f. Dissertação (mestrado em Estudos Linguísticos e Literários). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2004.

_____. *Unidades fraseológicas especializadas:* colocações e colocações estendidas em contratos sociais e estatutos sociais traduzidos no modo juramentado e não-juramentado. 2009. 282 f.

Tese (doutorado em Estudos Linguísticos). Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, 2009.

_____. Semelhanças e diferenças entre colocações e colocações especializadas. In: ORTIZ-ALVAREZ, M. L. (Org.). *Tendências atuais na pesquisa descritiva e aplicada em fraseologia e paremiologia*. Vol. 2. Campinas: Editora Pontes, 2012, p. 147-163.

_____. Collocations workbook: um material de apoio pedagógico on-line baseado em *corpus* para o ensino de colocações em inglês. *Revista de Estudos da Linguagem*, Belo Horizonte, v. 23, p. 833-881, 2015.

RUIZ GURILLO, E. *Aspectos de fraseología teórica española*. Cuadernos de Filología. Valencia: [s/n], 1997.

SCOTT, M. *WordSmith Tools* (version 6). Stroud: Lexical Analysis Software, 2012.

SINCLAIR, J. *Looking up:* an account of the COBUILD project in lexical computing. London: Collins Cobuild, 1987.

_____. The Search for Units of meaning. In: CICLE DE CONFERÈNCIES 95-96. *Lèxix,* corpus *I diccionaris*. Barcelona: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, 1996, p. 97-107.

THE BRITISH NATIONAL *CORPUS*: version 3 (BNC XML edition). Oxford, 2007. Disponível em: <http://www.natcorp.ox.ac.uk/>. Acesso em: 15 nov. 2015.

TAGNIN, S. E. O. *O jeito que a gente diz*. Expressões convencionais e idiomáticas. São Paulo: Disal, 2013.

_____. Collecting data for a bilingual dictionary of verbal collocations: from scraps of paper to *corpora* research. In: *PALC' 99*: practical applications in language *corpora*. Lodz: Lodz University Press, 1999.

TOGNINI-BONELLI, E. Towards Translation Equivalence from a *Corpus* Linguistics Perspective. In: SINCLAIR, J. et al. (Ed.). *Grammar patterns*. London: Collins COBUILD, 1996, p. 197-217.

_____. Corpus *Linguistic at works*. Amsterdam/Philadelphia: John Benjamins, 2001.

TOGNINI-BONELLI, E.; MANCA, Elena. Welcoming children, pets and guests: towards functional equivalence in the languages of 'agriturismo' and 'farmhouse holidays. *TRADTERM*, 10, p. 295-312, 2004.

UNCITRAL. A Guide to UNCITRAL: *Basic facts about the United Nations Commission on International Trade Law*. Vienna: United Nations, 2013. Disponível em: <http://www.uncitral.org/pdf/english/texts/general/12-57491-Guide-to-UNCITRAL-e.pdf>. Acesso em nov. de 2015.

UNCITRAL. Disponível em: https://www.uncitral.org/. Acesso em nov. de 2015.

ZANCHETTA, E.; BARONI, M.; BERNARDINI, S. *Corpora* for the masses: the BootCaT front-end. In: *CORPUS* LINGUISTICS 2011 CONFERENCE, 2011, Birmingham. *Abstracts*. Birmingham: University of Birmingham, 2011.

ZULUAGA, A. *Introducción al estudio de las expresiones fijas*. Frankfurt: Peter D. Lang, 1980.

O uso de *corpus* paralelo e comparável para descrever padrões de uso na tradução de abreviaturas e acrônimos de termos médicos

The use of parallel and comparable *corpus* to describe patterns of use in the translation of abbreviations and acronyms of medical terms

Márcia Moura da Silva
Gabriele Paparelli

Resumo: O objetivo do presente artigo é apresentar pesquisa em andamento que analisa o comportamento tradutório de abreviaturas e acrônimos de termos médicos no par linguístico português-inglês para então propor um glossário *on-line* que sirva como fonte para tradutores, revisores e pesquisadores. A pesquisa tem como base teórica a tradução técnico-científica e os Estudos da Tradução baseados em *corpus* e segue alguns princípios e técnicas desenvolvidas pela Linguística de *Corpus* para compilar um *corpus* paralelo e dois *corpora* comparáveis. Ainda que para facilitar a divulgação do conhecimento científico haja uma tendência de manter esses elementos em sua forma em língua inglesa, resultados preliminares nos mostram um número estatisticamente expressivo de abreviaturas e acrônimos que seguem os padrões de suas respectivas línguas.

Palavras-chave: Abreviatura. Acrônimo. Estudos da Tradução baseados em *corpus*. Tradução.

Márcia Moura da Silva – Professora adjunta do Instituto de Letras, Departamento de Línguas Modernas, da Universidade Federal do Rio Grande do Sul, doutora em Estudos da Tradução pela Universidade Federal de Santa Catarina – marcia.moura@ufrgs.br.
Gabriele Paparelli – Aluna de graduação do curso Bacharelado em Letras, habilitação português/inglês, da Universidade Federal do Rio Grande do Sul, bolsista de Iniciação Científica PROBIC FAPERGS UFRGS – gadifab@hotmail.com.

Abstract: The purpose of this paper is to present ongoing study that analyses the translation of abbreviations and acronyms of medical terms in the Portuguese-English pair in order to propose an on-line glossary to assist translators, revisers, and researchers. The study is based on Techno-Scientific Translation and *Corpus*-based Translation Studies. It borrows some principles and techniques developed within *Corpus* Linguistics to compile one parallel *corpus* and two comparable *corpora*. Although there is a tendency to keep these elements in English as means to facilitate the spread of scientific knowledge, preliminary results show a statistically significant number of abbreviations and acronyms that follow the patterns of their respective languages.

Keywords: Abbreviation. Acronym. *Corpus*-Based Translation Studies. Translation.

1 Introdução

O presente artigo tem por objetivo descrever pesquisa em andamento[1] que partiu da experiência de uma das autoras com tradução de textos médicos nos pares linguísticos português-inglês e inglês-português[2]. Durante seu trabalho, frequentemente se deparou com a necessidade de lidar com abreviações e acrônimos, elementos bastante recorrentes nesse gênero textual. A falta de padronização resultou em pesquisas demoradas por textos paralelos ou *sites* de abreviaturas à procura de correspondentes. Mesmo quando encontrados, era preciso decidir, em meio a uma profusão de informação na internet (nem sempre confiável), se determinada abreviatura era mais comumente traduzida ou deixada na forma que aparecia em língua inglesa, por exemplo. Assim, pensou-se na criação de material de pesquisa que pudesse facilitar o trabalho de tradutores e outros profissionais e pesquisadores de tradução.

Academicamente falando, basta uma busca pelas dissertações e teses defendidas nos programas de pós-graduação das universidades brasileiras para constatar que, embora as pesquisas com tradução de textos médicos tenham aumentado nos últimos anos, ainda há pouco interesse acadêmico em investigar as escolhas tradutórias para abreviaturas e acrônimos.

Entre as pesquisas existentes na área, podemos citar alguns trabalhos nos pares linguísticos inglês-português e português-inglês com variados enfoques, que incluem metodologia de *corpus* (COULTHARD, 2005), tradução de resumos médicos (PASQUALI; PINTO, 2013) e tradução de legendagem (COLLET, 2012) . Entre os trabalhos existentes, destacamos o de Paiva (2009), que tem por objeto as subáreas de cardiologia e cardiovascular. A pesquisadora vale-se dos Estudos

[1] *A tradução de abreviaturas e acrônimos de termos médicos no par linguístico português-inglês.*

[2] Por muitos anos, a professora Márcia M. da Silva traduziu material na área da saúde que incluía transcrições no par linguístico português-inglês e material impresso no par linguístico inglês-português. As traduções eram parte de grandes projetos que envolviam médicos, profissionais da saúde e representantes de instituições governamentais da área da saúde.

da Tradução baseados em *corpus* (ETC) como pressuposto teórico, tendo como objetivo observar as estratégias que evidenciam traços de simplificação e de explicitação[3]. Embora o estudo aqui descrito compartilhe com o trabalho de Paiva a observação de textos médicos com base nos ETC, o foco nas abreviaturas e acrônimos em uma subárea distinta (reumatologia) enriquece a discussão sobre a tradução do texto médico em geral, com destaque para a situação real de uso dos itens objeto desta investigação.

Ademais, acreditamos que a observação, por meio de *corpora*, de um maior número de especialidades amplia a pesquisa na área médica como um todo, possibilitando verificar se determinados padrões de uma especialidade se reproduzem em outras, por exemplo, ou mesmo mostrar particularidades de certas subáreas.

Ainda que haja, como nos lembra Franco Aixelá (2009), certa disposição para manter abreviaturas inalteradas em textos traduzidos do inglês para preservar o caráter de dada disciplina e facilitar a divulgação do conhecimento técnico-científico, como mencionado, a prática nos mostra que esses elementos estão longe de ser estanques e que tampouco há consenso em relação a como traduzi-los, o que acaba por dificultar o trabalho do tradutor. Como grande parte da divulgação do conhecimento técnico-científico é feita em língua inglesa, a tradução de artigos médicos para essa língua vem se tornando imprescindível. Assim sendo, o principal objetivo da pesquisa em andamento é verificar o comportamento tradutório em relação a abreviaturas e acrônimos de termos médicos da área da reumatologia, no par linguístico português-inglês, para que se possa propor um glossário *on-line* como material de consulta.

A pesquisa tem como base teórica a tradução técnico-científica (AZENHA, 1996, 1999; AIXELÁ, 2009) e os Estudos da Tradução baseados em *Corpus* (BAKER, 1993, 1995; OLOHAN, 2004; TYMOCZKO, 1998). Um *corpus* paralelo e dois *corpora* comparáveis, que serão descritos em maiores detalhes na metodologia, serão compilados seguindo alguns princípios e técnicas desenvolvidos pela Linguística de *Corpus* (BAKER, 2013; BERBER SARDINHA, 2002, 2004; BIBER, 1993; LEECH, 1991). A combinação desses *corpora* possibilitará verificar se determinadas abreviaturas e acrônimos são mantidos inalterados ou traduzidos, bem como seus padrões de uso, indicando se determinado padrão está restrito ao texto traduzido ou se é mais frequente nele. Espera-se que tal conhecimento possa auxiliar tradutores e pesquisadores na produção de textos que tenham mais aceitação na comunidade internacional. Para Olohan (2004), a influência da língua de partida (LP) em padrões de uso observados na língua de chegada (LC) seria um dos principais benefícios da combinação entre esses dois

[3] Na *explicitação*, o tradutor adiciona, de maneira explícita, no TC, componentes que estariam somente implícitos no TP e na *simplificação*, o tradutor usa uma linguagem mais simples no TC para facilitar a compreensão do leitor.

tipos de *corpora*. Porém, visto que as traduções objeto desta pesquisa são em língua inglesa, e dada sua influência na área aqui pesquisada, espera-se o cenário inverso, ou seja, maior influência da língua inglesa sobre as abreviaturas e acrônimos usados em língua portuguesa, embora resultados preliminares com o *corpus* paralelo já apontem um número estatisticamente relevante de abreviaturas e acrônimos que seguem os padrões das respectivas línguas, como mostramos nos resultados preliminares abaixo.

A escolha da subárea da reumatologia se deu pelo fato de grande parte do trabalho de tradução feito por uma das autoras ter sido nessa área, sobretudo em artrite reumatoide (AR). Segundo o Ministério da Saúde, em 2011, as doenças reumáticas já acometiam 12 milhões de brasileiros[4]. Como o tratamento medicamentoso dessas doenças é garantido pelo Sistema Único de Saúde (SUS), há bastante interesse por parte dos grandes laboratórios médicos em negociar a venda de produtos de última geração ao governo brasileiro, daí o número considerável de projetos de tradução na área. Embora o foco seja nas abreviaturas e acrônimos usados na subárea da reumatologia, não se desprezarão abreviaturas e acrônimos que sejam compartilhados com outras subáreas, como o exemplo do desfecho clínico[5] *QoL* (*quality of life/qualidade de vida*), ou micro-organismos que impactam doenças tratadas pela reumatologia, como é o caso de HIV (vírus da imunodeficiência humana) e HPV (papilomavírus humano.)

Em relação à escolha dos textos, essa se deu pelo fato de serem esses textos de revistas representativas dessa subárea e também por sua respeitabilidade. Por serem publicações de importantes sociedades médicas, pressupõe-se que estejam em conformidade com os padrões de aceitabilidade de cada língua.

É importante mencionar que, como acontece com o texto acadêmico, a publicação do texto médico em língua inglesa dá ao trabalho maior visibilidade internacional, visto ser o inglês *lingua franca* e ter, assim, abrangência global. Ademais, países como Estados Unidos e Inglaterra são detentores de grande conhecimento nessa área e financiadores de pesquisas robustas. Dessa maneira, são muito mais comuns textos traduzidos para o inglês do que o contrário, por isso o *corpus* paralelo consiste nos originais em língua portuguesa e suas traduções em língua inglesa.

A seguir, discutimos brevemente os pressupostos teórico-metodológicos que nortearão a pesquisa aqui apresentada.

---

[4] Disponível em: <http://www.blog.saude.gov.br/promocao-da-saude/29041-saude-alerta-para-prevencao-as-doencas-reumaticas>.

[5] "Evento de investigação supostamente causado pelo fator em estudo. Exemplo doença, complicação, efeito terapêutico" (disponível em: <http://pt.slideshare.net/flavioes/tipos-de-estudo-1814732>).

2 Tradução técnico-científica

Segundo Aixelá (2004), a tradução técnico-científica não recebe a mesma reverência que a tradução literária dentro dos Estudos da Tradução porque seus textos são vistos como menos criativos e elaborados, requerendo muito pouco além de conhecimento terminológico. Porém, o autor sugere que essa tradução se destaca entre outros gêneros, pertencendo a um campo de pesquisa independente. Para ele, diferentemente de outros tipos de traduções, a técnico-científica precisa respeitar tanto a função referencial da língua quanto as convenções da linguagem técnica no que diz respeito à precisão para melhor se entender o mundo.

Aixelá (2009) trabalha com a noção de que a tradução é regida por uma tensão dupla – a força centrípeta, que a leva em direção às propostas do texto de partida (TP), e a centrífuga, que a leva em direção ao texto de chegada (TC) no que diz respeito às noções do que vem a ser correto e também às expectativas do leitor. Segundo o autor, ao usar exclusivamente a primeira, o tradutor estaria criando um texto repleto do que denomina "interferência", que pode ser minimizada quando as forças são equilibradas. Porém, em consequência dos limites de entrega bastante curtos com os quais trabalham os tradutores, além de remunerações nem sempre ideais, Aixelá acredita que tradutores somente se desviam do TP se julgam estritamente necessário, pois é mais rápido produzir uma tradução mais próxima a ele.

Em relação a textos técnico-científicos, Aixelá (2009) sugere que pessoas que compartilham uma profissão tendem a criar sua própria terminologia, seja por necessidade (precisão e clareza), ou exclusividade, o que acaba gerando certa opacidade para as pessoas de fora, mas que fortalece o sentimento de pertencimento dos profissionais em questão. Outro aspecto importante é o papel da língua inglesa, como já mencionado, que acaba por fazer com que traduções a partir dessa língua tenham um grau maior de interferência, aumentando, assim, a probabilidade de aceitação, por estarem promovendo a internacionalização da terminologia, facilitando, assim, o fluxo de conhecimento técnico-científico. Como há vários neologismos em inglês nessa área, tradutores tendem a mantê-los para que seus textos ganhem prestígio e exclusividade. Por outro lado, as interferências podem ser rejeitadas porque existe a vigilância de autoridades na cultura de chegada que se preocupam com a pureza linguística.

Outro autor que discorre sobre o texto técnico-científico e sua tradução, Azenha (1999) também nos adverte sobre essa ideia generalizada de que os textos técnicos são estanques e que os problemas de tradução relacionados a eles estão restritos ao plano léxico-terminológico. Para o autor, eles são "formas híbridas expostas à ação de um número elevadíssimo de variáveis e a terminologia, longe de ser estática, é dinâmica e admite uma margem de subjetividade no tratamento de seu objeto" (p. 11). O sistema de pesos e medidas é um bom exemplo dessa variedade de textos técnicos. Ao procurar uma receita de bolo inglês na internet,

por exemplo, o brasileiro espera encontrá-la no sistema métrico, não no imperial, que é o mais usado na Grã-Bretanha. Assim, ao se traduzir essa receita para o português, deve-se levar essa particularidade em consideração e buscar estratégias apropriadas para melhor entendimento por parte do leitor de chegada, visto que a tradução literal por si só não daria conta desse objetivo.

Ainda que textos técnico-científicos estejam sujeitos à variação, Aixelá (2009) observa que os tradutores desse tipo de texto optam pela interferência, ou seja, por manter a terminologia inalterada, em vez de uma tradução mais "pura". Ele também nos lembra que muitos tradutores em formação se queixam de professores que exigem pureza na LC, quando, de fato, o mercado rejeita traduções que não se enquadram em modelos internacionais. Por outro lado, Azenha (1999) enfatiza que, pela natureza dinâmica desses textos, é de vital importância que os professores conscientizem seus alunos para a variabilidade desse tipo de tradução e também os encorajem a exercer o pensamento crítico, pois somente assim estarão mais bem preparados para as demandas do mercado.

Em relação a essas demandas, embora as pesquisas acadêmicas na área dos Estudos da Tradução tenham tradicionalmente o texto literário como foco, a demanda de mercado pela tradução técnico-científica é bem maior. Assim, ainda que possa existir uma tendência à não tradução de certos elementos dentro do texto técnico-científico, como aponta Aixelá (2009), para se conformar a determinados padrões, é importante que tradutores em formação sejam conscientizados de que existe mais de uma solução para a tradução de terminologia e que o grau de interferência depende do quanto ela é aceita na cultura de chegada. Dessa maneira, futuros tradutores estarão mais aptos a suprir essa demanda.

Como pesquisadoras, interessa-nos observar as variações inerentes ao texto técnico-científico e as soluções encontradas por tradutores na produção de um TC adequado. Porém, independentemente do fato de um tradutor traduzir ou não termos técnicos, engana-se quem pensa que a decisão de transpô-los inalterados para o TC seja uma decisão automática, pois, para tanto, o tradutor deve considerar aspectos que vão além da questão terminológica. É nesse sentido que vemos vantagem na combinação de estudos voltados a textos técnico-científicos e a metodologia de *corpus*, pois, graças aos avanços nas ferramentas de construção de *corpus*, a observação de outras traduções vem se tornando cada vez mais precisa e imediata e vem aumentando consideravelmente a confiabilidade dos resultados por permitir que um número representativo de textos seja incluído em um único *corpus*. Ademais, como essas ferramentas possibilitam a observação do objeto de estudo em contexto, tornou-se mais fácil expandir as discussões para além do plano linguístico. Porém, antes de descrevermos mais detalhadamente o processo de compilação dos *corpora* que estão sendo construídos para a análise aqui proposta, tecemos algumas considerações breves sobre a tradução de abreviaturas e acrônimos e sobre os Estudos da Tradução baseados em *Corpus*.

2.1 Tradução de abreviaturas e acrônimos

Trabalhamos com a noção de que a abreviatura se refere à "apresentação de uma palavra por meio de algumas de suas sílabas ou letras"[6], tendo em conta seus aspectos fônicos, como AAS (Ácido Acetilsalicílico); enquanto o acrônimo "é pronunciado como se fosse uma palavra única, como OTAN e NASA"[7].

Do ponto de vista morfológico, de um modo geral, as abreviaturas[8] se classificam em i) simples, formadas pelas iniciais ou qualquer outra letra de um grupo de palavras (e.g. ROM/*Read-Only Memory*); ii) palavras cortadas (*clipped words*), formadas por três ou mais letras contíguas de uma mesma palavras (e.g. del/delete); iii) combinações, que são junções de palavras previamente cortadas ou abreviadas (e.g. *modem* /*modulator-demodulator*) e iv) complexas, que combinam duas ou mais abreviaturas simples por meio de símbolos tipográficos, como é o caso do hífen, (e.g. CD-ROM). (Cf. BELDA, 2004).

Em relação à tradução dessas reduções de palavras, para Aixelá (2009), por exemplo, há certa disposição em mantê-las inalteradas em textos traduzidos do inglês para manter o caráter da disciplina a qual pertence e facilitar a divulgação do conhecimento técnico-científico:

> Assim, se você deseja que sua disciplina seja descrita em seus próprios termos, geralmente é mais fácil importar palavras e estruturas pré-fabricadas do que criar outras; além disso, há a vantagem da exclusividade devido ao fato de que termos importados tendem a ser mais opacos que outros derivados de palavras preexistentes na LC (sendo essa uma das razões pela qual os jargões técnicos modernos tendem a ser mais opacos para o leitor geral quando não estão em inglês). Esse motivo é também reforçado pelo argumento da promoção da internacionalização da terminologia, que facilitaria, assim, o fluxo de conhecimento técnico-científico. *Essa é uma justificativa importante e recorrente dada para a não tradução de abreviaturas, que provavelmente representam o grau máximo de interferência na tradução técnico-científica.* (AIXELÁ, 2009, p.80-81, grifo e tradução nossos[9])[10]

---

[6] Dicionário Aurélio.

[7] "is pronounced as if it were a single word, in the manner of NATO and NASA." (Farfex Dictionary: <http://www.thefreedictionary.com/acronym>).

[8] Ao apresentar tal classificação, Belda (2004) refere-se somente à abreviatura, porém, seguindo as definições que adotamos, alguns de seus exemplos incluem também acrônimos.

[9] Todas as traduções feitas das citações em língua inglesa foram feitas por nós.

[10] "Thus, if you want your discipline to be described in its own terms, it is generally easier to import ready-made words and structures than to create new ones, not forgetting the bonus of exclusivity due to the fact that imported terms tend to be more opaque than others derived from pre-existing TL words (one of the reasons why modern technical jargons tend to be more opaque for the general reader when not in English). This motive is also supported by the argument of promoting the internationalisation of your terminology, and thus facilitating the flow of scientific and technical knowledge. This is an important and often quoted reason for the nontranslation of abbreviations, which probably represent the maximum degree of interference in technical and scientific translation".

Já Belda (2004), que escreve sobre a tradução de abreviaturas na área da tecnologia da informática no par linguístico inglês-espanhol, chama a atenção para um dos principais aspectos que o tradutor deve considerar na tradução de abreviaturas – a ordem das palavras. A língua inglesa é flexível o suficiente para permitir a conversão de substantivos em adjetivos simplesmente colocando-os antes de outros substantivos. Em espanhol, e também em português, tal conversão não é comum, sendo que normalmente usa-se algum conector, sobretudo preposição, como é o caso de *CTS* (*carpal tunnel syndrome*), traduzido para o português como STC (síndrome do túnel do carpo)[11]. Como o *corpus* paralelo da pesquisa aqui descrita consiste em textos escritos originalmente em português com suas respectivas traduções para o inglês, tal conversão ocorre no texto traduzido.

O autor, porém, ratifica a ideia de que a prática tradutória segue o padrão de uso que, por sua vez, está atrelado à aceitação ou não por profissionais da LC de termos inalterados. O autor dá o exemplo da tentativa de usar a abreviatura espanhola MMM (*Malla Máxima Mundial*) para a inglesa WWW (*World Wide Web*), uma proposta feita pelo CVC (Centro Virtual Cervantes), que acabou não se concretizando.

3 Estudos da Tradução baseados em *corpus*

Baker (1995) define *corpus* como sendo uma

> coleção de textos armazenados em formato digital que podem ser analisados automaticamente ou semiautomaticamente de várias maneiras [...] O mais importante é que seja construído para um propósito específico seguindo critérios explícitos de desenho a fim de garantir que seja representativo da área ou língua que pretende estudar. (BAKER, 1995, p. 225)[12]

Para Baker (1993), que deu início às pesquisas de Estudos da Tradução baseados em *corpus*, os textos traduzidos *"registram eventos comunicativos genuínos e, como tal, não são inferiores nem superiores a outros eventos comunicativos em qualquer língua. Contudo, são diferentes, e a natureza dessa diferença precisa ser explorada e registrada*"[13] (p. 234).

---

[11] Também foram encontradas outras variações do termo em português como, por exemplo, *síndrome to túnel cárpico, síndrome do túnel carpal* e *síndrome do túnel de carpo*.

[12] "collection of texts held in machine-readable form and capable of being analysed automatically or semi-automatically in a variety of ways […] What is important is that it is put together for a particular purpose and according to explicit design criteria in order to ensure that it is representative of the given area or of language it aims to account for."

[13] "*record genuine communicative events and as such are neither inferior or superior to other communicative events in any language. They are however different, and the nature of this difference needs to be explored and recorded*."

Essa visão do texto traduzido tem como origem os Estudos Descritivos da Tradução, apresentados por James Holmes em seu artigo seminal, de 1972, *The Name and Nature of Translation Studies*. Segundo o autor, os estudos descritivos *"descrevem o fenômeno do ato tradutório e da(s) tradução/traduções como se manifestam no mundo de nossa experiência*"[14,15].

Os trabalhos de Baker apoiam-se sobretudo nos trabalhos de Even-Zohar e Gideon Toury[16]. Na teoria do polissistema de Even-Zohar, na qual um determinado polissistema literário é parte de um polissistema cultural maior, os textos traduzidos são vistos como parte de seus originais, assim, ganhando relevância e sendo considerados dignos de investigação como um sistema em si.

Os ETC emprestam da Linguística de *Corpus* alguns de seus princípios e algumas das técnicas por ela desenvolvidas sobretudo no que diz respeito à metodologia de compilação de *corpus*, mas firmam-se como uma abordagem com forte base descritiva e que dá espaço a investigações bastante abrangentes[17], como atesta Tymoczcko (1998, p. 653):

> Os Estudos da Tradução com *Corpus* (ETC) surgiram em um momento crítico na disciplina dos Estudos da Tradução. Emergindo da linguística de *corpus* e, portanto, inerentemente fiel às abordagens linguísticas da tradução, os ETC, ao mesmo tempo que marcam uma mudança das abordagens prescritivas para as descritivas, [...] levam em consideração tanto os pormenores do texto escolhido pelo tradutor individual, como padrões culturais mais amplos, sejam internos ou externos ao texto. Embora o material para os *corpora* estejam baseados nos meios linguísticos da tradução, as perguntas feitas a esses *corpora* podem servir para tratar não somente questões de língua e linguística, mas também questões de cultura, ideologia e crítica literária.[18]

Os estudos com *corpora* permitem, assim, estudar o fenômeno da tradução de diferentes pontos de vista e a partir de diferentes línguas e culturas, revelando tanto

---

[14] Cf. Shuttleworth, M; Cowie, M. Dictionary of Translation Studies. Manchester (1997, p. 39).

[15] "*describe the phenomena of translating and translation(s) as they manifest themselves in the world of our experience*".

[16] Toury's *Descriptive Translation Studies- and Beyond* (1995) "has become an indispensable reference point for those working in this area. Toury's early work came out of the context of polysystem theory, developed by his colleague Itamar Even-Zohar" (MUNDAY, 2009).

[17] Em geral, os ETC não estão restritos à descrição de fenômenos linguísticos. A pesquisa de Baker (2013) sobre a representação da palavra *mulçumano* (*Muslim*) na imprensa britânica, por exemplo, tem por base a Análise Crítica do Discurso.

[18] "*Corpus* translation studies (CTS) has emerged at a critical time in the discipline of Translation Studies. Growing out of *corpus* linguistics and thus inherently having an allegiance to linguistic approaches to translation, CTS at the same time marks a turn away from prescriptive approaches to translation toward descriptive approaches […] and takes into account the smallest details of the text chosen by the individual translator, as well as the largest cultural patterns both internal and external to the text [.A]lthough the materials of *corpora* are based upon the language medium of translations, interrogation of *corpora* can nonetheless serve to address not simply questions of language or linguistics, but also questions of culture, ideology, and literary criticism."

as similaridades como as diferenças. Tymoczko (1998, p. 253) aponta ainda que ao longo da história da teoria da tradução as diferenças eram vistas negativamente e esperava-se um TC mais fiel ao TP; porém, hoje as diferenças são mais valorizadas.

A presente pesquisa utiliza dois tipos de *corpora*, a saber:

a) *Corpus* paralelo – coleção de textos originais na língua A e suas traduções em língua B.
b) *Corpus* comparável – i) *corpus* comparável monolíngue: consiste em coleção de textos originais em língua A e textos traduzidos em língua A e ii) *corpus* comparável bilíngue: coleção de textos originais em língua A e textos originais em língua B.

Como mencionado, Olohan (2004) acredita que a combinação desses dois tipos de *corpora* seja bastante vantajosa, pois através dela pode-se observar a influência da LP em padrões de uso observados na LC:

> Uma abordagem metodológica que promete emergir em trabalhos futuros é a combinação de resultados de uma análise com *corpora* comparáveis e paralelos [...] Essa abordagem dupla pode ser particularmente frutífera para mensurar a influência do texto de partida sobre padrões de uso observados na língua traduzida. (OLOHAN, 2004, p. 192)[19]

3 Metodologia

Berber-Sardinha (2004, p. 18) aponta ser o *corpus* uma coleção de textos naturais grande e criteriosa que deve refletir apropriadamente a variedade escolhida "o mais fielmente possível". Em caso de variedade específica, sugere o autor, o pesquisador deve ter o cuidado de escolher material que reflita pontualmente tal variedade para evitar vieses e contaminações.

Tal variedade acaba sendo um critério sugerido também em relação ao tamanho ideal do *corpus* de estudo, sobre o qual ainda não há consenso. Sinclair (1997), cujo trabalho está voltado para a língua geral para propostos lexicográficos, sugere que o *corpus* deva ser o mais extenso possível. Embora seja da mesma opinião, Berber-Sardinha (2004, p. 29) destaca a impossibilidade de se incluir um idioma inteiro em um *corpus,* por isso a necessidade de "*delimitar ao máximo a variedade (tipo de texto, por exemplo) incluída no* corpus".

[19] "A methodological approach that is likely to come to the fore in future work is the combination of findings from comparable *corpus* analysis and parallel *corpus* work, [...] This dual approach proves particularly fruitful in measuring the extent of any source-text influence on patterns of usage observed in translated language" (OLOHAN, 2004, p. 192).

Embora o tamanho seja um dos principais critérios a se pensar ao compilar um *corpus*, sua representatividade é considerada ainda mais relevante. Para Leech (1991, p. 27), um *corpus* é representativo *"contanto que os resultados baseados em seus conteúdos possam ser aplicados a um* corpus *hipotético maior*"[20].

Já Biber (1993) defende uma pesquisa teórica criteriosa antes de se desenhar um *corpus* para melhor identificar as características linguísticas que serão analisadas. Berber-Sardinha (2004, p. 29) ressalta que o *corpus* "*deve ser adequado aos interesses do pesquisador, que deve ter uma questão a investigar para a qual necessite de um* corpus *específico*".

No caso dos *corpora* compilados para esta pesquisa, acreditamos serem representativos, pois temos claro nosso objetivo de observar o comportamento tradutório para abreviaturas e acrônimos na subárea da reumatologia, assim como o padrão de uso em textos escritos originalmente em língua portuguesa e em língua inglesa para servir a uma determinada população, ou seja, tradutores, revisores e pesquisadores dessa área. Tal representatividade é reforçada pelo fato de que os textos foram extraídos de periódicos bem conceituados e específicos da subárea de estudo, o que faz com que os *corpora*, ainda que não sejam grandes, tenham uma concentração maior de terminologia.

A seguir, descrevemos resumidamente o processo de compilação dos *corpora*.

3.1 Corpora

No presente estágio da pesquisa, estamos compilando um *corpus* paralelo com textos originais em língua portuguesa com suas traduções em língua inglesa da *Revista Brasileira de Reumatologia.* No próximo estágio, compilaremos dois *corpora* comparáveis: i) *corpus* com os textos traduzidos em língua inglesa dessa revista e textos escritos originalmente em inglês da revista *Rheumatology* (*corpus* comparável monolíngue), e ii) *corpus* com textos escritos originalmente em português da *Revista Brasileira de Reumatologia* e textos escritos originalmente em inglês da revista *Rheumatology* (*corpus* comparável bilíngue). A análise dos *corpora* está sendo feita utilizando as ferramentas ParaConc e AntConc, cujas funções são descritas na próxima seção.

Neste primeiro momento, optamos por trabalhar com as edições 2009 e 2010 (volumes nº 49 e 50, respectivamente) da *Revista Brasileira de Reumatologia*, sendo o volume nº 49 a primeira edição a ser traduzida para o inglês. Até o momento, nosso *corpus* paralelo conta com 117 textos em português e inglês, com um total de 284.315 *tokens*[21] em português e 273.259 em inglês e com 17.117 *types* em português e 12.732 *types* em inglês.

[20] "*to the extent that findings based on its contents can be generalized to a larger hypothetical* corpus."
[21] Enquanto *tokens* se referem ao número total de palavras em um *corpus, types* se referem às palavras diferentes.

3.2 Escolha das ferramentas

Após refletir sobre qual *software* atenderia as necessidades de nossa pesquisa, optamos por utilizar o programa ParaConc 1.0, que é um concordanciador bilíngue usado em análises contrastivas, desenvolvido por Michael Barlow[22]. O ParaConc possibilita gerar listas de frequência das palavras de texto alinhados, buscar por palavras específicas, além da possibilidade de alinhar até quatro textos paralelos, que podem ser em quatro línguas diferentes, ou um texto original mais três traduções diferentes. Apesar de sua interface pouco amigável e intuitiva, o ParaConc é uma das poucas ferramentas disponíveis que permite salvar o texto alinhado, além de oferecer a opção de salvar o andamento do trabalho (*workspace*). Ademais, devido à falta de atualizações que visem melhorar o programa, passamos por inúmeras dificuldades e a necessidade de refazer boa parte dos alinhamentos, pois quando o arquivo alinhado é exportado e salvo, algumas frases acabam perdendo seu alinhamento no arquivo de texto resultante. Sua utilização requer um alto grau de paciência e atenção redobrada por parte do usuário.

Já os *corpora* comparáveis serão processados com a ferramenta AntConc, uma ferramenta de livre acesso, desenvolvida por Laurence Anthony[23]. Essa ferramenta, em particular, permite ao pesquisador: i) gerar listas de palavras, em ordem de frequência, de todas as palavras, de todas as palavras que constam nos arquivos de textos selecionados; ii) gerar lista de clusters/N-grams, que permite observar como determinado termo se combina com outras palavras do *corpus* e iii) gerar lista de palavras-chave (Keyword List), resultante da comparação entre a lista de frequência das palavras do *corpus* de estudo com a do *corpus* de referência, um *corpus* maior que o *corpus* de estudo, como é o caso do BNC (British National *Corpus*), uma coleção de 100 milhões de palavras em inglês britânico que inclui tanto a língua escrita como a falada. Tal comparação permite que o pesquisador identifique palavras que são estatisticamente mais frequentes no *corpus* de estudo. Quanto maior for a frequência estatística de uma palavra no *corpus* de estudo, maior será sua especificidade.

3.3 Limpeza dos arquivos, catalogação e primeiras extrações

Uma vez que selecionamos os volumes da *Revista Brasileira de Reumatologia* que disponibilizam os artigos em língua portuguesa e sua tradução em língua inglesa, esses foram salvos em formato arquivo de texto (*.txt). Após diversos testes com o ParaConc, utilizamos a codificação ANSI para os arquivos do *corpus* paralelo, posto que a ferramenta funciona melhor com essa codificação (já para o

[22] Disponível em: <http://www.paraconc.com/>.
[23] Disponível em: <http://www.laurenceanthony.net/software/antconc/>.

corpus comparável, será utilizada a codificação UTF-8. O AntConc aceita diversas codificações, a UTF8 é a mais apropriada para línguas com muitos diacríticos, como é o caso do português. O primeiro passo foi remover dos artigos todos os elementos extratextuais, como tabelas, gráficos, imagens, agradecimentos e *links* externos. Elementos como sinais de porcentagem e abreviações latinas (e.g. et al. e apud), que entraram em conflito com a ferramenta ParaConc ou que geraram ruído nos resultados, também tiveram que ser removidos.

Em caso de necessidade de recuperação de informação, embora tenhamos optado por não usar o cabeçalho, estabelecemos o sistema de catalogação, exemplo do qual mostramos na Tabela 1, abaixo:

Tabela 1 – Catalogação de arquivos

Código	Autor,Título, Fonte	Disponível em
pt2010-01-01P	Elisabeth Gonzaga Canova Fernandes; Vanessa Ramos Guissa; Cynthia Saviolli; José Tadeu Tesseroli Siqueira; Marcelo Valente; Clovis Artur Almeida da Silva . Osteonecrose de mandíbula em pacientes com lúpus eritematoso sistêmico juvenil observada em exame de imagem. Rev. Bras. Reumatol. vol.50 no.1 São Paulo jan./fev. 2010	http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0482-50042010000100002&lng=pt&nrm=iso&tlng=pt (acesso em 20 Mar. 2017)
en2010-01-01P	Elisabeth Gonzaga Canova Fernandes; Vanessa Ramos Guissa; Cynthia Saviolli; José Tadeu Tesseroli Siqueira; Marcelo Valente; Clovis Artur Almeida da Silva. Osteonecrosis of the jaw on imaging exams of patients with juvenile systemic lupus erythematosus. Rev. Bras. Reumatol. vol.50 no.1 São Paulo jan./fev. 2010	http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0482-50042010000100002&lng=pt&nrm=iso&tlng=en (acesso em 20 Mar. 2017)

Fonte: Elaborada pelas autoras

Em relação à busca por abreviaturas e acrônimos, basta que se carreguem todos os arquivos alinhados na ferramenta ParaConc para se gerar a lista de frequência das palavras, por ordem alfabética ou por frequência. O ParaConc ainda oferece opções para refinar a lista gerada, como a opção que permite ignorar as palavras maiúsculas. Para esta pesquisa, porém, optamos por não ignorá-las simplesmente porque as abreviaturas e acrônimos aparecem notadamente em letras maiúsculas.

A lista de frequência gerada nos permite procurar candidatos a abreviaturas e acrônimos. Com esses candidatos em mãos, pode-se gerar uma lista específica (por meio da opção "search" da ferramenta), para observar esses elementos em seus contextos de uso nos pares linguísticos trabalhados. A partir dessa lista, pode-se também verificar as formas por extenso das abreviaturas e acrônimos, como é possível observar na Figura 1 o exemplo da abreviatura LES (Lúpus Eritematoso Sistêmico).

No caso das abreviaturas e acrônimos, interessa-nos extrair as formas por extenso, para que também possamos identificar como são traduzidas (*corpus* paralelo), e seu padrão de uso (*corpora* comparáveis). O conhecimento dessa relação entre abreviatura/acrônimo e sua forma por extenso poderá auxiliar tradutores e pesquisadores a escolherem colocações já consolidadas no campo.

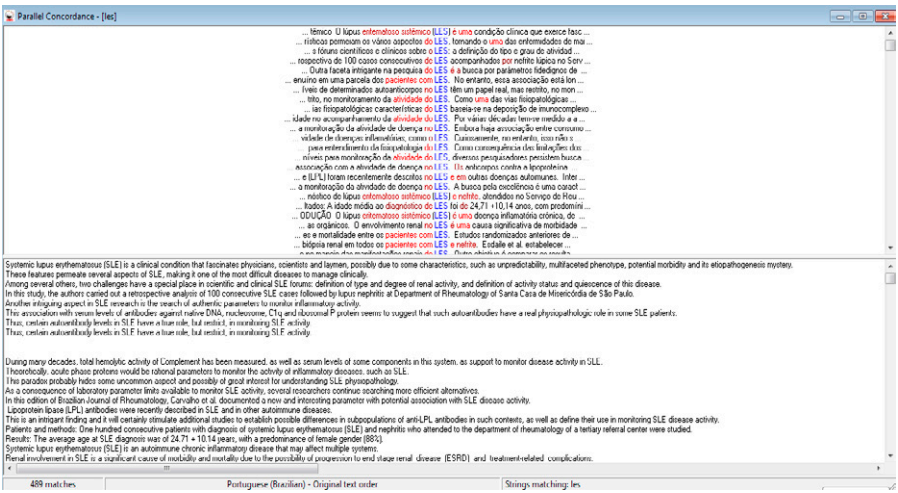


Figura 1 – Opção "search" do programa ParaConc
Fonte: Elaborada pelas autoras

4 Resultados preliminares

Como a pesquisa se encontra em estágio de compilação dos *corpora*, este trabalho não visa a apresentar resultados conclusivos. Contudo, de uma lista de frequência das cinquenta primeiras ocorrências de abreviaturas e acrônimos (a Tabela 2 mostra as cinco primeiras mais frequentes), já foi possível observar que há um número expressivo de abreviaturas que seguem os padrões das línguas em questão (54%), o que refuta em parte a sugestão de Aixelá (2009) de que haja certa disposição em manter abreviaturas inalteradas em textos traduzidos do inglês para preservar o caráter de determinada disciplina e também para facilitar a divulgação do conhecimento técnico-científico, que, como já mencionado, é feita sobretudo em língua inglesa. Em nosso *corpus* paralelo, por exemplo, a abreviatura LES/SLE (Lúpus eritematoso sistêmico / *Systemic lupus erythematosus*) aparece 386 vezes em português e 378 vezes em inglês, sendo a abreviatura mais frequente em nosso *corpus*. Já a segunda abreviatura mais frequente, *AR/ RA (Artrite Reumatoide/ Rheumatoid arthritis),* aparece 119 vezes em português e 128 em inglês. Em seguida, temos ES/SSc (Esclerose Sistêmica / *Systemic Sclerosis*), PCR/CRP (Proteína

C-reativa / *C-reactive protein*) e LESJ/JSLE (Lúpus Eritematoso Sistêmico/Juvenil / *Juvenile systemic lupus erythematosus)*, que aparecem, respectivamente, 117 vezes em português e 95 em inglês; 101 vezes em português e 72 em inglês e 97 vezes em português e 103 em inglês .

Tabela 2 – As 5 abreviaturas/acrônimos traduzidos mais frequentes

	Freq	PT	Forma por Extenso	Tradução	Forma por Extenso	Freq
1	386	LES	Lúpus eritematoso sistêmico	SLE	Systemic lupus erythematosus	378
2	119	AR	Artrite Reumatoide	RA	Rheumatoid arthritis	128
3	117	ES	Esclerose Sistêmica	SSc	Systemic Sclerosis	95
4	101	PCR	Proteína C-reativa	CRP	C-reactive protein	72
5	97	LESJ	Lúpus Eritematoso Sistêmico Juvenil	JSLE	Juvenile systemic lupus erythematosus	103

Fonte: Elaborada pelas autoras

Em relação às abreviaturas e acrônimos que são iguais nas duas línguas (46%), é interessante notar que, apesar de esses elementos terem se mantido iguais, as formas por extenso de três das cinco ocorrências mais frequentes (Tabela 3) seguem suas respectivas línguas, sendo que as outras duas, NK (*Natural Killer Cell* / *célula Exterminadora Natural*) e IgM (*Immunoglobulin M* / *Imunoglobulina M*) aparecem no *corpus* sem as formas por extenso (s/n).

Tabela 3 – As 5 abreviaturas/acrônimos inalterados mais frequentes

9	78	PRL	Proclatina	PRL	Proclactin	76
10	71	NK	s/n	NK	s/n	72
12	60	IgM	s/n	IgM	s/n	61
14	54	(anti-)LPL	Lipoproteína lipase	(anti-)LPL	Lipoprotein lipase	47
15	52	SLEDAI	Índice de atividade da doença	SLEDAI	Systemic Lupus Erythematosus Disease Activity Index	48

Fonte: Elaborada pelas autoras

Em uma análise parcial dessas primeiras 50 ocorrências, parece-nos que há uma tendência em usar as abreviaturas e acrônicos nas respectivas línguas quando esses se referem a denominações de doenças (e.g. LES, AR, ES e LESJ), enquanto permanecem inalterados quando se referem a elementos como hormônios (e.g. PRL), proteínas (e.g. IgM, LPL), escores mensuradores de atividade de doença (e.g. SLEDAI), vírus, como é o caso de HIV (vírus da imunodeficiência humana) e HPV (papilomavírus humano), respectivamente nas posições 22º e 43º do *corpus*, entre outros. Contudo, tal conjectura poderá somente ser confirmada com a expansão do *corpus*, que pretendemos ampliar para a marca de 500 mil palavras.

5 Considerações finais

Embora possamos observar um aumento gradual no número de pesquisas acadêmicas com textos médicos na área dos Estudos da Tradução, ainda é inexpressiva a quantidade de pesquisas que tenham como foco a tradução de abreviaturas e acrônimos. A partir da prática tradutória, tivemos a oportunidade de identificar esses elementos como um problema de tradução, visto não haver padronização na maneira de traduzi-los. Como o conhecimento científico é normalmente divulgado em língua inglesa, vem se tornando cada vez mais necessário que pesquisadores falantes de outras línguas, incluindo a portuguesa, submetam artigos nesse idioma para aumentar as chances de publicação. Com base nesse cenário, deu-se início à pesquisa aqui descrita, a fim de verificar padrões tradutórios e de uso desses elementos através da compilação de um *corpus* paralelo e dois *corpora* comparáveis.

Ainda que haja certa disposição em manter abreviaturas nas formas usadas e consolidadas em língua inglesa (AIXELÁ, 2009), observamos que um número considerável desses elementos são usados em suas respectivas línguas. Ainda que estejamos nas primeiras etapas do estudo, esses resultados já nos encorajam a nos perguntar se a tradução ou manutenção dessas abreviaturas e acrônimos estaria relacionada a algum elemento em particular, uma vez que observamos que há um número expressivo de abreviaturas e acrônimos traduzidos que se referem a nomes de doenças, por exemplo. Porém, estaremos em melhor posição para responder a essa e a outras perguntas que possam surgir ao longo do estudo quando completarmos a coleta e análise dos textos.

Uma vez concluído o processamento dos *corpora*, nosso objetivo é criar um glossário *on-line* nos pares linguísticos português-inglês e inglês-português para que seja disponibilizado a todos os interessados na tradução de abreviaturas e acrônimos em reumatologia. Não se objetiva criar um glossário definitório, visto que o problema detectado está sobretudo na questão de saber se esses elementos são ou não traduzidos, assim, ele será composto das abreviaturas e acrônimos seguidos das formas por extenso. Uma vantagem adicional de se criar um glossário desse tipo é ter esses elementos convenientemente em um único *site*, uma vez que a procura em diferentes fontes consome considerável tempo do tradutor. Finalmente, uma vez que todo o processo esteja concluído, nosso intuito é que que os *corpora* da subárea da reumatologia acompanhem os *corpora* existentes na plataforma do Termisul[24] da Universidade Federal do Rio Grande do Sul. Os *corpora* de outras subáreas da saúde (cardiologia, pediatria, enfermagem) e de química do Termisul já se encontram disponíveis para consulta.

[24] Disponível em: <http://www.ufrgs.br/termisul/>.

Referências

AZENHA, J. Jr. *Tradução Técnica e Condicionantes Culturais*. São Paulo: Humanitas, 1999.

_____. Tradução técnica, condicionantes culturais e os limites da responsabilidade do tradutor. *Cadernos de Tradução*, Florianópolis, v. 1, n. 1, p. 137-148, 1996.

BAKER, M. *Corpora* in Translation Studies. An overview and suggestions for future research. *Target*, 7(2), p. 223-243, 1995.

_____. *Corpus* Linguistics and Translation Studies: implications and applications. In: BAKER, M.; FRANCIS, G.; TOGNINI-BONELLI, E. (Org.). *Text and Technology*: in honour of John Sinclair. Amsterdam: John Benjamins, 1993, p. 233-250.

BAKER, P. Sketching Muslims: a *corpus* driven analysis of representations around the word 'Muslim' in the British press 1998–2009. *Applied Linguistics*, v. 34, n. 3, p. 255-278, 2013.

BELDA, J. R. M. Translating Computer Abbreviations from English into Spanish: main types and problems. *Meta: Translators' Journal*, v. 49, n. 4, p. 920-929, 2004.

BERBER SARDINHA, A. P. *Linguística de* Corpus. Barueri: Manola, 2004.

_____. *Corpora* Eletrônicos na Pesquisa em Tradução. Florianópolis: *Cadernos de Tradução*, v. 9, n. 1, p.15-59, 2002.

BIBER, D. Representativeness in *corpus* design. *Literary and Linguistic Computing*, v. 5, n. 4, p.243-257, 1993.

COLLET, Thaís. *Procedimentos tradutórios na legendagem de house: análise da terminologia médica referente a exames e aparelhos*. Dissertação (mestrado em Estudos da Tradução). Programa de Pós-Graduação em Estudos da Tradução, Universidade Federal de Santa Catarina, Florianópolis, 2012.

COULTHARD, R. J. *The application of* corpus *methodology to translation*: the JPED parallel *corpus* and the pediatrics comparable *corpus*. Dissertação (mestrado). Programa de Pós-Graduação em Estudos da Tradução, Centro de Comunicação e Expressão, Universidade Federal de Santa Catarina, Florianópolis, 2005.

FRANCO AIXELÁ, J. An overview of interference in scientific and technical translation. *The Journal of Specialised Translation*, n. 11, p. 75-88, 2009.

_____. The Study of Technical and Scientific Translation: An Examination of its Historical Development. *JoSTrans - The Journal of Specialised Translation*, v. 1, n. 1, 2004.

MUNDAY, J. (Org.). *The Routledge Companion to Translation Studies*. Revised Edition. Oxon: Routledge, 2009.

OLOHAN, M. *Introducing* Corpora *in Translation Studies.* London: Routledge, 2004.

PAIVA, P. T. P. *Uma investigação de traduções de textos da área médica sob a luz dos estudos da tradução baseados em* corpus. Tese (doutorado). Instituto de Biociências, Letras e Ciências Exatas, UNESP, São José do Rio Preto, 2009.

PASQUALI, A. B.; PINTO, P. T. A tradução de resumos médicos como meio de aprendizagem do processo tradutório e da terminologia especializada. *Caminhos em Linguística Aplicada*, v. 9, n. 2, p. 25-49, 2013.

SINCLAIR, J. *Corpus* evidence in language description. In: WICHMANN, A. S.; FLIGELSTONE, S.; MCENERY, T.; KNOWLES, G. (Ed.). *Teaching and language* corpora. London: Longman, 1997, p. 27-39.

TYMOCZKO, M. Computerized *Corpora* and the Future of Translation Studies. *Meta*, v. 43, n. 4, p. 652-660, 1998.

Identificação de termos no discurso literário de fantasia da série Harry Potter em uma abordagem direcionada por *corpus*

Term identification in the fantasy literary discourse of the Harry Potter series: a *corpus*-driven approach

Raphael Marco Oliveira Carneiro

Resumo: Este trabalho relata os resultados de uma investigação direcionada por *corpus* que objetivou descrever o uso de unidades lexicais ficcionais como termos dentro do universo de discurso literário de fantasia. Tal *corpus* é composto pelas sete obras da série literária Harry Potter e outros três volumes que detalham o mesmo mundo ficcional criado por J. K. Rowling. O programa Word Smith Tools 6.0 e suas três ferramentas principais, Concord, *keywords* e Wordlist, possibilitaram, respectivamente, a recuperação de contextos linguísticos para a identificação de enunciados definitórios, a identificação de candidatos a termos e o levantamento de dados estatísticos. A etiquetagem de itálicos também contribuiu para a identificação de termos. O estudo contribui para o reconhecimento do potencial conceptual de termos ficcionais e do valor semântico cultural desses termos. Além disso, revela como manifestações literárias contemporâneas fazem uso de elementos folclóricos e, ao mesmo tempo, fundam uma nova realidade imaginativa. O trabalho demonstra que a adoção do conceito de universo de discurso pode contribuir para a ampliação do escopo dos estudos terminológicos.

Palavras-chave: Terminologia. Linguística de *Corpus*. Discurso literário de fantasia. Harry Potter.

Raphael Marco Oliveira Carneiro – Mestre em Linguística e Linguística Aplicada pelo Programa de Pós-Graduação em Estudos Linguísticos (PPGEL) do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU) – raphael.olic@gmail.com.

Abstract: This paper reports on the results of a *corpus*-driven investigation whose aim was to describe the use of fictional lexical units as terms within the universe of fantasy literary discourse. Said *corpus* comprises the seven books of the Harry Potter literary series and other three volumes that are part of the same fictional world created by J. K. Rowling. The software Word Smith Tools 6.0 and its three main tools, Concord, *keywords* and Wordlist, enabled the retrieval of linguistic contexts for the identification of defining utterances, the identification of term candidates and the gathering of statistical data. The tagging of italics has also contributed to term identification. The study contributes to the acknowledgement of the conceptual potential of fictional terms and of their cultural semantic value; to the understanding of how contemporary literary manifestations make use of folkloric elements and, at the same time, engender a new imaginative reality; and points out that by adopting the concept of universe of discourse the scope of terminology studies can be expanded.

Keywords: Terminology. *Corpus* Linguistics. Fantasy literary discourse. Harry Potter.

1 Introdução

Considerando recentes desenvolvimentos dos estudos terminológicos (ESPERANDIO, 2015; CARNEIRO, 2016) em relação à investigação de terminologias no universo de discursos com baixo de cientificidade e tecnicidade, este trabalho pretende contribuir para a reflexão sobre o estatuto terminológico de unidades lexicais em uso no universo de discurso literário de fantasia. A propósito, Barros (2006, p. 26) constata que "[...] as investigações científicas sobre o léxico de obras literárias têm observado a presença marcante, nessas obras, de termos pertencentes a campos temáticos e especializados". Nesse seu artigo de revisão sobre aspectos epistemológicos e perspectivas científicas da Terminologia, Barros (2006) reconhece que a área caminha em direção ao estabelecimento de relações de cooperação com a literatura, apontando essa relação como uma interface emergente nos estudos terminológicos.

Por sua vez, as pesquisas de Barbosa (2006, 2007) sobre o universo de discurso etnoliterário postulam os fundamentos teóricos que caracterizam as unidades lexicais em uso nesse universo. Segundo Barbosa (2007, p. 434), "essas unidades lexicais têm sememas muito especializados, construídos com semas específicos do universo de discurso em causa, provenientes das narrativas, cristalizados, tornando-se verdadeiros símbolos dos temas envolvidos". Nesse sentido, Carneiro (2016), com vistas a contribuir para o reconhecimento das funções de usos terminológicos no universo de discurso literário de fantasia dentro da temática *Witchcraft and Wizardry* (Magia e Bruxaria), dadas as suas relações intertextuais e interdiscursivas com discursos etnoliterários, conduziu uma investigação direcionada por um *corpus* em inglês britânico composto pelas sete obras da série Harry Potter e pelos outros três volumes que detalham o mesmo mundo ficcional criado por J. K. Rowling[1].

[1] Título das obras que compõem o *corpus* de estudo: *Harry Potter and the Philosopher's Stone, Harry Potter and the Chamber of Secrets, Harry Potter and the Prisoner of Azkaban, Harry Potter and the*

A escolha do *corpus* se deu em vista da proeminência e influência que a série Harry Potter obteve na virada do milênio, de 1997-2007, constituindo um marco na história da literatura inglesa.

Dessa maneira, neste trabalho, organizado nas seções de referencial teórico, procedimentos metodológicos, recortes observacionais e considerações finais, relatamos parte dessa pesquisa com enfoque nos procedimentos metodológicos e nos critérios que contribuíram para a identificação de termos[2] em nosso *corpus* de estudo.

2 Referencial teórico

É de conhecimento que a identidade epistemológica da Terminologia se consolidou a partir de domínios científicos e tecnológicos com forte orientação onomasiológica e normalizadora das denominações de conceitos desses domínios. A partir da Teoria Comunicativa da Terminologia (TCT), contudo, houve um redirecionamento do pensamento em relação à constituição de terminologias que passaram a ser reconhecidas em sua poliedricidade, ou seja, em função de suas faces linguística, cognitiva e comunicativa. Desde então, surgiram diversas perspectivas que passaram a enfocar o fenômeno terminológico de um ponto de vista específico; são elas: Teoria Sociocognitiva da Terminologia (TST), Socioterminologia (ST), Terminologia Cultural (TC) e Terminologia Textual (TT). A Etnoterminologia, conforme proposta de Barbosa (2007), integra o conjunto dessas perspectivas que têm dinamizado o campo dos estudos terminológicos, com foco em discursos com baixo grau de cientificidade e tecnicidade e discursos etnoliterários, como fábula, folclore, lendas, mito, literatura de cordel e literatura popular.

Na perspectiva da Etnoterminologia, consideram-se universos de discurso[3] como *locus* da ativação do estatuto terminológico, em vez de domínios especializados e científicos. Entende-se por universo de discurso um

> [...] conjunto de discursos manifestados e manifestáveis, que tendem *ad infinitum*, reunidos por *critérios de equivalência*, ou seja, caracterizados por *constantes e coerções*, suscetíveis de configurar uma *norma discursiva frástica* e *transfrástica*, discursos que

Goblet of Fire, Harry Potter and the Order of the Phoenix, Harry Potter and the Half-Blood Prince, Harry Potter and the Deathly Hallows, Fantastic Beasts and Where to Find Them, Quidditch Through the Ages, The Tales of Beedle the Bard. Para simplificar as referências às obras neste texto utilizamos a seguinte notação respectivamente: HP 1, HP 2, HP 3, HP 4, HP 5, HP 6, HP 7, FB, QA, TB.

[2] Noção central para os estudos em Terminologia em referência às unidades lexicais temáticas de domínios especializados do conhecimento humano que será definida no contexto deste trabalho na seção 4.

[3] Tomamos a noção de discurso como processo de produção que implica uma enunciação de codificação e de decodificação, bem como um enunciado.

> mantêm entre si redes de relações intertextuais e interdiscursivas, inseridos num contexto linguístico e sociocultural, pertencentes à *macrossemiótica* de uma cultura. (PAIS; BARBOSA, 2004, p. 81)

Assim, Barbosa (2007, p. 439) entende que "toda classificação resulta dos entornos discursivos e dos condicionamentos das normas discursivas, dependente, portanto, dos universos de discurso e das situações de discurso". É em um eixo *continuum* de especialização que se deve situar o problema do estatuto de unidades lexicais, o que resulta na identificação de unidades mais ou menos especializadas. Em resumo, as unidades lexicais estabelecem relações de dependência com universos de discurso, configurando-se em vocábulo ou termo em função do ambiente textual-discursivo.

Ao considerar o universo de discurso literário de fantasia, observam-se certas características constantes que nos permitem agrupar discursos-ocorrência nessa classe de discursos. Uma dessas constantes são as relações de intertextualidade e interdiscursividade estabelecidas com discursos etnoliterários, como o folclore, assim como com outros discursos-ocorrência da própria fantasia. Hunt (2010) aponta que o reconto de mitos e lendas, elementos do folclore de um povo, é pouquíssimo encontrado fora da literatura infantil, segmento com o qual o discurso literário de fantasia geralmente se alinha. Outra constante que nos permite identificar a fantasia literária como um universo de discurso é a criação de mundos ficcionais[4], que não se limitam a supostas correspondências com o mundo real. Um mundo ficcional é definido como "um mundo possível construído por um texto ficcional ou outro meio performativo semiótico"[5] (DOLEŽEL, 1998, p. 280). A serialização, ou seja, apresentação da narrativa em múltiplos volumes, também é uma marca caracterizadora desse universo. Como exemplos, podemos citar as séries *The Lord of the Rings*, *The Chronicles of Narnia*, *Discworld*, dentre muitas outras (cf. CARNEIRO, 2016 para uma lista de séries literárias de fantasia). Assim, textos específicos do universo de discurso literário de fantasia tendem a engendrar detalhados mundos ficcionais que se particularizam devido a um estado possível de ações e conceitos próprios expressos e lexemizados[6] em unidades lexicais.

---

[4] Com base em Doležel (1998), define-se 'mundo' como a totalidade de entidades materiais e mentais que pode ser designada por meios linguísticos ou outros meios semióticos; "mundo real" como um mundo possível realizado que é percebido pelos sentidos humanos e fornece o palco para a atuação humana; "mundo possível" como um mundo que é pensável.

[5] No original: "*A possible world constructed by a fictional text or other performative semiotic medium*".

[6] "A lexemização [...] é aqui entendida como 'la mise em lexème', a configuração do fato virtual em grandeza-signo, no próprio ato de instaurar a significação. Trata-se de processo complexo que é precedido pelo de conceptualização. Este, por sua vez, parte do *continuum amorfo* (HJELMSLEV, 1968, p. 71-85). Nessa fase, os 'fatos' constituem substâncias estruturáveis, que são apreendidos e recortados pelos grupos linguísticos e socioculturais, de diferentes maneiras, e que mantêm um núcleo de percepção biológica universal" (BARBOSA, 2001, p. 43).

Doležel (1998) concebe a ficcionalidade como um fenômeno semântico no eixo *representação (signo) – mundo*. Assim, o paradigma de mundos ficcionais nos permite compreender que a base semântica de unidades lexicais em ambientes literários se sustenta na oposição do eixo semântico "mundo real" e "mundo ficcional". Em suas palavras, "[...] em uma direção, ao construir mundos ficcionais, a imaginação poética trabalha com 'material' retirado da realidade; na direção oposta, construtos ficcionais influenciam profundamente nossa imaginação e entendimento da realidade"[7] (DOLEŽEL, 1998, p. X). De forma similar, Ryan (2014), que compartilha do mesmo modelo teórico de Doležel, afirma que há uma pluralidade de mundos, de modo que o mundo em que vivemos é chamado de real e é o único mundo com existência autônoma. Os outros são mundos possíveis não reais, criações da imaginação.

> Textos não ficcionais se referem ao mundo real, enquanto ficcionais criam mundos possíveis não reais. Nesse modelo, a distinção entre ficção e não ficção é uma questão de referência: a não ficção faz alegações verídicas sobre o mundo real, enquanto a ficção faz alegações verídicas sobre um mundo possível alternado. (RYAN, 2014, p. 6)

Há que se considerar também que, o "descompromisso" da fantasia com o mundo real frequentemente produz sememas desviantes da experiência humana biofísica do real. Quando lexemizados no discurso, esses sememas[8] geram unidades específicas de um universo de discurso. Ao levar em conta a função referencial das unidades lexicais ficcionais, observa-se que os seus referentes devem ser encontrados em um mundo ficcional, em um tipo de dêixis cognitiva, em que a projeção mental do mundo ficcional permite ao leitor localizar as entidades ou particulares ficcionais designados pelas unidades lexicais. Isto é, os referentes das unidades lexicais não estão no mundo real, mas tampouco são inexistentes; eles existem em um mundo ficcional. Em outras palavras, a semântica ficcional de Doležel (1998) admite e explica a função referencial de termos ficcionais com base no conceito de mundo ficcional, diferentemente de outras abordagens como a mimese, em que os particulares ficcionais seriam imitações de entidades do mundo real.

---

[7] No original: "*In one direction, in constructing fictional words, the poetic imagination works with 'material' drawn from actuality; in the opposite direction, fictional constructs deeply influence our imagining and understanding of reality*".

[8] "Os semas nucleares definem os traços invariáveis em um lexema, aqueles traços que justificam a especificidade de seu significado, de seu valor, que permanece constante independentemente do contexto de aparição. Os semas contextuais, por sua vez, são aqueles que dependem do contexto no qual o lexema é inserido e servem para declinar o significado invariável segundo as particulares acepções que aquele lexema pode, de vez em quando, assumir. O significado de um lexema depende sempre da combinação de ao menos um sema nuclear com pelo menos um sema contextual. É esta combinação, variável evidentemente a cada inserção do lexema em um texto dado, que toma o nome de semema. [...] O semema, como se vê, reúne em si feixes de semas que, combinando-se, justificam as significações específicas de cada ocorrência" (VOLLI, 2012, p. 70-71).

Cabe mencionar também que a linguagem literária, segundo Steger (1987), é dotada de uma motivação pragmática específica, ou seja, motivação para a criação sintética de um mundo ficcional. Dessa forma, seus modos de ação, objetos e delimitação ontológica circunscrevem-se em relação a tudo aquilo que se torna manifesto por meio de uma forma linguística, sendo que a validade de seus enunciados refere-se ao que é verdadeiro por meio da forma estética. Nessa concepção, a linguagem literária realiza a modelização de um mundo secundário, de modo a expressar uma visão de mundo particular.

Retornando à questão terminológica propriamente dita, Barbosa (2006), a propósito dos discursos etnoliterários, caracteriza as unidades lexicais com valor terminológico, nesse universo de discurso, com base na interface entre linguagem especializada e linguagem literária. Em suas palavras, essas unidades "[...] são quase-termos técnicos, pois pertencem a uma linguagem especial/especializada. Seus sememas não correspondem, pois, nem aos sememas da língua comum, nem aos sememas das linguagens dos domínios científicos" (BARBOSA, 2006, p. 50). Nessas condições, "[...] essas unidades lexicais reúnem qualidades das línguas especializadas e da linguagem literária [...]" (BARBOSA, 2006, p. 48). Portanto, as unidades lexicais, conforme o recorte da Etnoterminologia, apresentam uma constituição sincrética de aspectos especializados e literários, garantindo, ao mesmo tempo, funções na tessitura do texto literário e valor semântico sociocultural, na medida em que constituem documentos do processo histórico de uma cultura, particularmente no tocante ao imaginário coletivo.

Um exemplo comumente encontrado nos textos de Barbosa refere-se ao rito Bumba-meu-boi, do Maranhão. Nesse rito folclórico, a unidade lexical "boi" não designa o "boi" da biologia nem da agropecuária; ela representa uma entidade mítica, que é morta, mas ressuscita ao final da narrativa, sendo inclusive interpretada como a morte e ressurreição de Cristo. Esse exemplo ilustra como a unidade "boi", em princípio uma unidade lexical da língua comum, adquire uma significação especial e particular quando usada no contexto de um rito folclórico, no qual não se trata mais de um animal em termos biológicos, mas sim de uma criatura mítica. Mesmo sendo uma unidade lexical popularizada, ela mantém uma relação de exclusividade semântica em função de um rito folclórico, conservando assim um valor especializado pertinente a um universo de discurso etnoliterário. Em termos referenciais, o "boi" do rito Bumba-meu-boi não deve ser encontrado no mundo real, mesmo que concretamente representado, mas na projeção mental de um mundo ficcional caracterizado por um estado particular de coisas, em que a ressurreição de um boi é possível. Essa unidade não mantém o mesmo nível de especialização de um termo científico, mas sua especificidade semântica já lhe confere estatuto de termo dentro do universo de discurso etnoliterário.

Assim, a perspectiva que se observa na Etnoterminologia é a da transdisciplinaridade. O tratamento transdisciplinar de unidades lexicais exige a adoção de

uma lógica diferente. Nicolescu (1999, p. 29 apud SANTOS, 2008, p. 74) assim caracteriza a lógica clássica: "1. O axioma da identidade: A é A; 2. O axioma da não contradição: A não é não A; 3. O axioma do terceiro excluído: não há um termo T, que é, ao mesmo tempo, A e não A". Santos (2008, p. 75) explica que, "por esses axiomas, a lógica clássica admite um único nível de realidade, uma vez que o axioma número 3 exclui a possibilidade de articulação". A transdisciplinaridade, ao contrário, permite a articulação, e reconhece a possibilidade de um *terceiro termo incluído*, de modo que A pode ser não A; "[...] transdisciplinaridade significa transgredir a lógica da não contradição, articulando os contrários [...]" (SANTOS, 2008, p. 75).

Considerando que "vocábulo" é o contrário de "termo" na relação língua comum x linguagem especializada, respectivamente, ao aplicarmos os axiomas da lógica clássica à problemática do estatuto de unidades lexicais, temos que: "termo" é "termo"; "termo" não é "vocábulo"; *não há um termo T, que é, ao mesmo tempo, "termo" e "vocábulo"*. Em outras palavras, na lógica clássica, há uma separação total e nítida entre essas duas instâncias, de modo que os termos e as linguagens especializadas são tomados como independentes dos vocábulos e da língua comum, como água e óleo, elementos que não se misturam, perspectiva essa sob a qual opera a TGT. Contudo, ao aplicarmos a lógica transdisciplinar obtemos o seguinte: "termo" é "termo"; "termo" não é "vocábulo"; *há um termo T, que é, ao mesmo tempo, "termo" e "vocábulo"*. Nessa lógica, aceitam-se as articulações entre língua comum e linguagens especializadas, inclusive entre linguagem literária; a dinâmica dos movimentos entre vocábulos e termos; e a hibridização dessas duas funções em uma, vocábulos-termos, ou unidades multifuncionais.

Nesse breve apanhado teórico, buscamos delinear o campo de atuação e o objeto de estudo da Etnoterminologia e as contribuições da semântica de mundos ficcionais para essa abordagem. Na seção seguinte, apresentamos os procedimentos metodológicos mais relevantes para a identificação de termos no universo de discurso literário de fantasia.

3 Procedimentos metodológicos

Para a identificação dos termos da temática *Witchcraft and Wizardry* presente na série Harry Potter, utilizamos procedimentos usuais da Linguística de *Corpus* (LC) que não serão descritos em detalhes aqui (cf. CARNEIRO, 2016 para detalhes). Basta dizer que fizemos uso do programa Word Smith Tools 6.0 (SCOTT, 2012) e suas três ferramentas principais, Concord, Keywords e Wordlist, para obter, respectivamente, os contextos linguísticos de uso dos termos, identificar os candidatos a termos e quantificar os dados do *corpus*. Daremos enfoque ao procedimento da metodologia que, por meio da etiquetagem de itálicos, automatizou

a identificação de termos grafados dessa forma. Após a compilação do *corpus* e conversão de arquivos em *.docx realizamos a etiquetagem de itálicos por meio do programa Microsoft Word. Em sequência, apresentamos os critérios e as características do *corpus* de estudo:

Quadro 1– Tipologia do *corpus* de estudo

Critérios	Características
Língua	Monolíngue (inglês)
Modo	Escrito (narrativas)
Tempo	Sincrônico (textos de 1997-2008) Contemporâneo
Seleção	Amostragem (amostra de textos literários ficcionais) Estático (seleção não renovável)
Conteúdo	Especializado (textos de uma série de ficção literária de fantasia infantojuvenil)
Autoria	Apenas um autor; língua nativa (inglês britânico)
Tamanho	1.156.126 itens (médio-grande)[9]
Nível de Codificação	Uso de cabeçalhos; etiquetado (itálicos)
Uso na pesquisa	Estudo (análise e descrição linguística)

Fonte: CARNEIRO, 2016, p. 132

A seguir apresentamos dados numéricos obtidos pela ferramenta Wordlist, bem como outros dados que caracterizam o *corpus* de estudo:

[9] De acordo com a classificação de Berber Sardinha (2004, p. 26), os *corpora* são assim classificados conforme suas extensões: pequeno (menos de 80 mil palavras); pequeno-médio (80 a 250 mil palavras); médio (250 mil a 1 milhão de palavras); médio-grande (1 milhão a 10 milhões de palavras); grande (10 milhões ou mais de palavras).

Tabela 1 – Características dos livros que compõem o *corpus* de estudo

Livros	Ano de publicação (edição do *corpus*)	Páginas	*Tokens* (itens)	*Types* (formas)	Itálicos
HP 1	1997 (2004)	223	80.190	5.838	405
HP 2	1998 (2004)	251	87.964	6.995	661
HP 3	1999 (2004)	317	110.895	7.562	458
HP 4	2000 (2004)	636	196.277	10.466	919
HP 5	2003 (2003)	766	264.251	12.481	1.297
HP 6	2005 (2005)	607	174.171	10.438	603
HP 7	2007 (2007)	607	203.795	11.245	805
FB	2001 (2001)	64	14.416	3.266	106
QA	2001 (2001)	63	11.504	2.721	60
TB	2008 (2008)	126	12.663	2.881	33
Total[10]	1997-2008 (2001-2008)	3.660	1.156.126	23.841	5.384

Fonte: CARNEIRO, 2016, p. 132

As ocorrências em itálico (5.384) foram etiquetadas para automatizar a identificação de termos que são grafados dessa forma. Na opção "localizar" do programa Word, escolhemos o "formato: itálico" e inserimos o código " ^& " na opção substituir, em que "" é a etiqueta de abertura que indica "início do segmento textual grafado em itálico", "^&" é o símbolo que indica "qualquer expressão incluindo espaços" e "" é a etiqueta de fechamento que indica "término do segmento textual grafado em itálico". A Figura 1 ilustra o procedimento de etiquetagem.

[10] Para os cálculos, determinamos que palavras hifenizadas correspondem a um item. Os valores totais das ocorrências de itens, formas e itálicos devem ser vistos como aproximados. Mesmo com o cuidado de fazer correções, não podemos ter certeza de que o *corpus* está isento de inadequações ortográficas geradas durante o processo de conversão de *.pdf para *.txt, o que altera os valores computados pelo WST. Por isso, são valores relativos e não absolutos.

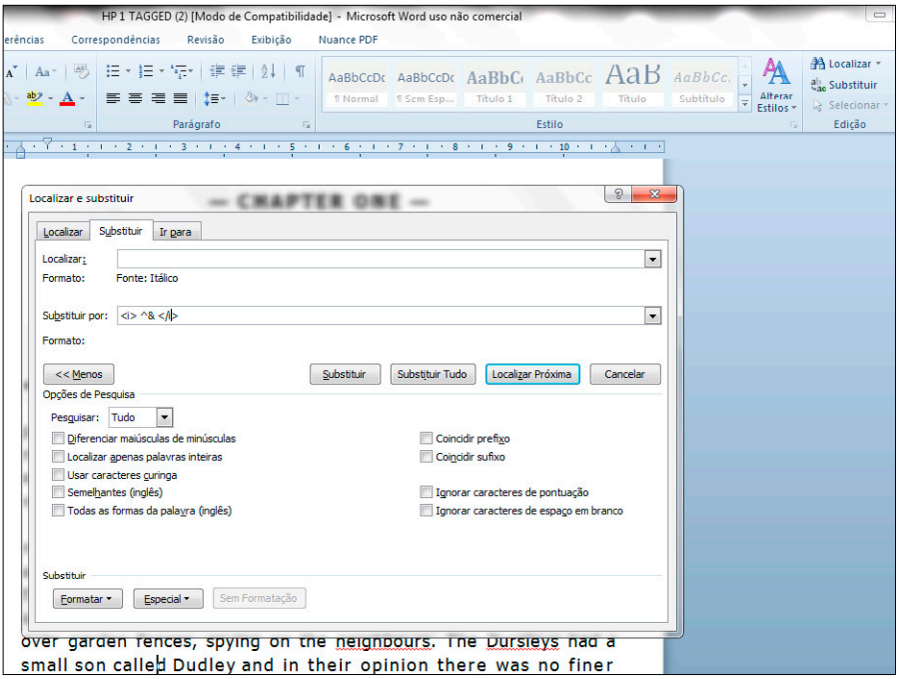


Figura 1 – Etiquetagem de itálicos
Fonte: CARNEIRO, 2016, p. 138

Uma observação rápida das ocorrências em itálico na Figura 2 revela que o itálico é um recurso gráfico usado com funções diversas nos textos do *corpus*, que não cabem ser analisados aqui. Assim, as concordâncias feitas por meio das etiquetas foram realizadas para identificarmos os termos particularizados pelo uso do itálico.

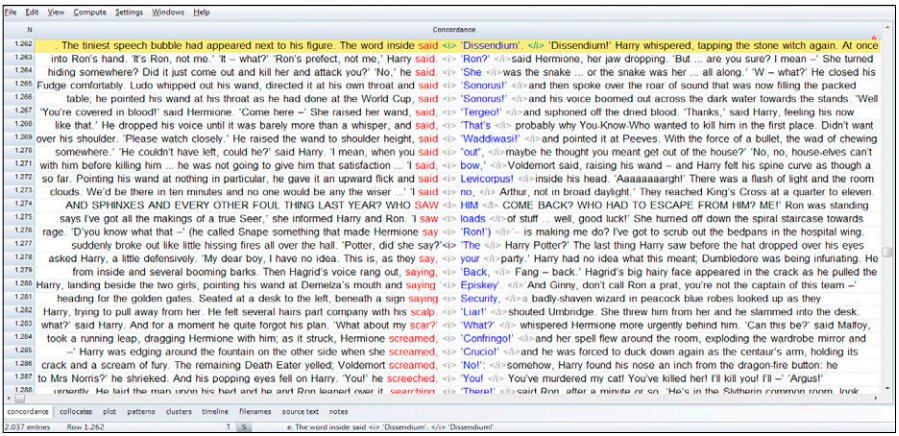


Figura 2 – Lista de linhas de concordâncias automatizadas por meio da busca pela etiqueta *
Fonte: CARNEIRO, 2016, p. 148

Na Figura 2, detectamos os termos *Dissendium*, *Sonorus, Tergeo, Waddiwasi, Levicorpus, Episkey, Confringo* e *Crucio*, todos grafados em itálico e com iniciais maiúsculas, traços distintivos da diferenciação de uso dessas unidades lexicais. Todos eles podem ser classificados como "encantamentos" utilizados para lançar feitiços. Observe-se que esses termos são geralmente usados após verbos de elocução como *said* e *screamed*, e indicam a força ilocucionária dos encantamentos que visam a provocar uma ação ou mudança.

Além dessa identificação de candidatos a termos por meio da etiquetagem de itálicos utilizamos o critério de pertinência temática do termo, com frequência em menor escala. Assim, mesmo um termo que se configure como *hapax legomenon* (apenas uma ocorrência no *corpus*), pode ser incluído na estruturação conceptual do universo de discurso caso seja tematicamente pertinente. Entendemos que "pertinência temática [...] significa a propriedade de um termo pertencer a uma terminologia *stricto sensu* pelo fato de vincular-se a um conceito que faz parte do campo cognitivo do domínio inventariado" (KRIEGER; FINATTO, 2004, p. 138). Assim, dado o domínio semântico-conceptual estabelecido pelo campo *Witchcraft and Wizardry*, selecionamos os termos semanticamente vinculados a ele.

Em sequência, para ilustrar, apresentamos, na Figura 3, um recorte da hierarquia conceptual elaborada com base no *corpus* dos termos-conceitos inventariados dentro da temática *Witchcraft and Wizardry.*

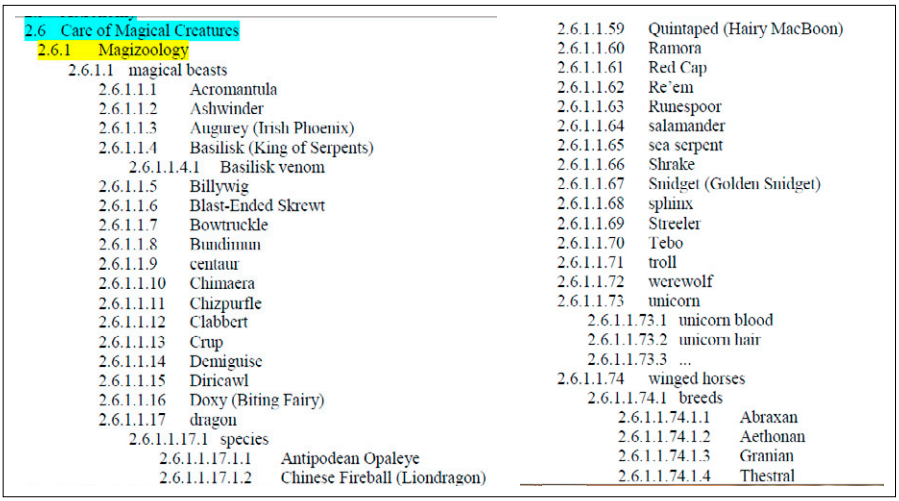


Figura 3 – Recorte da hierarquia conceptual
Fonte: CARNEIRO, 2016, p. 151-152

No recorte da hierarquia conceptual anteriormente apresentado, encontramos o campo *Magizoology*, que se refere ao campo de estudos de criaturas mágicas. É possível observar, nesse conjunto de termos, criaturas presentes em manifestações etnoliterárias e outras manifestações do universo de discurso literário de fantasia, como *Chimaera*, *centaur*, *dragon*, *sphinx*, *troll*, *werewolf*, *unicorn*. Esses termos, também presentes em outras manifestações discursivas, atestam a interdiscursividade no interior do universo de discurso literário de fantasia. São termos que já fazem parte do imaginário coletivo. Ao lado deles, encontramos criações lexicais exclusivas do mundo ficcional de Harry Potter, como *Billywig*, *Blast-Ended Skrewt*, *Bowtruckle, Demiguise, Thestral,* dentre outros. Esses termos neológicos fundam uma nova dimensão imaginária, de modo que criaturas já consagradas na mitologia e no folclore são ressignificadas com acréscimos de semas e convivem em um mesmo mundo ficcional com animais mágicos cujos conceitos são engendrados a partir do mundo ficcional de Harry Potter.

Após esse breve percurso dos procedimentos metodológicos que possibilitaram a identificação e recolha de termos a comporem a área temática do *corpus*, passaremos a comentar alguns exemplos de uso de unidades lexicais ficcionais em relação à presença de contextos linguísticos definitórios e ao engendramento do conceito dessas unidades.

4 Recortes observacionais

Nesta seção, apresentamos e comentamos alguns resultados que corroboram a identificação de termos no *corpus* analisado.

Principiamos com o cenário comunicativo que se estabelece nas obras contidas no *corpus*, simplificadamente caracterizado como a comunicação entre um enunciador que conhece um mundo ficcional e um enunciatário que o desconhece, mas que é parcialmente familiarizado com esse mundo devido às suas experiências com outros discursos do mesmo universo de discurso. Tal cenário é estabelecido de forma mais explícita em TB, em que, após a introdução da autora, consta a seguinte nota sobre as notas de rodapé:

> **A Note on the Footnotes**
> Professor Dumbledore appears to have been writing for a wizarding audience, so I have occasionally inserted an explanation of a term or fact that might need clarification for Muggle readers. JKR

Consideramos essa nota como uma indicação do cenário comunicativo em que a obra se insere. Uma vez que Dumbledore estaria escrevendo para o público bruxo, há a necessidade de que a autora esclareça certos termos e fatos para o público não bruxo. Em outras palavras, assume-se que o leitor não bruxo não tem familiaridade com a terminologia que constitui o universo de discurso bruxo, surgindo a necessidade de explicações. Tais explicações, em grande medida, são feitas em enunciados definitórios, apresentados a seguir, que particularizam e circunscrevem termos como *Squib*, *warlock*, *wizard*, *Necromancy* e *Inferi* dentro do domínio semântico-conceptual ativado no universo de discurso de Harry Potter:

> 2 [A Squib is a person born to magical parents, but who has no magical powers. Such an occurrence is rare. Muggle-born witches and wizards are much more common. JKR]

> 2 [The term "warlock" is a very old one. Although it is sometimes used as interchangeable with "wizard", it originally denoted one learned in duelling and all martial magic. It was also given as a title to wizards who had performed feats of bravery, rather as Muggles were sometimes knighted for acts of valour. By calling the young wizard in this story a warlock, Beedle indicates that he has already been recognised as especially skilful at offensive magic. These days wizards use "warlock" in one of two ways: to describe a wizard of unusually fierce appearance, or as a title denoting particular skill or achievement. Thus, Dumbledore himself was Chief Warlock of the Wizengamot. JKR]

> 1 [Necromancy is the Dark Art of raising the dead. It is a branch of magic that has never worked, as this story makes clear. JKR]

> 4 [Inferi are corpses reanimated by Dark Magic. JKR]

Trata-se de notas autorais assinadas com as iniciais da autora, supostamente inseridas posteriormente aos textos do Professor Dumbledore. As notas para *Squib*, *Necromancy* e *Inferi* possuem elementos de definições formais simples, de acordo com Pearson (1998). As três notas são iniciadas com o termo, que é então definido em um gênero próximo (*person, Dark Art* e *corpses,* respectivamente) e uma característica específica (*born to magical parents, but who has no magical powers; of raising the dead; reanimated by Dark Magic*). As características específicas são introduzidas com o uso de particípio, preposição e particípio, respectivamente. A nota para *warlock* apresenta mais traços de uma definição enciclopédica, devido à sua configuração mais extensa e explicativa.

A partir da análise realizada, podemos afirmar que as notas de rodapé constituem um elemento do peritexto literário que fornece explicações pertinentes às especificidades do mundo ficcional ao qual as obras fazem referência. Enquanto discurso manifestado, as notas são marcas caracterizadoras da ficcionalidade das obras, ao chamarem atenção para termos específicos do mundo ficcional. As notas nos permitem afirmar o valor terminológico assumido por certas unidades lexicais ficcionais ao serem explicitamente definidas. Se elas são assim definidas em virtude do cenário comunicativo é porque há algo de especial sobre elas, caso contrário esse fazer teria sido desnecessário.

Outro termo usado em um contexto linguístico explicitamente definitório é o termo *Horcrux*. No contexto apresentado a seguir, há, inclusive, referências metalinguísticas (*understand the term* e *is the word*) ao termo, o que revela um fazer denominativo explícito do conceito de *Horcux.*

> "Well," said Slughorn, not looking at Riddle, but fiddling with the ribbon on top of his box of crystallised pineapple, "well, it can't hurt to give you an overview, of course. Just so that you understand the term. A **Horcrux** is the word used for an object in which a person has concealed part of their soul." (grifo nosso)

Após análise semântico-conceptual, realizada por meio dos contextos linguísticos de ocorrência para *Horcrux,* chegamos à seguinte representação proposicional da definição: {[Horcrux] (to be) [object] [animal] [person] + (used for) [concealing] [soul]}, em português {[Horcrux] (ser) [objeto] [animal] [pessoa] + (servir para) [esconder] [alma]}. A definição pode então ser formulada como *object, animal or person that conceals part of a person's soul*.

Apesar de o termo *Horcrux* ser uma unidade lexical neológica (não há ocorrências anteriores à série HP conforme atestado por meio do Google Books Ngram Viewer, na Figura 4), o conceito por trás da denominação não é um conceito novo dentro do universo de discurso literário de fantasia e dos discursos etnoliterários. Kronzek e Kronzek (2010, p. 130-131) afirmam que, "apesar de a palavra *Horcrux* ser exclusiva do mundo mágico de Harry, a ideia por trás dela – de que a alma, ou pedaços dela, podem ser guardados e protegidos em objetos materiais – faz

parte de histórias folclóricas, mitos e práticas ao redor do mundo"[11] (KRONZEK; KRONZEK, 2010, p. 130-131).

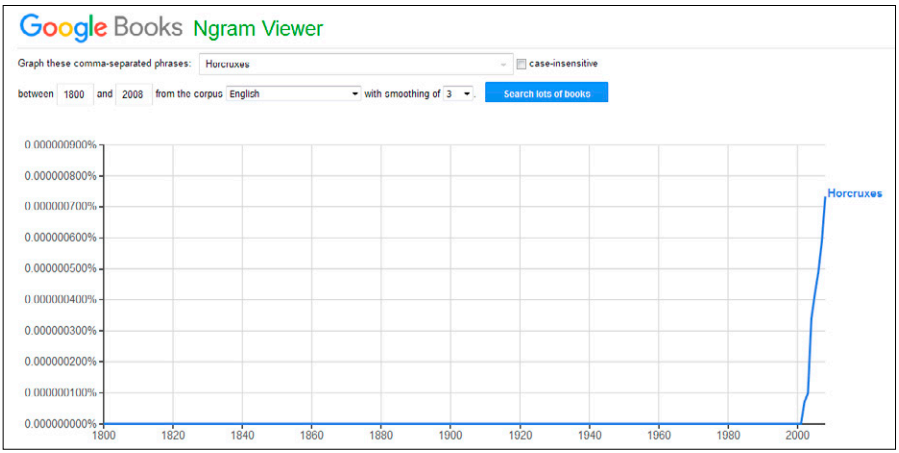


Figura 4 – Visualização do n-grama *Horcruxes*
Fonte: Google Books Ngram Viewer[12]

Kronzek e Kronzek (2010, p. 131-132) registraram noção semelhante ao termo *Horcrux* no conto folclórico russo "Koschei the Deathless", em um conto das *Arabian Nights* e no *role-playing game Dungeons and Dragons*. O conceito de *Horcrux* também foi identificado fora da ficção. Os mesmos autores encontraram esse conceito em sociedades tribais da Sibéria e da América do Sul, nas quais pessoas doentes são tratadas por xamãs por meio da transferência de suas almas para uma bolsa medicinal até que os pacientes se recuperem. Em certas culturas, também, acredita-se que o pedaço da alma de uma pessoa pode ser acidentalmente aprisionado em um objeto, como para os artesãos Navajo. Também há a crença, na prática de vodu no Haiti, de que uma pessoa pode ser escravizada ao ter a alma aprisionada em uma garrafa; o uso dessas "garrafas de espírito", por sua vez, remonta ao Congo Africano. Acrescentamos a essas identificações o fato de que na trilogia *O Senhor dos Anéis*, a sobrevivência da personagem Sauron está ligada a um objeto material, o Um Anel. Sua força vital depende da sobrevivência do Anel. Quando o Anel é destruído por Frodo na Montanha da Perdição, ele é acometido a Sauron, e seu reino maléfico é esfacelado.

[11] No original: "*Although the word Horcrux is unique to Harry's wizarding world, the idea behind it – that the soul, or pieces of it, can be stored and protected in material objects – is part of folk tales, myths, and practices from around the world*".
[12] Disponível em: <https://books.google.com/ngrams/graph?content=Horcruxes&year_start=1800&year_end=2008&corpus=15&smoothing=3&share=&direct_url=t1%3B%2CHorcruxes%3B%2Cc0#t1%3B%2CHorcruxes%3B%2Cc0-1>. Acesso em: 25 maio 2016.

Os exemplos citados revelam intersecções conceptuais entre diferentes manifestações discursivas, tanto etnoliterárias quanto no universo de discurso da literatura de fantasia. Atestamos, por meio desses exemplos, a interdiscursividade do universo de discurso literário de fantasia com outras manifestações discursivas, sugerindo um provável conjunto arquiconceptual (conjunto intersecção entre conceitos).

É possível recuperar também investimentos semânticos que contribuem para a axiologização[13] das unidades lexicais. O termo *Horcrux*, por exemplo, é axiologizado negativamente ao longo do texto narrativo. A atribuição de semas disfóricos como [*evil*] e [*banned subject*] ao termo atribui um juízo de valor negativo. Nos movimentos discursivos do enunciador, a prática de criação de *Horcruxes* não é encorajada, visto que são objetos amaldiçoados produzidos pelas artes das trevas, relacionados ao mal e à infelicidade. O objetivo da criação de *Horcruxes*, de se tornar imortal, portanto, é condenado no todo do *corpus*. Outro termo que também é axiologizado negativamente no *corpus* é *Avada Kedavra*, que provoca morte instantânea àqueles que são atacados por essa maldição. É classificada como uma maldição imperdoável (*Unforgivable Curses*), o que já determina um caso claro de axiologização negativa. Em outras palavras, o ato de matar alguém é um comportamento condenado na narrativa. Esses exemplos demonstram como os termos podem receber investimentos axiológicos e, então, configurarem sistemas de valores que, segundo Barbosa (2007, p. 444), "[...] determinam pensamentos e comportamentos, [...] formas de ver o mundo, [...] maneiras de agir recomendáveis ou condenáveis, no fazer social".

Outro fato importante a ser destacado é que unidades lexicais identificadas com valor terminológico no universo de discurso literário de fantasia da série Harry Potter passaram a integrar denominações científicas de espécies zoológicas e botânicas no mundo real. Citamos como exemplo, o termo *Thestral,* que é usado em Harry Potter para designar um tipo de cavalo alado invisível para aqueles que não possuem um entendimento emocional da morte. Esse termo passou a ser usado na denominação *Thestral incognitus* para designar uma nova espécie de inseto descoberta no Chile.

Em artigo, os pesquisadores apresentam a seguinte explicação etimológica:

> **Etimologia:** *Thestral* (gênero masculino), da criatura ficcional criada por J. K. Rowling na saga Harry Potter. Thestrals referem-se a uma raça de cavalos alados

[13] "Os investimentos axiológicos, isto é, o modo com o qual um determinado termo é posto em relação com o positivo ou com o negativo, com a felicidade ou com a infelicidade, com o bem ou com o mal [...] são fundamentais para a constituição dos objetos de valor" (VOLLI, 2012, p. 132). A axiologização positiva ou negativa de um termo é operacionalizada pela oposição da categoria semântica fundamental chamada tímica, entre euforia e disforia, o que, de acordo com a semiótica greimasiana, é a raiz somática dos nossos juízos de valor (VOLLI, 2012, p. 131).

> com corpo esquelético. A *carinae* de marfim e os calos no dorso do novo gênero são semelhantes ao corpo esquelético do thestral de Rowling. Adicionalmente, thestrals não podem ser vistos por todos; a espécie desse novo gênero vem de localidades muito bem coletadas e, mesmo assim, a escassez de espécimes pode ser devido ao fato de não serem facilmente visualizadas por todos [...] Etimologia: *incognitus* do latim, significa desconhecido, referente às poucas coleções da espécie, que é aparentemente rara entre os pentatomídeos do Chile[14]. (FAÚNDEZ; RIDER, 2014, p. 395-397)

Os cientistas que nomearam essa espécie o fizeram com base em uma intersecção semântica entre semas percebidos como semelhantes entre as duas criaturas. Tem-se, portanto, um caso claro de motivação cultural para a denominação do novo inseto, prova de que a subjetividade do cientista e do acervo cultural de uma sociedade influenciam denominações científicas. Em outras palavras, o mundo não se apresenta objetivamente aos homens, mas sim pelo crivo do sujeito que se encontra sob as mais variáveis influências sociocognitivas e culturais. Ademais, esse caso revela o potencial conceptual do termo ficcional em questão, já que ele foi utilizado não por mera referência ao universo de Harry Potter, mas por questões semântico-conceptuais.

Vale destacar também as denominações de conceitos ficcionais que fazem uso da nomenclatura binomial latinizada, grafada em itálico, com gênero iniciado em letra maiúscula e espécie em letra minúscula, tipicamente encontrada em termos das ciências zoológica e botânica, como em *Gernumbli gardensi*, para designar uma espécie de gnomo e *Mimbulus mimbletonia* para designar uma espécie de planta mágica. Entendemos que o enunciador em HP não está buscando a identificação e sistematização de novas espécies como faz um cientista. Contudo, ao atribuir um nome a modo de uma designação científica, o enunciador faz uso de um modo de dizer tipicamente encontrado no universo de discurso científico, o que confere certa legitimidade e cientificidade para uma espécie própria de um mundo ficcional.

Enquanto integrantes de uma linguagem literária especialmente organizada que modeliza uma concepção de mundo semioticamente construído, os termos ficcionais contribuem para a eficácia dessa concepção no plano da expressão. O uso desses termos atribui uma dimensão a mais à narrativa, de modo que, para além dos eventos narrados, há um conhecimento específico próprio de um

[14] No original: "***Etymology:*** *Thestral (Gender masculine), from the fictional creature created by J. K. Rowling in her saga of Harry Potter. Thestrals are a breed of winged horses with a skeletal body. The ivory carinae and calluses on the dorsum of the new genus resemble the skeletal body of Rowling's thestral. Additionally, thestrals cannot be seen by everyone; the specimens of this new genus come from localities that have been fairly well collected, and yet the scarcity of specimens may be due to their not being easily seen by everyone.* [...] ***Etymology:*** *incognitus from Latin, means unknown, referring to the few collections of this species, which is apparently rare among the Chilean pentatomids*" (FAÚNDEZ; RIDER, 2014, p. 395-397).

mundo ficcional que fundamenta a narrativa. Em outras palavras, o itálico em HP, tanto nas denominações de espécies quanto nas denominações de feitiços, dentre outras, assim como outros recursos gráficos passíveis de serem utilizados, como sublinhado, negrito e aspas, é um traço suprassegmental[15] importante para a análise-descrição de termos em textos. Esses recursos agregam uma camada de significação para além do significado das unidades lexicais em si, apontando para usos lexicais que os distinguem dos demais. Portanto, eles são cruciais para um reconhecimento terminológico textual.

Em vista da temática das obras (*Witchcraft and Wizardry*), os termos ficcionais gozam de proeminência semântica ao acionarem uma rede de inter-relações estabelecidas, tanto no interior da obra quanto com outras manifestações literárias. Assim, reiteramos o posicionamento de Barbosa (2006, p. 51) de que "as unidades lexicais atualizadas nos textos mantêm uma rede de relações semânticas específicas – no interior do universo de discurso – e têm funções particulares, quanto à designação e à referência. Por essa razão, são multifuncionais" (BARBOSA, 2006, p. 51). Além disso, também referendamos as considerações da pesquisadora ao concluir que, "é preciso estar familiarizado com as histórias, conhecer o pensamento e o sistema de valores da cultura em questão, para poder compreendê-los bem. De fato, é outra linguagem, que é preciso aprender, para interpretá-los corretamente" (BARBOSA, 2006, p. 50).

Tendo essas considerações em mente propomos a seguinte definição de termos ficcionais:

> [...] unidades lexicais que, na maioria dos casos, designam elementos não pertencentes ao mundo experimentado fisicamente [...], de forma que a existência dos elementos por elas designadas está condicionada ao texto, além de depender parcial ou totalmente da cognição; são também unidades lexicais semanticamente representativas de uma temática, usadas para a composição de um texto literário, tendo em vista a criação de um mundo ficcional. Em outras palavras, essas unidades lexicais habitam o imaginário humano e fazem parte do acervo cultural de dada sociedade. (CARNEIRO, 2016, p. 107)

Ao mesmo tempo em que os termos ficcionais exercem uma função preponderante na composição textual de um mundo ficcional notadamente caracterizado pela fantasia, eles conservam um valor semântico cultural. Em outras palavras,

[15] Esclarecemos que o uso do termo "suprassegmental" para se referir a uma característica da escrita é tomado de acordo com a segunda acepção desse termo encontrada no dicionário Aulete: "2. Situado acima de um segmento" (Disponível em: <http://www.aulete.com.br/suprassegmental>. Acesso em: 26 ago. 2016). Em outras palavras, esse termo, que é comumente empregado para se referir a fenômenos fonológicos, é aqui empregado para se referir a uma dimensão da significação de unidades lexicais que está acima da significação segmental dessas unidades, o que contribui para a proeminência semântica delas.

esses termos atestam como, no processo de evolução histórica da cultura, diferentes simbolizações vão se constituindo e (re)formulando o imaginário humano. São fontes, portas de acesso para o entendimento de uma função importante da linguagem, assim caracterizada por Benveniste (1988, p. 27): "[...] o poder fundador da linguagem, que instaura uma realidade imaginária, anima as coisas inertes, faz ver o que ainda não existe, traz de volta o que desapareceu".

Em seguida, para concluir o que pontuamos ao longo do texto, traçamos algumas aplicações e implicações de estudos terminológicos em obras do universo de discurso literário de fantasia.

5 Considerações finais

Buscamos apresentar alguns aspectos teóricos e metodológicos mínimos que contribuíssem para a identificação de termos no universo de discurso literário de fantasia conforme manifestado na série infantojuvenil Harry Potter, de J. K. Rowling, podendo ser replicados para outras obras literárias desse universo de discurso. Acreditamos ter demonstrado que há especificidades denominativas e conceptuais nessas unidades lexicais que as configuram como termos. O estudo de termos em uso em discursos literários contribui para uma melhor compreensão da cultura, em como manifestações literárias contemporâneas fazem uso de elementos folclóricos e, ao mesmo tempo, fundam uma nova realidade imaginativa, um mundo secundário, de modo que estudar os termos ficcionais é descrever a dimensão do imaginário humano. Assim, descrições de terminologias no universo de discurso literário, como a exemplificada neste trabalho, podem ter aplicações para estudos folclorísticos e literários. O estudo terminológico na ficção também apresenta implicações para os estudos estilísticos de textos literários, bem como para a tradução literária.

É preciso salientar também que, dadas as relações intertextuais entre literatura, cinema, televisão e jogos virtuais, que também se baseiam na criação de mundos ficcionais narrativos, é possível que se considerem manifestações textuais audiovisuais e multimodais para um reconhecimento terminológico da fantasia, da ficção-científica, dentre outros universos de discurso. Assim, se a noção de universo de discurso passa a ser adotada para reconhecimento terminológico, em vez de domínio especializado, entendemos que o foco dos estudos terminológicos pode ser ampliado no sentido de considerar discursos que não são prototipicamente científicos, mas que também fazem uso de unidades lexicais que já apresentam alguma especificidade de conteúdo, com um valor específico atualizado nos textos. Essa ampliação do escopo é benéfica, visto que oportuniza análises e descrições relativas à dinamicidade de unidades lexicais em transição entre distintos níveis de especialização, levando as pesquisas a considerarem também

relações terminológicas interuniverso de discurso. Além disso, de um ponto de vista sociocognitivo, cultural e textual, descrições terminológicas que levem em conta textos literários ficcionais podem vir a ter mais condições de explicar motivações culturais e sociocognitivas de denominações científicas, por exemplo, como no caso comentado da unidade *Thestral*. Logo, torna-se importante que a teoria terminológica de um modo geral amplie o seu escopo no sentido de considerar outros domínios da experiência e atividade humana, além dos científicos, tecnológicos e profissionais, passíveis de fazerem uso de conjuntos lexicais com algum valor terminológico.

Agradecimentos

Agradeço à CAPES pela bolsa de pesquisa de mestrado concedida de 2014/2-2016/1. Agradecimentos também se fazem necessários à profa. dra. Ana Luiza Pires de Freitas pelos comentários e sugestões que contribuíram para o esclarecimento de certos aspectos do texto. Quaisquer imprecisões que persistirem são de minha responsabilidade. Também agradeço à profa. dra. Maria José Bocorny Finatto por acreditar neste trabalho.

Referências

BARBOSA, M. A. Etno-terminologia e Terminologia Aplicada: objeto de estudo, campo de atuação. In: ISQUERDO, A. N.; ALVES, I. M. (Org.). *As ciências do léxico*: lexicologia, lexicografia, terminologia. Campo Grande: UFMS; São Paulo: Humanitas, 2007, p. 433-445.

_____. Para uma etno-terminologia: recortes epistemológicos. *Cienc. Cult.,* São Paulo, v. 58, n. 2, jun. 2006. Disponível em: <http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200018&lng=en&nrm=iso>. Acesso em: 02 ago. 2013.

BARROS, L. A. Aspectos epistemológicos e perspectivas científicas da terminologia. *Cienc. Cult.,* São Paulo, v. 58, n. 2, jun. 2006. Disponível em: <http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200011&lng=en&nrm=iso>. Acesso em: 15 jul. 2015.

BENVENISTE, É. *Problemas de Linguística Geral I*. 2. ed. Campinas: Editora da UNICAMP, 1988.

BERBER SARDINHA, T. *Linguística de* Corpus. Barueri: Manole, 2004.

CARNEIRO, R. M. O. *Discurso literário de fantasia infantojuvenil*: proposta de descrição terminológica direcionada por *corpus*. 2016. 281f. Dissertação (mestrado em Linguística e Linguística Aplicada). Universidade Federal de Uberlândia, UFU, Uberlândia, 2016. Disponível em: <https://repositorio.ufu.br/handle/123456789/18082>. Acesso em: 10 out. 2017.

DOLEŽEL, L. *Heterocosmica*: fiction and possible worlds. Maryland: The Johns Hopkins University Press, 1998.

ESPERANDIO, Isabela. B. Legendas de seriados de tema sobrenatural : uma abordagem terminológica para tradutores. 2015. 2229f. Dissertação (mestrado em Letras). Universidade Federal

do Rio Grande do Sul, UFRGS, Porto Alegre - RS, 2015. Disponível em: <http://hdl.handle.net/10183/131767>. Acesso em: 04 abr. 2018.

FAÚNDEZ, E.; RIDER, D. Thestral incognitus, a new genus and species of Pentatomidae from Chile (Heteroptera: Pentatomidae: Pentatominae: Carpocorini). *Zootaxa,* v. 3884, n. 4, 2014, p. 394-400. Doi: http://dx.doi.org/10.11646/zootaxa.3884.4.9

HUNT, P. *Crítica, teoria e literatura infantil*. São Paulo: Cosac Naify, 2010.

KRIEGER, M. da G.; FINATTO, M. J. B. *Introdução à terminologia*: teoria e prática. São Paulo: Contexto, 2004.

KRONZEK, A. Z.; KRONZEK, E.. *The sorcerer's companion*. 3. ed. New York: Broadway Books, 2010.

PAIS, C. T.; BARBOSA, M. A. Da análise de aspectos semânticos e lexicais dos discursos etno-literários à proposição de uma etno-terminologia. *Matraga*, Rio de Janeiro, v. 11, n. 16, p. 79-100, jan./dez. 2004.

PEARSON, J. *Terms in Context*. Amsterdam: John Benjamins, 1998.

RYAN, M.-L. Mundos impossíveis e ilusão estética. *Revista Contracampo*, Niteroi, v. 29, n. 1, p. 4-25, abr./jul. 2014. Doi: http://dx.doi.org/10.22409/contracampo.v0i29.658

SANTOS, A. Complexidade e transdisciplinaridade em educação: cinco princípios para resgatar o elo perdido. *Rev. Bras. Educ.*, Rio de Janeiro, v. 13, n. 37, p. 71-83, abr. 2008. Doi: http://dx.doi.org/10.1590/S1413-24782008000100007.

SCOTT, M. *WordSmith Tools*. Versão 6. Liverpool: Lexical Analysis Software, 2012.

STEGER, H. O que é linguagem literária? *Fragmentos*, Florianópolis, n. 3, p. 101-140, jan./dez, 1987.

VOLLI, U. *Manual de semiótica*. 2. ed. São Paulo: Loyola, 2012.

Linguagem oral e variação dialetal

Pommersche korpora: um conjunto de *corpora* dialetais da variedade brasileira do pomerano

Pommersche korpora: a set of dialectological *corpora* of the brazilian pomeranian variety

Neubiana Silva Veloso Beilke

Resumo: Nosso objetivo geral foi a compilação de *corpora* do pomerano no Brasil, utilizando um amplo conjunto de fontes a fim de contribuir para a documentação do dialeto. Nossos objetivos específicos foram tentar identificar o uso ativo do pomerano no Vale do Rio Doce (MG) e do Vale do Rio Pardo (RS) e comparar os dados coletados nessas regiões. Fundamentamo-nos na Lexicologia, na Linguística de *Corpus* e na Sociogeolinguística. Nossa metodologia envolveu a compilação, transcrição e tratamento de dados. Ao final, constituímos um conjunto de *corpora* denominado *Pommersche Korpora*, que contém quatorze *corpora* escritos e um *corpus* oral. Constatamos a existência de uma variedade brasileira do pomerano. Acreditamos ter contribuído por meio de procedimentos metodológicos desenvolvidos que podem ser replicados para outros dialetos.

Palavras-chave: Pomerano. Linguística de *Corpus*. Sociogeolinguística. *Pommersche Korpora*.

Neubiana Silva Veloso Beilke – Doutoranda pelo Programa de Pós-Graduação em Estudos Linguísticos do Instituto de Letras e Linguística da Universidade Federal de Uberlândia/UFU, mestre em Estudos Linguísticos pela Universidade Federal de Uberlândia, bolsista CAPES – neubianabeilke@ufu.br.

Abstract: Our main objective was to compile *corpora* of Pomeranian in Brazil, with the adoption of a wide range of sources to contribute to the Pomeranian documentation. Our specific objectives were to identify the survival of Pomeranian in the regions of Vale do Rio Doce (MG) and Vale do Rio Pardo (RS), and to compare Pomeranian collected in these regions. The study is theoretically based on Lexicology, *Corpus* Linguistics and Sociogeolinguistics. Our methodology was the compilation, transcription and processing of data. Finally, this resulted into a set of *corpora* called *Pommersche Korpora* that contains fourteen textual *corpora* and one spoken *corpus*. We identify the existence of a Brazilian Pomeranian variety. We believe this study leaves some contributions for dialectological studies.

Keywords: Pomeranian. *Corpus* Linguistics. Sociogeolinguistics. *Pommersche Korpora*.

1 Introdução

Este artigo é um resumo da nossa pesquisa de mestrado. Nele, delineamos o tema de estudo, o pomerano, bem como o objetivo geral e os específicos, a justificativa da nossa pesquisa, as hipóteses levantadas, os procedimentos metodológicos desenvolvidos e aplicados, os resultados, além de algumas análises preliminares esboçadas a partir da observação dos resultados da compilação dos *corpora* processados em programa de análise lexical. Por fim, apresentamos algumas perspectivas futuras para a exploração do banco de dados linguísticos que criamos. Listamos, além disso, algumas contribuições que ele pode oferecer para pesquisas em Linguística de *Corpus* (LC), em Dialetologia e em Sociolinguística.

O pomerano é uma variedade linguística germânica, que pertence ao tronco indo-europeu e à família das línguas germânicas, estando situado no grupo do baixo-alemão, proveniente das terras planas do norte da Europa. Enfocamos especificamente o *Pommersches Plattdeutsch (*VOLLMER, 2008; HERRMANN-WINTER, 1998, 2003, 2013), ou seja, o baixo-alemão pomerano, e suas formas presentes no Brasil. Convencionamos aqui que iremos nos referir a essa variedade em alguns momentos apenas como pomerano, em outros, como variedade brasileira do pomerano (VBP), ou como baixo-alemão pomerano (BAP) a fim de evitar o sobreuso do referente.

2 Objetivos

Nosso objetivo geral foi compilar materiais autênticos, os quais continham amostras linguísticas provenientes do dialeto pomerano falado no Brasil com a finalidade de compor um conjunto formado por *corpora* escritos e um *corpus* oral (transcrito) e, consequentemente, constituir um banco de dados linguísticos (*corpus*) para futura exploração e descrição por meio da abordagem-metodologia da Linguística de *Corpus* e também estudos em Lexicologia e Sociogeolinguística.

Nossos objetivos específicos foram identificar a sobrevivência ou o desaparecimento da variedade nas regiões alvo de nossa coleta de dados – o entorno do Vale do Rio Doce (MG) e o interior do Rio Grande do Sul (RS) – e comparar brevemente os dados coletados nessas duas regiões.

3 Justificativa

Nossa iniciativa parte da consciência de que precisamos agir frente à possibilidade de desaparecimento da variedade pomerana, devido às constatações, feitas por meio do contato pessoal com usuários dessa língua, pelas visitas de campo e pelos estudos de autores como Pessoa (1995), Vogt (2001), Seyferth (2003), Dück (2005; 2008), Maltzahn (2011), Heinemann (2013), entre outros. Esses estudiosos indicam que o pomerano corre risco de desaparecimento, visto que as novas gerações não estão fazendo uso do dialeto e que, ao morrerem os falantes mais idosos, o falar entra cada vez mais em declínio e pode cair em desuso, pela diminuição gradual do número de falantes.

Outra razão em contribuir para a documentação linguística está no fato de que, na região da antiga Pomerânia oriental, onde hoje é a Polônia, o pomerano oriental não é mais falado (GRANZOW, 2009), salvo por alguns descendentes dos pomeranos orientais que ainda se reúnem em encontros periódicos e fazem uso linguístico dele (WANGERINER TREFFEN, 2015).

As problemáticas de pesquisa impõem-se a partir das lacunas deixadas por outros trabalhos e por questões que não tenham sido satisfatoriamente estudadas em determinadas localidades. Não foram localizados meios de promoção e de estudo do pomerano em algumas regiões do Brasil, como no entorno do Vale do Rio Doce, no leste de Minas Gerais, e no Vale do Rio Pardo, no Rio Grande do Sul. Por isso, um estudo comparativo entre os dialetos falado nessas regiões pôde ser realizado, com base na variação do uso de itens lexicais, encontrados por meio de *corpora*, tais como os que nos propusemos a reunir. É necessário que se identifique melhor a situação nessas regiões e que também sejam reconhecidas as iniciativas que promovem o resgate e o fomento nas regiões menos pesquisadas. Dessa forma, escolhemos enfocar o pomerano praticado em localidades nas quais a população é remanescente da imigração pomerana, mesmo que não tenha ainda recebido grandes menções e reconhecimentos pela mídia ou pelas pesquisas.

No caso de Minas Gerais, é recente a descoberta desses descendentes, e há poucos trabalhos acerca do assunto. De forma geral, os pomeranos ficaram isolados em seu grupo por muito tempo, até serem descobertos pela mídia (SEIBEL, 2010). Granzow (2009, p. 161-162) também relatou ter encontrado comunidades bem fechadas e isoladas. No caso específico de Itueta (MG) e de Vila Neitzel (MG), houve um isolamento geográfico, devido à necessidade de se atravessar

o Rio Doce de balsa para se chegar até eles, em períodos em que as estradas da zona rural norte de Itueta não eram transitáveis. O pomerano falado por eles pode apresentar peculiaridades – como o contato com o regionalismo mineiro do português – que mereçam um estudo comparativo com aquele falado na região do Vale do Rio Pardo (RS) – espaço povoado até hoje por muitos descendentes de imigrantes pomeranos, que preservam suas manifestações linguísticas, que precisam ser verificadas e descritas.

Compilar *corpora* dessa variedade germânica é importante para a comunidade falante e para os pesquisadores. Como variedade de língua, comporta toda uma forma de conceber o mundo e expressa um modo de pensar, de sentir, de organizar, de nomear, de significar, de interagir e de experienciar novas situações. Possibilitar meios de estudo e de descrição do pomerano pode contribuir para a preservação desse dialeto. Resgatar e registrar o dialeto é uma forma de oportunizar o conhecimento do vocabulário pomerano por outras comunidades, na qualidade de herança cultural, de tesouro linguístico dos pomeranos e de patrimônio da sociedade. É importante para conhecermos a forma que o léxico assume na cultura pomerana. Salvaguardar o léxico também é uma forma de ampliar o conhecimento das relações entre língua, sociedade e cultura. Tudo isso justifica a razão de sua documentação.

Consideramos que é possível o desenvolvimento de futuras descrições e análises com base no levantamento de dados empíricos reunidos em *corpora*, oferecendo condições e opções para estudo, congregando, inclusive, dados que permitam a produção de materiais didáticos para o ensino do pomerano.

4 Fundamentação teórica

Para alcançar os referidos objetivos, o estudo teve como fundamentação teórica a Lexicologia e a Linguística de *Corpus*, além de dialogar com campos interdisciplinares, como a Sociogeolinguística, e outras áreas da Linguística, como a Dialetologia.

Nosso objeto de estudo, o léxico pomerano, situa-se, na Linguística, dentro da área do conhecimento da Lexicologia. A Lexicologia, por sua vez, é o ramo da Linguística que se dedica ao estudo científico do léxico, uma ciência que abarca tanto as unidades lexicais quanto a organização do léxico.

Nós, seres humanos, sujeitos, etnias, comunidades linguísticas, nomeamos e conceituamos nossas experiências por meio da nossa "cognição da realidade" e *Weltanschauung* (concepção de mundo), expressas por meio do léxico. De acordo com Biderman (2001, p. 13-14), "o léxico de uma língua natural constitui uma forma de registrar o conhecimento do universo" e "[...] pode ser identificado com o patrimônio vocabular de uma dada comunidade linguística ao longo de sua história" (BIDERMAN, 2001, p. 14).

Assim, o léxico assume formas distintas nas várias línguas e culturas, com dimensões significativas próprias e organizações em sistemas semântico-léxico-gramaticais diferentes. Ele é o conjunto de palavras pertencentes a uma língua e, quando elas são materializadas em um texto, na modalidade oral ou na escrita, são chamadas de "vocabulário". De acordo com Biderman (1992, p. 399), o léxico é "o tesouro vocabular de uma língua". O léxico é também definido, conforme o conceito formulado por Zavaglia e Welker (2013), como "o conjunto rico e dinâmico de todas as palavras de uma língua que possuem organização enunciativa interdependente" (ZAVAGLIA; WELKER, 2013, s/p).

Quanto à Linguística de *Corpus*, nós nos embasamos, precisamente, nas teorias que fundamentam a abordagem-metodologia da LC. Realizamos um levantamento de diversas definições formuladas a respeito da LC e, posteriormente, nos posicionamos. Nas definições pesquisadas a LC aparece como "campo de criação e análise" (BERBER SARDINHA, 2009, p.7), "campo disciplinar" (MELLO, 2012, p. 53), "área de pesquisa" (FEITOSA, 2005, p. 34), "área do conhecimento" (GONZALEZ, 2007, p. 8) e "ramo específico do saber ou disciplina" (TAGNIN, 2005, p. 21). Nessas definições, a LC é vista de uma maneira ampla, como fomentadora de pesquisas e produtora de conhecimento, não se reduzindo a uma metodologia, embora seja definida também como "metodologia para investigações empíricas" (PARODI, 2010, p. 15). Em Gonzalez (2007), encontramos uma conceituação clara, de que a LC "é uma área do conhecimento que estuda a linguagem por meio da utilização de grandes quantidades de dados empíricos relativos ao efetivo uso da linguagem, com o auxílio do computador" (GONZALEZ, 2007, p. 8).

Parodi (2010), vê a LC como uma metodologia investigativa com princípios reguladores poderosos. No entanto, para outros autores, a LC pode ser definida apenas como metodologia, como é o caso de McEnery e Wilson (1996), embora também a considerem o "estudo da língua baseado em exemplos de usos linguísticos na 'vida real' […]" (McENERY; WILSON, 1996, p. 1-2).

Os linguistas alemães Lemnitzer e Zinsmeister (2006) afirmam que, por meio da LC, realizamos a descrição de enunciados de línguas naturais, seus elementos e estruturas e, a partir disso, podemos formular constructos teóricos, pois, para eles, a LC permite a "elaboração teórica com base na análise de textos autênticos" (LEMNITZER; ZINSMEISTER, 2006, p. 10).

Por tudo isso, acreditamos que a LC pode ser concebida de uma maneira ainda mais ampla, reconhecendo seus atributos e contribuições para o conhecimento linguístico. Assim, a partir de nossa reflexão sobre as definições de LC mobilizadas, definimos que a LC é uma abordagem-metodologia de princípios descritivos, que se fundamenta em dados autênticos e se relaciona com as evidências de maneira ampla. Ela permite a produção de conhecimentos variados, ancorados na realidade linguística, além de nos guiar para investigar hipóteses não premeditadas e para

permitir a descoberta e a comprovação de fatos linguísticos. A LC dá primazia à observação prévia dos dados levantados. Esse empirismo, um de seus pressupostos, é um meio de fundamentar a pesquisa objetivamente, em detrimento da especulação. Sob sua perspectiva, a análise dos dados permite verificar traços que se repetem, padrões de comportamento linguístico, variações recorrentes e, assim, atestar se existem regularidades sistemáticas. A partir de então, torna-se possível quantificá-las, descrevê-las e analisá-las confirmando a hipótese de que não são aleatórias, o que contribui para esclarecer suposições a respeito do funcionamento da língua. Consideramos a LC como um campo interdisciplinar que, ao ser ao mesmo tempo abordagem e metodologia, produz inúmeros conhecimentos inovadores e os mais diversificados estudos e olhares sobre a linguagem em geral.

Desse modo, acreditamos que a LC não se restringe a um método, mas é também uma abordagem da língua. Para corroborar nossa visão, lembramos a afirmação de Berber Sardinha de que a LC é "uma perspectiva, isto é, uma maneira de se chegar à linguagem" (2004, p. 37). Novodvorski e Finatto (2014) também legitimam nossa concepção de LC, de que se trata de uma forma diferenciada de ver e abordar a língua, pois afirmam que a LC também é um modo de compreender a língua "[...] que permite apreciar um objeto de estudo sob um ângulo diferenciado, que se constitui uma nova área de pesquisa, com abordagens e métodos próprios" (NOVODVORSKI; FINATTO, 2014, p. 9-10).

Ademais, dada a natureza interdisciplinar desta pesquisa e com o objetivo de atender às suas peculiaridades, adotamos o aporte teórico-metodológico do campo de estudos da Sociogeolinguística (doravante SGL). A SGL, por sua nomenclatura, já sugere abranger os princípios e as variáveis da Sociolinguística e da Geolinguística. Portanto, recorremos a elementos de ambas, as quais subsidiaram nossa coleta de dados orais para a compilação de um *corpus* oral dialetal do pomerano no Brasil.

Adotando a definição objetiva de Cardoso, podemos dizer que a Geolinguística ou Geografia Linguística é um método da Dialetologia que se incumbe de "recolher de forma sistemática o testemunho das diferentes realidades dialetais refletidas nos espaços considerados" (CARDOSO, 2010, p. 46). Ou seja, é o estudo empírico das variedades linguísticas na relação com as diatopias.

A Sociolinguística foi adotada como uma abordagem que considera as relações entre língua e sociedade, não se restringindo às variáveis linguísticas, mas também considerando as variáveis sociais e relacionando ambas. Segundo Von Borstel (2014), é essencial observar o contexto social, o cultural, o étnico, o religioso, o político e o econômico em uma sociedade ou em uma comunidade de fala.

O termo "Sociogeolinguística" (SGL), segundo afirmam Cristianini e Encarnação (2009, p. 91 apud CRISTIANINI, 2012, p. 26), surgiu em 2004 para designar os estudos geolinguísticos que consideram fatores tanto geográficos quanto sociais para coleta, para registro e para análise de dados linguísticos, pois

muitos trabalhos utilizaram as duas abordagens juntas, considerando as variáveis sociais e unindo o método da Sociolinguística ao método da Geolinguística. Conforme Cristianini (2012), essas pesquisas "revelam aspectos que se reportam à atualização do léxico num processo de mudança linguística e à compreensão das subjacências presentes a cada designação" (CRISTIANINI, 2012, p. 30).

Quando é feita a transcrição e se inicia o tratamento e a análise dos dados, a abordagem da SGL não se restringe a aspectos quantitativos, mas também abrange os qualitativos. Os dados também podem ser aproveitados para outras análises, pois foram coletados sob uma base metodológica criteriosa e possuem legitimidade. Eles constituem dados atuais e são evidências empíricas para o desenvolvimento de estudos descritivos. Foram unidas as abordagens-metodologias da SGL e da LC, no que se refere a coletar dados orais por meio da primeira e analisar esses dados por meio da segunda. Em nosso trabalho, estabelecemos uma relação entre a LC e a SGL, pois são abordagens empíricas e descritivas. Consideramos que ambas oferecem meios eficientes para coleta e para a compilação de conteúdo lexical; permitem a percepção das variações linguísticas e também a obtenção de resultados produtivos e analisáveis com as ferramentas da LC, que opera com grandes quantidades de dados.

Concluímos que a SGL, definida como campo interdisciplinar e como abordagem-metodologia, foi fundamentação eficiente para atender à coleta de dados orais para a constituição do *corpus* oral.

5 Hipóteses

Formulamos quatro hipóteses para esta pesquisa:

i) a cogitação de que o pomerano não é ágrafo, pois a coleta e a compilação de *corpora* demonstram a existência de formas escritas, com traços dialetais, presentes em cartas e em textos advindos de descendentes de pomeranos. O conhecimento de uma grafia germânica foi se perdendo ao longo das gerações que se sucederam pela não continuidade do estudo em língua alemã, devido ao processo de escolarização brasileiro e à aprendizagem do português. *Corpora* escritos do pomerano demonstram a presença de alguma escrita pomerana; portanto, defendemos que não houve caso de agrafia, mas uma perda de cultura escrita;

ii) no início da imigração, as comunidades pomeranas estabeleceram-se no campo, em áreas rurais, para se dedicarem aos trabalhos na área da agricultura. Como os pomeranos ainda são considerados um grupo predominantemente camponês, acreditamos que os falantes são encontrados, na sua maioria, no meio rural, podendo ser inexistente, em algumas localidades na zona urbana, a presença

de falantes. Essa hipótese aplica-se aos dois principais pontos de coleta: Vale do Rio Pardo (RS) e Vale do Rio Doce (MG);

iii) existência de interferências de caráter linguístico, a serem verificadas, sofridas pelos pomeranos em contato com outras etnias, pois, nos pontos de pesquisa selecionados como principais – Vale do Rio Pardo (RS) e Vale do Rio Doce (MG) –, houve contato de pomeranos com outras etnias. Por essa razão, os dados de *corpora* podem apresentar evidências desse contato interdialetal, bem como evidências de variação lexical na comparação entre essas regiões; e

iv) a consideração de que o contato dos descendentes de pomeranos no Brasil com a língua portuguesa e o seu prestígio como língua oficial do país permitiram uma interferência português-pomerano, a ponto de o pomerano falado atualmente no Brasil já representar uma variedade brasileira do pomerano, na qual o contato de línguas tenha propiciado usos com empréstimos lexicais do português no sistema pomerano. A hipótese, nesse caso, é a de que os *corpora* do pomerano apresentam indícios na formação de frases e na inovação de itens lexicais que demonstrem essa relação de "mistura" pomerano-português. A essa variedade chamamos *Brasilianisch-Pommersch*.

6 Metodologia

Nossa metodologia, de modo geral, envolveu a coleta, a compilação, a transcrição e o tratamento de dados escritos e orais, bem como a convenção e a conversão da escrita em um padrão único (uniformizado, dentro do contexto da nossa pesquisa) para as amostras linguísticas pomeranas, tanto aquelas coletadas na primeira fase, da constituição dos *corpora* escritos, quanto na segunda fase, da coleta, transcrição e composição do *corpus* oral.

Para nós, tratamento de dados é todo e qualquer trabalho de manuseio e/ou ação com a fonte linguística após sua coleta. O tratamento de dados é uma série de ações necessárias para constituição de *corpora* para estudos em LC. Dentre essas ações, listamos a própria compilação, além da transcrição, da conversão e da convenção da escrita (caso necessário, devido às variações) e da análise de dados. Também consideramos como tratamento dos dados a inserção de cabeçalhos, qualquer modificação no arquivo (sem alterar o conteúdo lexical), a organização, o salvamento em extensão *.txt, a limpeza, a etiquetagem e quaisquer anotações de análise e comentários do compilador e/ou analista, dentre outras alterações, a partir da forma exata na qual um texto foi coletado originalmente.

Em suma, realizar o tratamento dos dados é deixá-los organizados e em condições de serem legíveis por programas de computador e operáveis pelo linguista de *corpus*.

6.1 Primeira fase: constituição dos corpora escritos

Coletamos textos escritos em pomerano, fotografando descrições de museus pomeranos, alguns *Wandschoener* (enfeites colocados na parede com inscrições), *Huusspruch* (inscrições com provérbios pintados, geralmente, acima da porta principal das casas) encontrados em residências e lápides de cemitérios de imigrantes – aquelas que se referiam a túmulos de pomeranos e que continham inscrições com indícios de traços dialetais.

Também coletamos cartas, letras de músicas, poemas, lembranças, anotações e receitas culinárias doadas por pomeranos em cidades como Canguçu (RS), São Lourenço do Sul (RS) e na zona rural de Itueta (MG). Fotografamos algumas páginas de bíblias antigas em baixo-alemão pomerano, coletamos matérias jornalísticas e calendários pomeranos na internet, assinamos o jornal *Folha Pomerana*, transcrevemos registros eclesiásticos, extraímos legendas de documentários. Enfim, buscamos os mais diversificados meios de obter textos, partes de textos e/ou algum material escrito em pomerano.

A compilação desses materiais foi feita de diversas formas: digitando manualmente e conferindo atentamente, procedimento adotado quando se tratava de material com conteúdo menor. No caso de livros inteiros em BAP, os textos foram escaneados em formato *.pdf e, já em via digital, foram tratados por um programa leitor de caracteres ópticos. Em casos de textos da internet foram compilados e salvos, primeiramente em extensão *.doc, depois foram identificados por meio de cabeçalhos e, em seguida, salvos em formato *.txt. A esse procedimento, chamamos de "compilação direta". Porém, a maioria dos dados em pomerano obtidos na forma escrita foram coletados pessoalmente em viagens de campo realizadas para as localidades pesquisadas.

No nosso caso, cujo objetivo principal foi compor os *corpora* para permitir a descrição lexical, o tratamento da escrita pomerana, denominado por nós de "conversão e convenção", foi indispensável. Essa estratégia tornou possível a obtenção de um conjunto de *corpora* que fosse legível pelos programas de análise lexical, e identificável quando das buscas pelos mesmos itens lexicais. Esses itens, dispostos em diferentes grafias, não seriam facilmente percebidos como se tratando das mesmas unidades. Por essa razão, a reescrita em uma grafia padronizada foi uma solução que desenvolvemos no contexto deste trabalho.

6.1.1 Conversão da escrita

Chamamos de "conversão da escrita" a ação de reescrever os textos pomeranos conforme o padrão germânico, sem, contudo, utilizar caracteres escandinavos, principalmente a letra /å/ (lê-se a-*ablaut*), chamada popularmente de "a coroado", em português. Essa letra está presente na língua dinamarquesa, por exemplo. Os

textos com a escrita pomerana dicionarizada no Brasil por Tressmann (2006) que utiliza o /å/ não puderam ser legíveis pelo programa WST, conforme os testes que realizamos durante a compilação.

No nosso caso, o programa precisava ser configurado em língua alemã. Mas, como a letra /å/ pertence ao dinamarquês e o WST não configura duas ou mais línguas ao mesmo tempo, notamos que o programa apresentou caracteres distorcidos e que alguns itens lexicais foram perdidos, impedindo a legibilidade dos textos. Por isso, fizemos a conversão de todos os casos de /å/ por "oo", pois o som em pomerano é de um /o/ longo e alto. Além disso, encontramos na literatura pomerana casos de duplo /o/ nas posições em que Tressmann (2006), decidiu grafar como /å/. Citamos, por exemplo, o caso de *Hågel* (TRESSMANN, 2006, p. 182), ou seja, *Hoogel* (granizo/chuva de pedra), em pomerano, e, *Hagel* em alemão.

Fizemos a conversão da escrita dicionarizada nos casos de itens que encontramos escritos de outras formas na literatura pomerana, anteriores ao dicionário, visto que a estética gráfica da escrita pomerana dicionarizada no Brasil aparenta uma distância do pomerano em relação ao alemão maior do que ela realmente é. Esses casos estão convencionados e justificados em um documento auxiliar que criamos, como um passo do procedimento metodológico de convenção.

Também realizamos a conversão da escrita no caso de textos pomeranos encontrados em escrita transliterada ou escrita conforme o método de Wiesemann (2008), ou seja, escritas conforme a interpretação gráfica de falantes do pomerano alfabetizados somente em língua portuguesa. Nesse caso, os testes demonstraram que essas escritas eram legíveis pelo WST, porém, decidimos converter a escrita para padronizar somente uma forma de escrita nos *corpora*, pois, do contrário, não conseguiríamos detectar as repetições nas listas de frequência da Wordlist.

Inserimos abaixo um pequeno excerto do nosso trabalho de conversão da escrita, à guisa de exemplificação:

a) Fonte original:
Deych yexiht fó Noé is pasit zeya fél yóó […]
Noé vee a gaua meyx, vat léva dee up ayna veylt
vat ful xléht lüya vee (SOCIEDADE BÍBLICA DO BRASIL – SBB, 2012).

b) Tratamento da fonte:
Dees geschichte vo <von> Noeh is passiert seehr vel joohr […] Noeh wär a <ein> gauer mensch, wat leewa de up einer werlt wat vull schlecht lüür wär (BEILKE, 2014).

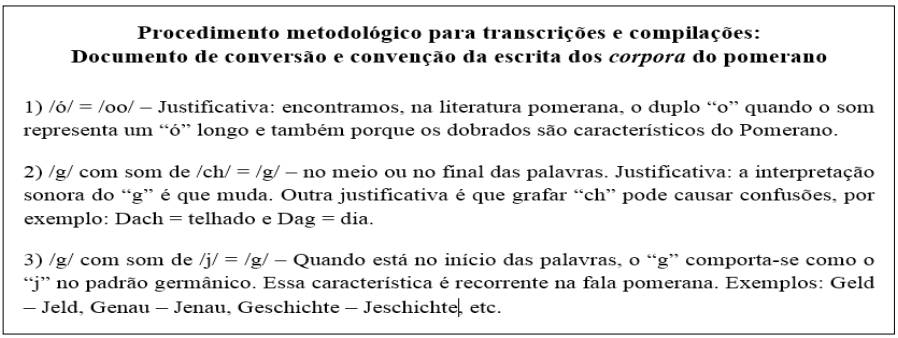
O procedimento de conversão reescreve todas as lexias de conteúdo semântico igual, porém com variações na forma de grafia, para uma mesma forma de escrita, padronizada no processo paralelo à conversão, que é a convenção, ou seja, a normatização das decisões de escrita.

6.1.2 Convenção da escrita

A convenção é a padronização, dentro do nosso trabalho, das decisões que tomamos em relação à conversão da escrita. Esse procedimento mantém, em um mesmo padrão, a conversão que fizemos. Na convenção é que justificamos (motivados, inclusive, pela recorrência das formas de escrita dos *corpora* "crus") a escrita que decidimos adotar. A convenção da escrita é um procedimento metodológico que, além de explicar as escolhas que fizemos e os critérios para fazê-las, dá coerência e uniformidade ao trabalho de compilação dos nossos *corpora*.

Esclarecemos que não criamos uma forma de escrita para o pomerano e nem propusemos uma nova forma de escrever a variedade. Nossa reescrita é apenas um recurso necessário, dentro do contexto da nossa pesquisa, de acordo com os objetivos que definimos e para conseguir alcançá-los.

A seguir, podemos visualizar um trecho do documento, ilustrado pela Figura 1, no qual constam as decisões que tomamos na execução dos procedimentos de conversão e de convenção da escrita pomerana para a compilação dos nossos *corpora*.



Procedimento metodológico para transcrições e compilações:
Documento de conversão e convenção da escrita dos *corpora* do pomerano

1) /ó/ = /oo/ – Justificativa: encontramos, na literatura pomerana, o duplo "o" quando o som representa um "ó" longo e também porque os dobrados são característicos do Pomerano.

2) /g/ com som de /ch/ = /g/ – no meio ou no final das palavras. Justificativa: a interpretação sonora do "g" é que muda. Outra justificativa é que grafar "ch" pode causar confusões, por exemplo: Dach = telhado e Dag = dia.

3) /g/ com som de /j/ = /g/ – Quando está no início das palavras, o "g" comporta-se como o "j" no padrão germânico. Essa característica é recorrente na fala pomerana. Exemplos: Geld – Jeld, Genau – Jenau, Geschichte – Jeschichte, etc.

Figura 1 – Parte do documento de conversão e de convenção da escrita pomerana
Fonte: Elaboração da autora

Segue, no Quadro 1, um comparativo de algumas formas de escrita que encontramos nos *corpora* e em obras lexicográficas.

Quadro 1 – Quadro comparativo de formas de escritas pomeranas

COMPARAÇÕES ENTRE FORMAS DE ESCRITA DO POMERANO			
***Corpora* escritos**	**Das Pommersche Wörterbuch, Hermann-Winter (1999)**	**Pomerisch-Portugijsisch Wöirbauk, Tressmann (2006)**	**Projeto Pomerando, Kuhn Silva (2012)**
Haus, Hause, Hus, Huus	Huus	Huus	Hus
Hagel, Hoogel, Hågel, Hoochel	Hågel	Hågel	Hóhal
Wiet, Weiss, Witt	Wiet	Wit	vit
Futéla, votela, fortela, fotela	vertellen	fortela	futéla
Frau, Fruh, Fruu, Fruuh, Fruch, Fruug	Fru	Fruug	Fruch
Kerl, Keirl, Keyrl, Kejrl	Kierl	Keirl	Kêil
Schaffe, Schoppe, Schopa, Schoppa	Schoppe	Schåp	Schop

Fonte: Elaboração da autora

No Quadro 1, é possível visualizar, à esquerda, as diversas formas de escrita dos mesmos itens lexicais, porém com as variações encontradas nos textos pomeranos; à direita, também é possível visualizar as diferentes propostas de escrita já publicadas. Por isso, foi necessário reescrever todo o material coletado, padronizando todos os dados em uma única forma de grafia.

Esclarecemos que salvamos uma versão em *.doc, antes da conversão e da convenção das formas de escrita encontradas, a fim de permitir um estudo comparativo das formas de escritas pomeranas originalmente encontradas, bem como para preservar os dados dialetais originais. Também salvamos uma versão em *.doc após os procedimentos supracitados, e uma versão totalmente tratada, em *.txt, além de salvarmos a Wordlist dos *corpora* escritos e alguns exemplos de linhas de concordâncias dos nossos testes.

Os procedimentos acima descritos permitiram a coleta e a compilação de dados para a constituição dos nossos *corpora* escritos. Após todas essas etapas de coleta, de compilação e de tratamento de dados, os *corpora* foram organizados. A organização dos *corpora* escritos foi realizada juntamente com a organização do *corpus* oral, pois foi necessário agrupar a vasta gama de materiais coletados sob um critério comum, bem como separar os *subcorpora* por suas respectivas diamesias.

6.2 Segunda fase: constituição do corpus oral

Para a coleta de dados orais, adaptamos o Questionário Sociolinguístico (QS) e o Questionário Semântico Lexical (QSL), este último para a versão em alemão *Lexikalisch-Semantischer Fragebogen* (LSF); selecionamos algumas localidades para realizarmos visitas prévias, a fim de fazer a caracterização histórico-geográfica das

localidades; organizamos um plano de recrutamento com critérios de inclusão e de exclusão dos candidatos às entrevistas, segundo as variáveis da SGL e as exigências do CEP (Comitê de Ética em Pesquisas com Seres Humanos); realizamos, também, procedimentos para a efetivação das entrevistas, como, por exemplo, a preparação dos instrumentos de coleta (questionários e gravadores) e testagem do método por meio da realização de entrevistas piloto.

Fizemos a coleta nas seguintes localidades: Vila Neitzel (MG); Itueta (MG); Arroio do Tigre (RS); São Lourenço do Sul (RS); Canguçu (RS) e Santa Maria de Jetibá/ES. Decidimos realizar coleta de dados também em Santa Maria de Jetibá/ES, por ser conhecida como a capital pomerana do Brasil e para permitir futuras comparações com o pomerano encontrado em outras localidades.

Após essa etapa, efetuamos o tratamento dos dados orais, quando foi necessário desenvolver um método de transcrição, o qual foi feito com base na ortografia germânica do dicionário *Duden* e da bíblia pomerana (a *Barther Bibel*). Ademais, realizamos os procedimentos de etiquetagem parcial e de exclusão de dados antroponímicos. A forma de escrita adotada para a transcrição das entrevistas e constituição do *corpus* oral foi a mesma escrita convertida e convencionada para os *corpora* escritos.

Ainda em relação aos procedimentos metodológicos para a constituição do conjunto dos *corpora* pomeranos, tanto escritos quanto oral, organizamos todo o material compilado como uma unidade, processo durante o qual realizamos a codificação, a nomenclatura, o agrupamento, a separação e o reagrupamento dos dados, segundo suas classificações em diamesias ou gêneros ou domínio discursivo ou suportes.

6.3 Recursos utilizados e outros procedimentos

Os recursos utilizados foram o OmniPage®, o ArcGis® e o Word Smith Tools®. Também utilizamos pontualmente um extrator de legendas, o SubExtractor®, versão de 2013, com o auxílio de um colaborador, para compilar as legendas dos documentários pomeranos.

Para compilar textos de livros, recorremos ao conversor de imagem em texto, ou seja, um programa reconhecedor de caracteres óticos – OCR (*optical character recognition*), o OmniPage Professional, versão 17.0 de 2009 – e fizemos a revisão e a correção dos caracteres que apresentaram distorções.

Para realizar o mapeamento das localidades específicas em que existe a presença de descendentes de pomeranos no Brasil, utilizamos o ArcGis (2013, versão 10.2), desenvolvido pela ESRI (*Environmental Systems Research Institute*).

O ArcGis é um *software* de geoprocessamento, ou melhor, um sistema de informação geográfica que permite a elaboração e a manipulação de mapas, de

modo que o usuário pode escolher uma base cartográfica e marcar as localidades desejadas. Esse programa foi muito útil para nós, pois a metodologia da SGL implica a identificação dos pontos de pesquisa em mapas (ou cartogramas, na linguagem da SGL). A identificação das localidades foi realizada para documentar a variação dos fenômenos linguísticos no espaço geográfico.

Destacamos que também etiquetamos parcialmente os nossos *corpora* escritos e o nosso *corpus* oral, identificando alguns dos principais verbos e substantivos e também casos de variações diatópicas do pomerano documentadas nos *corpora*.

Além disso, elaboramos uma listagem dos textos que compõem o conjunto dos *corpora* escritos, informando detalhes a respeito deles, tais como a origem, o local de coleta, a forma de obtenção, o título e a data (quando localizada). Por fim, realizamos o arquivamento de segurança do PK via *drive* virtual *on-line*.

6.4 Princípios considerados na compilação de corpora: representatividade, extensão e balanceamento

No âmbito da LC, para que um conjunto de dados linguísticos seja um *corpus*, existem alguns princípios que devem ser considerados. Segundo Berber Sardinha (2004, p.18-19), esses princípios ou critérios são: a origem, pois os dados devem ser autênticos e escritos [ou falados] por falantes nativos [da língua-alvo]; o propósito: os dados devem ser objeto de estudo linguístico; a composição: os dados devem ser escolhidos e colhidos com critério; a formatação: os dados devem ser legíveis por computadores; a representatividade: os dados devem ser representativos de uma língua ou de uma variedade linguística, o que na prática significa dizer que o *corpus* deve ser o maior possível; e a extensão: o material deve ser vasto para ser representativo.

Quanto à representatividade, consideramos que deve ser avaliada do ponto de vista dos objetivos aos quais os *corpora* se referem. Tagnin (2010) esclarece que não há um consenso sobre esse conceito, mas que o *corpus* deve ser representativo daquilo que pretende estudar. Assim, entendemos que o *corpus* deve atender minimamente à necessidade para o qual foi criado e que, nas palavras da referida autora, "cabe ao criador do *corpus* estabelecer os critérios que garantam essa representatividade" (TAGNIN, 2010, p. 360). A nosso ver, um dos principais critérios que garante a representatividade é a autenticidade dos dados coletados.

Desse modo, acreditamos que nem sempre a representatividade estará diretamente ligada à extensão do *corpus*, pois um *corpus* de especialidade, de uma área nova, como a Linguística Forense, por exemplo (segundo Berber Sardinha, 2000; 2004), não poderá ser muito extenso, até que novos materiais sejam produzidos de forma autêntica sobre o assunto. No caso de *corpora* dialetais, quando se tratar de variedades em processo de desaparecimento, dificilmente se obterá um *megacorpora*

em extensão, porém ainda assim se fará importante reunir material sobre o dialeto em um *corpus*, visto que se trata de uma forma de acervo, com material legítimo que poderá documentar o mesmo.

A respeito do critério do *corpus* ser o mais vasto e extenso possível, observamos que, atualmente, são considerados os objetivos de cada pesquisa antes de determinar o tamanho do *corpus*. Outrossim, se um *corpus* pretende salvaguardar, descrever e estudar um dialeto em processo de desaparecimento, seu tamanho provavelmente não será extenso, mais ainda assim será significativo e, desde que atenda aos objetivos propostos pelo pesquisador, poderá ser considerado, sim, um *corpus* representativo.

Em relação ao princípio de um *corpus* precisar ser balanceado, contendo cada um de seus textos um tamanho igual ou aproximado, observamos que, em casos específicos, não é possível atingir uma mesma quantidade de textos de cada tipo e gênero para a coleta e nem que esses textos sejam do mesmo tamanho.

Temos em vista que, no caso da compilação de *corpus* ou *corpora* de uma variedade em risco, qualquer texto que se obtenha é importante para permitir o conhecimento do léxico em questão. Observamos que, no nosso caso, o não balanceamento do conjunto de *corpora* que compilamos é diretamente proporcional à sua representatividade, ou seja, quanto mais diversificados são os conteúdos e tipologias das amostras, mais representatividade conseguimos obter acerca do pomerano, tendo em vista os objetivos de documentar o léxico, resgatar a variedade linguística e permitir sua descrição.

7 Resultados

Como resultado principal, conseguimos constituir um conjunto de dados linguísticos autênticos do pomerano no Brasil. E, a partir disso, compusemos um conjunto de *corpora*. Falamos em conjunto porque nos referimos aos *corpora* escritos e ao *corpus* oral como uma unidade, a qual denominamos de *Pommersche Korpora* – PK, que é a união dos *corpora* escritos do pomerano, nomeados individualmente como *Pommersche Korpora Escritos* – PKE e do *corpus* oral do pomerano, nomeado individualmente como *Pommersch Korpus Oral – PKO*.

Adotamos o singular por se tratar de um todo, um conjugado, o *Pommersche Korpora* (o conjunto dos *corpora* do pomerano), e grafamos o nome dos *corpora* com K, para remeter à forma alemã, conforme grafado pelo IDS – *Institut für Deutsche Sprache* (Instituto para a Língua Alemã).

Para obter uma forma de visualização geral dos tipos de arquivos que compõem os nossos *corpora* da VBP, dada sua diversidade, bem como para representar o modo como sua estrutura está armazenada no computador, elaboramos um organograma do PK, conforme apresentamos na Figura 2, abaixo.

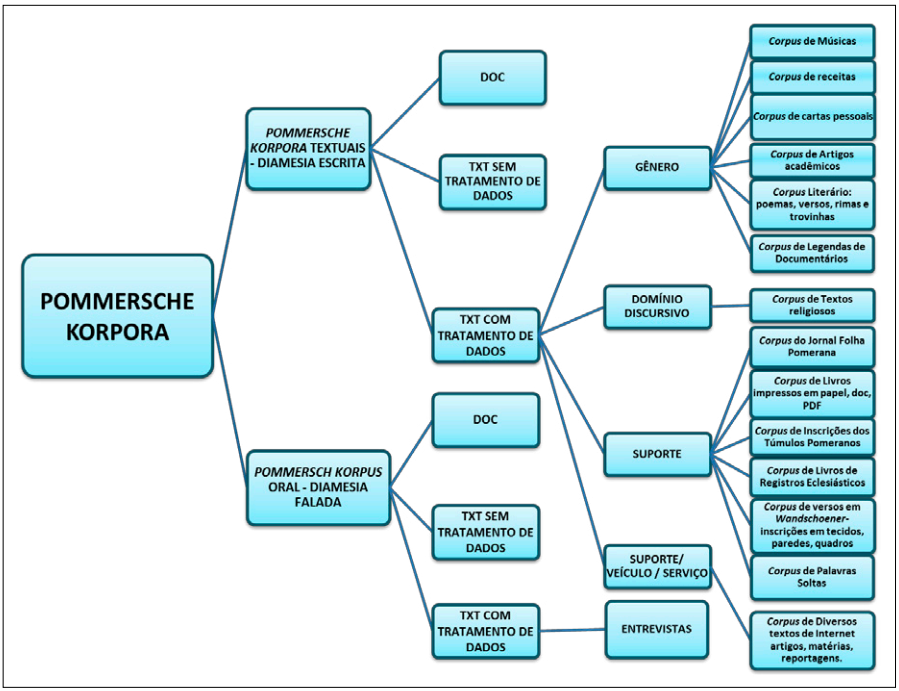


Figura 2 – Organograma com o detalhamento da estrutura do *Pommersche Korpora*
Fonte: Elaboração da autora

Conforme verificamos acima, o organograma permite observar como os *corpora* foram separados, organizados e desenhados dentro do conjunto, ou seja, a arquitetura "virtual- concreta" do PK.

Ao final, o PK contém quatorze *corpora* escritos, a saber: o (i) *Corpus* de Inscrições dos Túmulos Pomeranos; o (ii) *Corpus* de Livros de Registros Eclesiásticos; o (iii) *Corpus* de Cartas Pessoais; o (iv) *Corpus* de Receitas; o (v) *Corpus* do Jornal Folha Pomerana; o (vi) *Corpus* de Textos Diversos da Internet; o (vii) *Corpus* de Legendas de Documentários; o (viii) *Corpus* de Trabalhos Acadêmicos ; o (ix) *Corpus* de Textos Religiosos; o (x) *Corpus* de Músicas Pomeranas; o (xi) *Corpus* Literário Pomerano; o (xii) *Corpus* de Livros; o (xiii) *Corpus* de *Sprüche* Diversos; o (xiv) *Corpus* de Palavras Soltas; e um *corpus* oral, a saber o (xv) *Corpus* Oral de Entrevistas Interativas.

Apresentamos, a seguir, o Gráfico 1, com a distribuição dos *corpora*, incluindo o nosso *corpus* oral, dentro do conjunto do PK, a fim de permitir uma visão proporcional de todo o conjunto de dados que compilamos.

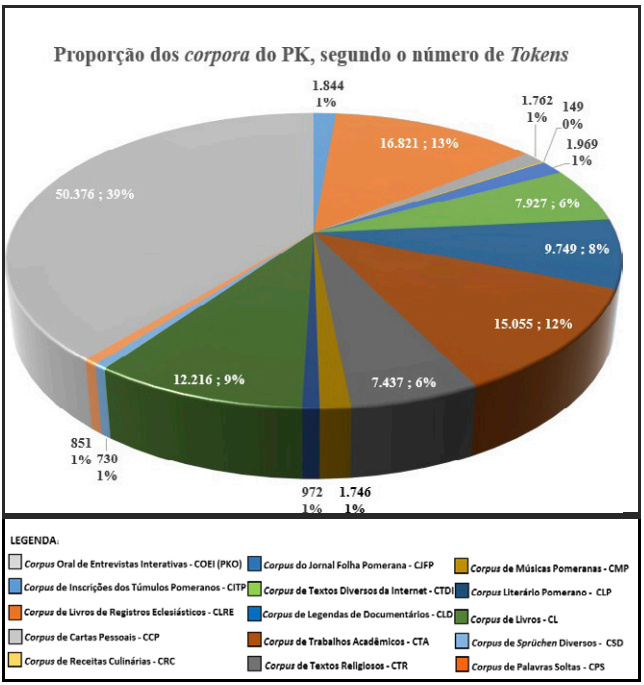


Gráfico 1 – Proporção dos *corpora* do PK, segundo o número de *Tokens*
Fonte: Elaboração da autora

Consideramos útil a elaboração de uma tabela, na qual foram reunidos os dados quantitativos dos *corpora* compilados, além de auxiliar na identificação dos dados do Gráfico 1, acima, no qual cada *corpus* pode ser identificado, pelas cores ou pelo número de *tokens*. Por isso, expomos a seguir, a Tabela 1.

Tabela 1 – Dados quantitativos parciais e totais do *Pommersche Korpora*

Corpus	***Tokens***	***Types***
1. *Corpus* de Inscrições dos Túmulos Pomeranos - CITP	1.844	640
2. *Corpus* de Livros de Registros Eclesiásticos - CLRE	16.821	1.558
3. *Corpus* de Cartas Pessoais - CCP	1.762	763
4. *Corpus* de Receitas Culinárias - CRC	149	81
5. *Corpus* do Jornal Folha Pomerana - CJFP	1.969	980
6. *Corpus* de Textos Diversos da Internet - CTDI	7.927	2.942
7. *Corpus* de Legendas de Documentários - CLD	9.749	2.470
8. *Corpus* de Trabalhos Acadêmicos - CTA	15.055	3.554
9. *Corpus* de Textos Religiosos - CTR	7.437	1.759
10. *Corpus* de Músicas Pomeranas - CMP	1.746	647
11. *Corpus* Literário Pomerano - CLP	972	531
12. *Corpus* de Livros - CL	12.216	3.861
13. *Corpus* de *Sprüche* Diversos - CSD	730	405
14. *Corpus* de Palavras Soltas - CPS	851	698
15. *Corpus* Oral de Entrevistas Interativas - COEI (PKO)	50.376	7.309
Pommersche Korpora Escritos -PKE	79.290	15.515
Pommersch Korpus Oral - PKO	50.376	7.309
***Pommersche Korpora* - PK**	129.666	20.672

Fonte: Elaboração da autora

O *Pommersche Korpora* Escritos – PKE totalizou 79.290 *tokens* e 15.515 *types*. O *Pommersch Korpus* Oral – PKO totalizou 50.376 *tokens* e 7.309 *types*. O conjunto de todos os *corpora* – o *Pommersche Korpora* ou PK – em todas as suas diamesias, modalidades e gêneros – constitui um acervo estatístico de 129.666 *tokens* e 20.672 *types*. Ele contém dados que permitem identificá-lo como um conjunto de *corpora* dialetais multilíngues contatuais, visto que reúne léxico pomerano autêntico, tanto do dialeto ativo quanto de textos históricos, contendo inclusive amostras linguísticas em pomerano, em alemão-padrão, em hunsriqueano e em português, todos dentro de um mesmo texto.

Portanto, o PK é um banco de dados linguísticos que constitui um conjunto de *corpora*, reunidos sob os critérios da Linguística de *Corpus*, com a finalidade de ser objeto de estudos da VBP. Trata-se de uma coletânea de textos de diversos tipos e autorias, provenientes de diferentes origens, suportes e localidades.

Quanto aos dois objetivos específicos a que nos propomos, foram (i) tentar identificar a sobrevivência ou o desaparecimento do dialeto nas regiões abrangidas pela coleta, principalmente nos Vales do Rio Doce (MG) e do Rio Pardo (RS) e (ii) comparar, brevemente, o pomerano coletado nas duas regiões principais que a pesquisa abrangeu (MG e RS). Podemos avaliar os resultados, observando que, quanto ao primeiro objetivo específico, afirmamos que, ao buscar material em pomerano, constatamos que essa língua ainda sobrevive (está em uso) no Vale do Rio Doce, em Minas Gerais. Nesse local, encontramos falantes ativos,

encontramos materiais escritos, bem como notamos que um sujeito-entrevistado, de 93 anos de idade, sabia como escrever palavras, por exemplo, *Schoope*, sem consulta ao dicionário.

Embora seja raro encontrar crianças e adolescentes falantes nessa região, notamos que algumas delas ainda sabem pequenas músicas, "trovinhas" e também algumas palavras, pois circunstanciaram durante as entrevistas algumas interrupções, com manifestações espontâneas. Ressaltamos que representantes dessa faixa etária não foram entrevistados, em virtude de não termos solicitado essa autorização ao CEP. O sistema de ensino público no leste de Minas Gerais não abrange o pomerano e, portanto, a tendência é que as novas gerações tenham dificuldades em manter a fala pomerana.

Já no Vale do Rio Pardo (RS), não conseguimos localizar falantes dessa variedade linguística para realizar entrevistas na localidade de Vera Cruz (onde já comprovamos a descendência majoritária de pomeranos). Quanto aos materiais escritos, só encontramos os registros e as inscrições em lápides, que datam de mais de um século atrás. Então, inferimos que o BAP não está tão preservado nessa região, embora o alemão-padrão esteja ativo na cidade vizinha, Santa Cruz do Sul. Contudo, esclarecemos que nossa avaliação não é conclusiva sobre esse ponto específico, pois precisaríamos voltar à localidade e insistir, inclusive na zona rural, na busca por falantes do BAP. Assim, quanto a identificar a sobrevivência do dialeto nessa região, não podemos fazer afirmações categóricas como a de que essa variedade não tenha sobrevivido em Vera Cruz. Apenas constatamos que, na zona urbana, o pomerano não está preservado, pela dificuldade em encontrar evidências atuais do mesmo na referida localidade. É necessário um aprofundamento das investigações na localidade.

Quanto ao segundo objetivo específico, podemos responder, principalmente com base nos dados do PKO, que, ao comparar as duas regiões abrangidas (todas as localidades de Minas Gerais e todas as localidades do Rio Grande do Sul pesquisadas), encontramos muitos itens lexicais idênticos, mas também alguns itens lexicais que apresentaram variações.

Como exemplos de itens pomeranos idênticos em Minas Gerais e Rio Grande do Sul, citamos a pergunta *Wie nennt man dies körperlich teil?* (Como se chama esta parte do corpo?, apontando para a unha do dedo da mão). A resposta em ambas as regiões foi *Finhanoogel*. Observamos que, para a unha do pé, o nome é outro, *Teehnanoogel*. Para a pergunta *Der Blume ... gelb, rund, mit einer Scheibe mit Samen in der Mitte? Wie nennt man das?* (Uma flor... amarela, redonda, com uma rodela de sementes no meio? Como se chama isso?), a resposta em ambas as regiões foi *Süünnblaume*, para designar o girassol. Dentre muitos outros casos em que coincidiram as respostas entre as regiões, temos *Komellatee* (para "camomila"), *Oossvoogel* (para "urubu").

Como exemplos de itens lexicais com variações entre Minas Gerais e Rio Grande do Sul, citamos a variação, principalmente na pronúncia da sílaba final, com a presença das consoantes procunciadas "nd" no Rio Grande do Sul. Para a pergunta *Wie nennt man dies körperlich teil?* (Como se chama esta parte do corpo?, apontando para o olho), em Minas Gerais, obtivemos a resposta *Oucho*, e, no Rio Grande do Sul, obtivemos a resposta *Ouchend*. O acréscimo do "nd" também foi observado em outros itens lexicais coletados no Rio Grande do Sul para o *corpus* oral.

Para a pergunta *Die Früchte ... kleiner als Orangen, die mit der Hand geschält werden, und in der Regel einen Duft auf der Hand lassen? Wie nennt man das?* (As frutas pequenas como a laranja, que se descascam com a mão, e, normalmente, deixam um cheiro na mão? Como se chama isso?), correspondentes em português à mexerica ou bergamota, obtivemos o seguinte resultado: em Minas Gerais, a resposta foi *Krawe* e, no Rio Grande do Sul, *Peechamota*.

Em Minas Gerais, notamos que alguns pratos da culinária brasileira e mineira tinham seus nomes em português desconhecidos pelos falantes, tais como curau de milho e mingau de milho. O mais próximo da denominação que conseguimos obter foi *miechpapa* (papa de milho), *miechmelka* (milho com leite) e *miechtouce* (milho doce).

Dessa forma, pudemos comparar o pomerano falado nas duas regiões de escopo da nossa pesquisa e avaliamos que existem algumas variações da fala pomerana entre as regiões, mas também existe uma norma pomerana, pois, para a maioria dos itens lexicais perguntados, as respostas foram iguais.

Esses foram os resultados alcançados por nossa pesquisa, de forma resumida. A seguir, trataremos brevemente da análise dos resultados, da classificação dos *corpora*, da apreciação das nossas hipóteses e a respeito de alguns fatos para os quais os *corpora* nos direcionaram.

7.1 Análise de Resultados

Realizamos a taxonomia ou classificação do PK, a fim de obter uma tipologia geral para o conjunto de *corpora* escritos – o PKE e para o nosso *corpus* oral – o PKO. A classificação, conforme os critérios de Berber Sardinha (2004) e Teixeira (2008), pode ser visualizada por meio da Figura 3, a seguir.

a. Tipo: dialetal e dialetal-regional. **b. Conteúdo**: multilíngue e multivarietal **c. Modos**: escrito. **d. Autorias**: de língua de "falantes nativos" do pomerano. **e.Seleções**: por amostragem (*sample corpus*), composto por porções de textos ou de variedades textuais, *corpus* geral de língua; **f.Tamanho e Representatividade**: pequeno (PKT: 79.290 *tokens*). Significativo e representativo; **g.Finalidade**: de estudo, de registro e de descrição do pomerano, e para futura composição de banco de dados para estudos em LC; **h.Tempo**: histórico e contemporâneo. Os diversos recortes temporais presentes nos *corpora* permitem estudos diacrônicos, com vários períodos de tempo não lineares. Permite estudos sincrônicos (que recortam um período de tempo específico) e estudos diacrônicos (que utilizam textos representativos de diferentes períodos de tempo). **i.Balanceamento**: não balanceado; **j. Integralidade**: composto de textos integrais, textos curtos, trechos de livros e alguns *corpora* de fragmentos como frases, versos e até palavras soltas (para resgatar o léxico); **k.Fechamento/Status**: estático, composição encerrada no contexto desta pesquisa. Aguardando validação; **l. Disposições internas**: comparável multilíngue contatual. **m. Nível de codificação**: parcialmente etiquetado.	**a. Tipo**: dialetal e dialetal-regional **b. Conteúdo**: bilíngue e/ou multilíngue – contém amostras conjuntas, dentro dos mesmos diálogos, de pomerano, alemão e português, o que é característico da variedade brasileira do pomerano e do fato dos pomeranos serem bilíngues em pomerano-português e/ou trilíngues em pomerano-alemão-português; **c. Modos**: falado, não-alinhado em som-escrita, composto pelas transcrições das falas; **d. Autorias**: de língua de "falantes nativos" do pomerano. **e. Seleções**: por amostragem (*sample corpus*), composto por porções de fala. **f.Tamanho e Representatividade**: pequeno (PKO:50.376 *tokens*), porém significativo e representativo do contexto atual do pomerano. O tamanho do PKO é considerável em se tratando das peculiaridades de compilação de *corpus* oral, do tempo de transcrição e da quantidade de compiladores (apenas um, no nosso caso); **g. Finalidades**: de estudo; **h. Tempo**: Contemporâneo. Permite estudos sincrônicos. **i.Balanceamento**: não balanceado **j. Integralidade**: composto de entrevistas integrais. **k. Fechamento/Status**: estático. Aguardando validação; **l. Disposições internas**: comparável multilíngue contatual. **m. Nível de codificação**: parcialmente etiquetado.

Figura 3 – Classificação do PKE e do PKO
Fonte: Elaboração da autora

Observamos, quanto ao tamanho, que a classificação dos *corpora* foi realizada de forma individual, por isso, o PKE e o PKO foram considerados de tamanho pequeno, por estarem dentro do parâmetro de até 80 mil palavras. Mas, se considerarmos o tamanho total do conjunto dos *corpora* (PKE + PKO), nosso PK é classificado como pequeno-médio, conforme parâmetro de Berber Sardinha (2004).

7.1.1 Testagem e avaliação das hipóteses de pesquisa

1) Primeira hipótese – O pomerano não é ágrafo – Confirmada

O pomerano não é e não foi ágrafo no Brasil, pois a coleta e a compilação de *corpora* demonstrou que existem formas de escrita anteriores à publicação do dicionário pomerano-português no Brasil, em 2006 (TRESSMANN, 2006). Os textos pomeranos coletados confirmam a hipótese já baseada na literatura pomerana preexistente à imigração, que data do século XVI. Além de havermos encontrado textos com traços dialetais dessa variedade, presentes em cartas e em

textos antigos, também encontramos bíblias em BAP no leste de Minas Gerais, entre descendentes de pomeranos, que ainda as utilizam para sua leitura.

Não nos compete julgar o mérito das formas de escrita, mas tão somente constatar que o fato de termos encontrado textos escritos com traços dialetais pomeranos comprovam que ele não é ou era ágrafo. Verificou-se que o que aconteceu no Brasil foi um processo de perda da cultura escrita, perda do conhecimento e da prática da escrita na forma germânica que os imigrantes trouxeram da sua terra de origem. Esse processo foi influenciado, em parte, pelo fim das escolas coloniais alemãs, a partir de sua proibição, em 1938, pelo Decreto-Lei nº 383. Também houve uma desconsideração de fontes primárias produzidas espontaneamente por pomeranos, no que tange a vestígios de escrita pomerana.

A própria constatação da diversidade de escritas pomeranas que encontramos durante o processo de coleta sugere a não agrafia. Essas formas de escrita vão desde materiais mais antigos que seguem traços da família germânica, próximas à grafia do *Hochdeutsch*, até outras que se distanciam com elementos da escrita moderna do pomerano, proposta por Tressmann (2006). E, ainda outras vezes, foram encontradas escritas que seguem a interpretação e a intuição dos falantes, aqueles que não tiveram oportunidade de alfabetização em escrita alemã, mas que grafam seu falar seguindo as regras ortográficas do português, como fez Kuhn Silva (2012).

Outro fato que reforça a ideia de que não se trataria de um caso de língua ágrafa é todo um conjunto de produção escrita, existente há vários séculos, se considerarmos, por exemplo, a *Barther Bibel*, publicada em 1588 e conhecida como a bíblia pomerana, bem como toda a literatura pomerana preexistente, já referida. Assim, acreditamos que os pomeranos que vieram para o Brasil já conheciam algum tipo de escrita, apesar de não padrão entre si e mesmo apresentando traços linguísticos que se diferenciassem do alto-alemão. Lembramos, ainda, do já referido sujeito-entrevistado do leste de Minas Gerais, que demonstrou conhecimento da escrita pomerana, ao grafar sua fala sem consulta a qualquer forma de dicionário.

2) *Segunda hipótese – Inexistência de falantes de pomerano na zona urbana – Negada*

No início da imigração, as comunidades pomeranas estabeleceram-se no campo, em áreas predominantemente rurais. Ainda hoje os pomeranos são considerados um grupo majoritariamente camponês e o censo demográfico das localidades pesquisadas apontaram para a maioria de população rural. Por isso, acreditávamos que os falantes de pomerano seriam encontrados exclusivamente na zona rural, podendo ser inexistente, em algumas localidades, a presença de falantes na zona urbana. Entretanto essa hipótese foi negada, visto que havia falantes de pomerano na zona urbana de Itueta (MG) e também na zona urbana

de Canguçu (RS). Portanto, consideramos que nossa segunda hipótese não se confirmou, pois não é possível generalizar e nem afirmar que só se encontram falantes dessa variedade na zona rural.

Os pomeranos estão atualmente muito ativos nas redes sociais, nos diversos meios de comunicação e muitos são os que ingressam no ensino superior. Portanto, acreditamos que a visão do pomerano como camponês tende a mudar, no sentido de que eles não estão mais tão isolados em áreas rurais, onde existe, inclusive, acesso à internet.

3) *Terceira hipótese – Contato do pomerano com outras variedades – Confirmada*

Nos pontos de pesquisa selecionados como principais – Vale do Rio Pardo (RS) e Vale do Rio Doce (MG) – houve, de fato, o contato de pomeranos com outras etnias. Os dados do PK apresentam evidências desse contato interdialetal e também evidências de variação lexical na comparação entre essas regiões.

No PKE, encontramos, na análise dos registros eclesiásticos de Vera Cruz (RS), dados das origens dos imigrantes, relatando inclusive o nome das localidades onde nasceram na Pomerânia. Os dados do *Corpus* de Registros Eclesiásticos confirmaram que entre os pomeranos havia silesianos, bávaros, renanos, embora em menor quantidade em relação aos pomeranos, que eram maioria. No entanto, esse fato já é um indício do contato interétnico e interdialetal.

Outro fato que confirma essa hipótese é a presença de itens lexicais não pomeranos e provavelmente hunsriqueanos, presentes no PKO, pois duas irmãs entrevistadas eram filhas de pai pomerano e mãe hunsriqueana. Ao responderem ao LSF/QSL, apresentaram algumas respostas não-pomeranas, como organizamos no Quadro 2, a seguir:

Quadro 2 – Respostas ao LSF/QSL com interferências de outras variedades germânicas

Respostas que apresentaram interferência de outras variedades germânicas				
Para o Português	**Em pomerano seria**	**Responderam**	**Hunsrückisch**	**Alemão-padrão**
Granizo ou chuva de pedra	Hoogel	Haagela	–	Hagel
Diarista	Dagelohner ou arbeira up Dag	schaft uf dag	schaft	Tagelöhner
Galinha	Huihna	Hinge	Hinkel/Hünkel	Huhn
Galinha sem rabo, cotó	Klitterhuihna	Schotterhingel	–	Ein Huhn ohne Schwanz
Fantasma ou Assombração	Spuicka, Spauck	Geschpenz	Gespenst	Geist
Feitiço ou macumba	Behexa	Das eingetan	–	Hexen, Zauber
Benzedeira	Besprecka	Braucher	–	Besprechen

Fonte: Elaboração da autora

E, ainda, um terceiro fato que evidencia a confirmação dessa hipótese é que tivemos que descartar duas entrevistas, pois um dos sujeitos-entrevistados falava hunsriqueano e outro falava suábio, e não pomerano, embora se considerassem pomeranos e estivessem inseridos em uma comunidade pomerana.

Por tudo isso, consideramos confirmada nossa hipótese de que houve contato de pomeranos com outras etnias alemãs em algumas regiões do Brasil, de modo que os textos dos *corpora* podem apresentar evidências desses contatos.

Observamos, ainda, que não foi possível encontrar nenhum texto ou fala pomerana, em nenhuma região, que tivesse o léxico totalmente diferenciado do alemão-padrão, visto que sempre havia algumas palavras que coincidiam tanto em pomerano quanto em alemão-padrão, tais como *Fluss*, *Blitz*, *Licht*, *Blind*, *Karre*, *Schnaps*.

4) *Quarta hipótese – Interferência do português e existência de uma variedade Brasileira do pomerano – Confirmada*

O contato dos descendentes de pomeranos no Brasil com o português e o prestígio desta língua como sendo oficial do país pode ter permitido uma interferência português-pomerano. Isso fez com que o pomerano falado atualmente no Brasil já represente uma variedade brasileira, cujo contato linguístico tenha propiciado usos com empréstimos lexicais do português no sistema pomerano:

encontramos indícios na formação de frases e na inovação de itens lexicais que demonstram essa relação de "mistura" pomerano-português.

Essa hipótese se confirma ao considerarmos alguns exemplos extraídos do PK, pois, nos textos e nas falas, ocorreram casos de "mistura" de pomerano-português ou pomerano-alemão-português, em um mesmo texto ou em uma mesma frase. Demonstramos, por meio do Quadro 3, a seguir, alguns exemplos.

Quadro 3 – Indícios do *Brasilianisch-Pommersch*, a variedade brasileira do pomerano

Indícios da Variedade Brasileira do Pomerano, o *Brasilianisch-Pommersch* no PK		
Forma encontrada	**Português**	**Alemão-padrão:**
Bolaspiela (3 ocorrências) Bolinhaspela (2 ocorrências)	Jogo bolinha-de-gude	Klicker
Klockacht (2 ocorrências) Kockacht (2 ocorrências)	Galinha D'Angola, cocá, cocár, angolista	Angola-huhn, Guinea-huhn
Pomdock (2 ocorrências)	Bodoque ou estilingue	Schleuder, Schlinge
Pomitta (6 ocorrências)	Palmito	Palmetto
Padariabrout, (3 ocorrências) Bäckariabrout, (1 ocorrência)	Pão-francês (Stuteln/stuuta)	Französisches Brot
Moskitt, Moskitte e Moskitta (15 ocorrências, itens lematizados)	Pernilongo	Stelzenläufer
Komella (6 ocorrências) Komellatee (3 ocorrências)	Flor de camomila	Kamille
Sarawä/Sarawee (5 ocorrências)	Sorigué, Saruê[1]	Stinktier

Fonte: Elaboração da autora

Além dos exemplos de substantivos mencionados no Quadro 3, percebemos também a presença de orações nas quais os sujeitos-entrevistados respondiam "misturando" português e pomerano, alguns exemplos podem ser visualizados por meio da Figura 4, a seguir.

<RHT-F-III-RS> 136. Wie nennt man das Objekt, das an den Wänden ist und dazu dient, die Lampe einzuschalten? „Ah eh dei, lichtshuolter oder lichtansticka **nois semp chamemo**“.

<RB-M-III-MG> QS-5. (2) Por gentileza, fale um pouco sobre o seu local de nascimento, de infância... Você sempre morou aqui? „[...] **nunca nóis** aah, weerde treffte deet, häwa hier, wära ümmer ick mit mien pappa in da meehr wohna, in Cassadiera deet [...]“.

<HG-M-II-MG> 90. Die Person, die mit der linken Hand isst, und alles mit dieser Hand tut? „**Canhoto num sei não,** dat **seria** ah mit der linker arm..., ah **só assim**, det schriewa, **eu tenho um irmão assim** nee, schriewa mit linker arm, hand“.

<HG-M-II-MG> 70. Kleine Insekten mit langen Beinen, die in der Nacht um unsere Ohren fliegen... „[...] **aí eu falo assim na roça, lá eu falo** nee moskitte lass uns nichta, keiner schloofa, **num deixa dormi**, moskitte [...] schloop, schloofe [...]“.

Figura 4 – Amostras transcritas do *Brasilianisch-Pommersch* na fala pomerana
Fonte: PKO - *Pommersch Korpus Oral* (BEILKE, 2016)

Acreditamos que a interferência do português na fala pomerana já esteja em um nível no qual os falantes transitam livremente entre as formas na sua comunicação, produzindo sentidos, nomeando objetos, falando um pomerano-brasileiro, por isso acreditamos que exista um *Brasilianisch-Pommersch*. Observamos que as transformações linguísticas ao longo do tempo permitiram um ambiente favorável ao surgimento de uma variedade com características próprias de escrita, pronúncias, usos e semas.

7.1.2 Outros resultados e hipóteses direcionados pelos *corpora*: esboços

Nossa pesquisa, a princípio, não estava voltada para a análise de *corpora*, mas apenas para a sua compilação e as questões nela envolvidas. A partir daqui, esboçamos algumas análises preliminares a respeito de fatos que foram percebidos por meio dos resultados alcançados.

Embora inicialmente tenhamos formulado hipóteses que limitariam nosso estudo a uma abordagem *corpus-based* (baseada em *corpus*), além de responder às hipóteses, os dados nos direcionaram para a observação de outros fatos linguísticos, por isso consideramos nosso estudo como sendo também uma abordagem *corpus-driven* (direcionada por *corpus*).

Ao realizarmos um experimento com uma pequena amostra extraída do PK, o *corpus* de música pomeranas, percebemos alguns padrões de uso frequentes. Analisamos as recorrências de uma mesma lexia por meio do agrupamento das

várias formas encontradas (lematização), além de checar os usos, por meio das linhas de concordâncias.

Ao gerar listas de palavras e concordâncias e, ao estudá-las, chegamos a alguns resultados acerca do léxico pomerano, fatos linguísticos para os quais os *corpora* nos direcionaram, conforme apresentamos nas alíneas a seguir.

a) Observamos que tanto na fala quanto na escrita pomerana, um dos itens mais frequentes é *deet*; no PK ele obteve 201 ocorrências. Convencionamos sua escrita como *deet*, forma encontrada na literatura. Porém, não foi possível converter todas as formas de escrita, pois houve casos de flexão do *deet*. Assim, consideramos mais produtivo para a análise fazer a lematização. Desse modo, foi possível observar melhor o comportamento do *deet* nas linhas de concordâncias que geramos a partir dele.

A Figura 5, a seguir, comprova a presença do *deet* no PK e demonstra sua lematização.

3.364	DEECH	1		1	1,05
3.365	DEECHA	1		1	1,05
3.366	DEEL	2		1	1,05
3.367	DEENE	1		1	1,05
3.368	DEENKEN	1		1	1,05
3.369	DEERA	4		3	3,16
3.370	DEES	4		1	1,05
3.371	DEESE	2		1	1,05
3.372	DEESEM	1		1	1,05
3.373	DEESEMBER	1		1	1,05
3.374	DEESES	1		1	1,05
3.375	~~DEEST~~	~~12~~		~~3~~	~~3,16~~ɔest[10]
3.376	DEET	201	0,17	26	27,37ɔet[183]
3.377	DEETCH	1		1	1,05
3.378	~~DEETS~~	~~4~~		~~2~~	~~2,11~~ɪeets[2]
3.379	DEEWIL	4		3	3,16
3.380	DEEWÜLL	5		1	1,05

Figura 5 – A partícula "*deet*" na Wordlist e lematizada no *Pommersche Korpora*
Fonte: Elaboração da autora

A Figura 6, a seguir, mostra as linhas de concordâncias de *deet* lematizado, ou seja, agrupada em todas as formas de flexão em que esse item ocorreu nos *corpora*.

Figura 6 – Linhas de concordâncias de *deet* lematizado
Fonte: Elaboração da autora

Observamos, por meio das linhas de concordâncias e dos contextos, que o *deet* parece ser usado com frequência como partícula enfática, fomos verificando fatos sobre o pomerano que não havíamos imaginado antes, e nem mesmo poderíamos prever que encontraríamos ao nos deparar com a tela de leitura vertical do PK.

Conjecturamos que o *deet* assemelha-se ao comportamento do *to do* em inglês e que é utilizado para fazer o passado simples. Inicialmente, pensamos que ele era usado sempre no final das orações, no entanto, ao observar os dados das linhas de concordâncias do PK, percebemos que essa não é uma posição fixa para ele, pois encontramos o *deet* em outras posições dentro das orações e, assim, verificamos que o *deet* não é utilizado de forma tão restrita como imaginávamos *a priori*.

Observamos que ele também é usado na segunda posição das orações, conforme linha 4 do *Concordance*, evidenciada na Figura 6 (linha 4), em "*Du deest a kint geboura*", "você terá um filho", onde embora o *deet* realmente funcione como partícula enfática, pois nesse caso não possui sentido sem o *geboura* (nascer), a sua flexão *deest* indica um tempo futuro.

Conforme visualizamos nas linhas de concordâncias, o *deet* é de fato muito utilizado como auxiliar para enfatizar uma ação ocorrida em um passado recente, como *gäwa deet* (deu), *moocka deet* (fez) e *kooma deet* (veio).

Outro exemplo de uso do *deet* em pomerano é *Dag waara deet*, expressão utilizada para falar que "amanheceu". Essa forma é característica da oralidade, pois se recuperarmos a expressão escrita teremos: *As dat dag waara deet*, que, em uma tradução literal, significa: "Assim o dia se fez fazer", mas com o sentido de informar

que o dia "amanheceu". Nesse caso o *deet*, apenas enfatiza a ação já ocorrida (no tempo verbal passado) realizado pelo verbo *waara*.

b) Observamos também no PK que alguns traços que distinguem o pomerano do alemão-padrão são recorrentes nas mesmas formas e posições, fazendo parecer que o pomerano é uma variação sistemática em relação ao alemão-padrão, pois existem padrões recorrentes de variação (nas mesmas formas e posições), por exemplo, entre os fonemas do alemão-padrão /ei/ e do pomerano /ie/ (*mein* – HD e *mien* – BP), exceto em itens que são iguais em alemão e pomerano.

Outro padrão observado é o grafema /z/: em posição inicial, em pomerano tem o fonema de /t/ (*tied* – BP e *Zeit* – HD) e o grafema /t/ duplo tem fonema de /r/, assim como ocorre no inglês (*better*, *berer*). Desse modo, em *Hochtiedsbirer*, por exemplo, substituindo os grafemas, conforme os casos acima expostos, teremos *Hochzeitsbitter* para a forma em alemão-padrão do mesmo substantivo, que em PTB é "convidador de/para casamentos". Observamos inúmeros casos de repetição desses padrões nos *corpora*; são fatos linguísticos que carecem de um estudo específico e detalhado.

Há também casos em que só a pronúncia pomerana é que varia, e quando são grafados diferentes para tentar captar essa variação de som é que fazem com que aparentem ser mais distantes do alemão do que realmente são. Explicamos, a seguir:

a) Toda vez que aparece o grafema /g/, quando está posicionado no início das palavras, ele pode ser identificado como aquele que se comporta como o fonema do /j/ no alemão-padrão. Ou seja, o /g/ inicial em pomerano teria o fonema de /i/ no português brasileiro. Exemplos: *Geschichte*, *Geld*, *Gelb* etc., que na VBP tem som de /i/ no início das palavras.

b) Em palavras como *Hoogel*, *Noogel*, *Voogel*, – o grafema /g/ sempre se comporta como o fonema /ch/ gutural do alemão-padrão, quando estão posicionados no meio da palavra.

c) Nas posições finais como em *lustig*, por exemplo, ocorrem variações na pronúncia de /g/, com o som de /k/, ou de /ch/ ou de /sch/. Essas foram as três variações sonoras que identificamos para /g/ nas entrevistas, porém ainda não presenciamos um /g/ nas posições finais com o fonema do /g/ como em PTB.

Esses casos exemplificados acima, observados por meio do PKO, merecem um estudo mais aprofundado. Acreditamos que, em alguns casos, não é a escrita que deva ser diferente, porque já existe um padrão de pronúncia convencional, ligando um determinado som para um sinal gráfico convencionado e que, comumente, nas línguas, variedades, variações e diatopias, as interpretações sonoras das grafias também variam. Nossa intenção não é discutir fonética e fonologia, mas, sim, levar à reflexão sobre alguns fatos linguísticos que encontramos nos nossos *corpora*, tendo em vista que todos os aspectos da língua são importantes e inseparáveis.

As percepções registradas acima indicam uma diversidade de estudos que podem ser feitos por meio da abordagem-metodologia da LC. A existência de tais fatos linguísticos sobressalta ao estudioso de *corpus* quando ele realiza uma leitura vertical das linhas de concordâncias em uma abordagem direcionada pelo *corpus*.

8 Considerações finais

Ao final, expusemos o conjunto dos *corpora* constituídos e sua arquitetura, demonstramos o potencial dos *corpora*, apresentamos dados estatísticos, fizemos descobertas sobre fatos linguísticos observados por meio de dados autênticos e esboçamos análises prévias sobre esses fatos para os quais fomos direcionados pelos *corpora*. Classificamos os *corpora*, avaliamos os resultados e também testamos nossas hipótese, além responder aos objetivos específicos e lançar um olhar sobre outras questões, dada a riqueza lexical contida nos dados.

Em suma, organizamos diversos *corpora,* considerados em sua totalidade de tamanho pequeno-médio ao reunir 14 *corpora* escritos provenientes de diversos gêneros e/ou suportes, e compilamos um *corpus* oral, constituindo um conjunto classificado como *corpora* dialetais multilíngues contatuais, acervo que compõe a unidade que denominamos *Pommersche Korpora* – PK. Portanto, podemos afirmar que obtivemos êxito ao cumprir nossa proposta, pois, segundo o levantamento estatístico do PK, coletamos 129.666 palavras corridas e 20.672 palavras distintas, de modo que pudemos alcançar uma considerável quantidade de amostras contendo o léxico pomerano.

Avaliamos que, por meio dos resultados alcançados em nosso trabalho de mestrado: conseguimos preencher a lacuna no que diz respeito à ausência de pesquisas acerca de regiões (como, por exemplo, a região leste de Minas Gerais) com presença de fala pomerana no Brasil e também à ausência de crítica ao trabalho de Tressmann (2008); apresentamos documentos históricos para abordar a imigração pomerana; debatemos a questão acerca de o pomerano ser língua ou dialeto; desenvolvemos procedimentos de tratamentos de dados orais dialetais e um método de consulta à bíblia pomerana de 1588; ampliamos o leque de fontes de coleta de dados linguísticos e fragmentos dialetais; fizemos transcrições de documentos históricos e fontes primárias (inclusive em letra gótica).

Quanto à análise qualitativa dos resultados alcançados, verificamos, com base nas evidências encontradas no PK, a existência de um pomerano que apresenta influências portuguesas, alemãs e dialetais (de outras variedades germânicas), por isso o chamamos de *Brasilianisch-Pommersch*, ou seja, uma variedade brasileira, a VBP.

Encerramos nosso trabalho lançando perspectivas futuras para o estudo do pomerano, com sugestões diretas de desdobramentos futuros possíveis por meio do trabalho iniciado, tais como: expansão dos *corpora* do PK; transformação do PK em

monitor; etiquetagem completa; estudo dos verbos e substantivos mais frequentes nos *corpora*; estudo dos dados do ponto de vista do contato de línguas; produção de um atlas linguístico parcial; produção de materiais didáticos com base em evidências empiricamente coletadas (evitando a formulação de frases artificiais para exemplificações); utilização dos dados como ferramenta auxiliar na tradução pomerano-alemão-padrão; levantamento de hipóteses sobre a sintaxe pomerana; estudos de perdas de alguns fonemas e grafemas nas posições finais das palavras; auxílio na produção de obras lexicográficas; análise sociolinguística com base nas respostas obtidas por meio da aplicação dos questionários; comparação com os dados de outros *corpora* (exemplo: *Corpus of Historical Low German* (*Corpus do Baixo-Alemão Histórico*) e/ou *Referenzkorpus Mittelniederdeutsch* (*Corpus* de Referência do Baixo-Alemão Médio).

Portanto, acreditamos e almejamos ter deixado uma contribuição para a compilação de outros *corpora* dialetais, pois as necessidades que nosso objeto impôs fez com que tivéssemos que desenvolver procedimentos metodológicos.

A LC foi o grande norte do nosso trabalho e a responsável pelo nosso produto, pois acreditamos que, de outro modo, não teríamos a visão do pomerano e nem a abrangência que pudemos alcançar com a utilização dessa metodologia-abordagem que funcionou como um "microscópio lexical". Os instrumentais da LC (estatísticas, linhas de concordâncias, leitura vertical, colocados etc.) permitem realizar micro e macroanálises, pois além de vermos minúcias em grande escala, podemos enxergar mais longe, como um telescópio, devido a grande quantidade de amostras que reúne e das quais nos aproxima. Em palavras pomeranas, poderíamos reiterar que "*Dai Teleskopa häwa ous joo veel weesa, wat man früüher ni vörher hät*" – "Os telescópios nos fizeram ver coisas que antes não percebíamos" –, metáfora que norteou nosso trabalho, pois traduz um dos grandes atributos da LC: ampliar nossa percepção sobre os fatos linguísticos e permitir a descrição do objeto investigado com riqueza de detalhes.

Referências

BERBER SARDINHA, T. B. *Linguística de* Corpus. Barueri, SP: Manole, 2004.

_____. Linguística forense. In: BERBER SARDINHA, T. *Pesquisa em linguí*stica de *Corpus* com *WordSmith Tools*. Campinas: Mercado das Letras, 2009, p. 69-81.

BIDERMAN, M.T.C. *Dicionário contemporâneo de português*. Petrópolis: Vozes, 1992.

_____. *Teoria linguística*. São Paulo: Martins Fontes, 2001.

CRISTIANINI, A. C. *Atlas semântico-lexical da região do Grande ABC*: um estudo geolinguístico. 2007. 635 f. Tese (doutorado em Semiótica e Linguística Geral). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007.

_____. Sociogeolinguística: uma abordagem para o estudo do léxico. In: SANTOS, I. P.; CRISTIANINI, A. C. (Org.). *Sociogeolinguística em questão*: reflexões e análises. São Paulo: Paulistana, 2012, p. 21-32.

DÜCK, E. S. *Witmarsum uma comunidade trilíngue*: Plautdietsch, Hochdeutsch e português. 2005. 152 f. Dissertação (mestrado em Linguística). Universidade Federal do Paraná, Curitiba, 2005. Disponível em: <http://acervodigital.ufpr.br/bitstream/handle/1884/2981/Disserta;jsessionid=DE71EFF071261AC1AD8D8F22CA244072?sequence=1> Acesso em: 16 fev. 2015.

_____. O trilinguismo no Colégio Fritz Kliewer de Witmarsum (Paraná). In: CELSUL, 8., Porto Alegre, 2008. *Anais*... Porto Alegre: [s/n], 2008. Disponível em: < http://www.leffa.pro.br/tela4/Textos/Textos/Anais/CELSUL_VIII/trilinguismo_col_fritz_kliewer.pdf >. Acesso em 04 fev. 2015.

ENCARNAÇÃO, M. R. T. *Estudo geolinguístico de aspectos semântico-lexicais nas comunidades tradicionais do município de Ilhabela*. 2005. 167f. Dissertação (mestrado em Linguística). Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, São Paulo, 2005.

FEITOSA, M. P. *Uma proposta de anotação de* corpora *paralelos com base na Lingüística Sistêmico-Funcional*. 2005. 177f. Dissertação (mestrado em Linguística Aplicada). Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

FERREIRA, C.; CARDOSO, S. *A dialetologia no Brasil*. São Paulo: Contexto, 1994.

GRANZOW, K. *Pomeranos*: sob o cruzeiro do sul, colonos alemães no Brasil. Vitória: Arquivo Público do Estado do Espírito Santo, 2009. Disponível em: <https://ape.es.gov.br/Media/ape/PDF/Livros/pomeranos_sob_o_cruzeiro_do_sul.pdf>. Acesso em 04 fev. 2015.

GONZALEZ, Z. M. G. *Linguística de* Corpus *na análise do Internetês*. 2007. 123 f. Dissertação (mestrado em Linguística Aplicada e Estudos da Linguagem). Pontifícia Universidade Católica, São Paulo, 2007.

HERRMANN-WINTER, R. *Plattdeutsch-hochdeutsches Wörterbuch für den mecklenburgisch- vorpommerschen Sprachraum*. Rostock: Hinstorff, 2003.

_____. *Sprachatlas für Rügen und die vorpommersche Küste*. Rostock: Hinstorff, 2013.

_____. Zur Geschichte der Dialektgeographie in Pommern. In: ASMUS, I. et al. (Hg.). *Geographische und historische Beiträge zur Landeskunde Pommerns*. Schwerin: Thomas Helms Verlag, 1998, S. 299-304.

KUHN SILVA, D. *Projeto Pomerando*: língua pomerana na Escola Germano Hübner. São Lourenço do Sul: Mais Cultura nas Escolas, 2012.

LEMNITZER, L.; ZINSMEISTER, H. *Korpuslinguistik*: eine Einführung. Tübingen: Narr, 2006.

MALTZAHN, G. M. *Família, ritual e ciclos de vida*: estudo etnográfico sobre narrativas pomeranas em Pelotas (RS). 2011. 151 f. Dissertação (mestrado em Ciências Sociais). Universidade Federal de Pelotas, Pelotas, 2011. Disponível em: <http://repositorio.ufpel.edu.br:8080/handle/123456789/1563>. Acesso em: 15 out. 2016.

MCENERY, T.; WILSON, A. Corpus *Linguistics an Introduction*. Manchester: Edinburgh University Press, 1996. Digital text – CD.

MELLO, H. Os *corpora* orais e o c-oral-brasil. In: RASO, T.; MELLO, H. (Org.). *C-Oral- Brasil*: *corpus* de referência do português brasileiro falado. Belo Horizonte: Ed. UFMG, 2012. CD1.

MÜNCHOW, A.; VORPAGEL, M.; WENDLER, H. *Bíbliha aventures*: aventuras da Bíblia em pomerano. Barueri: Sociedade Bíblica do Brasil, 2012.

NOVODVORSKI A.; FINATTO, M.J.B. Linguística de *Corpus* no Brasil: uma aventura mais do que adequada. *Letras & letras*, Uberlândia, v. 30, n. 2. p. 7-16, jul./dez. 2014. Disponível em: <http://www.seer.ufu.br/index.php/letraseletras>. Acesso em: 02 mar. 2016.

OMNIPAGE PROFESSIONAL. *Omnipage Professional*. Versão 17.0. Nuance Communications, 2009 – CD.

PARODI, G. *Linguística de* Corpus: de la teoría a la empiria. Madrid: Iberoamericana, 2010.

PESSOA, M. do S. *Ontem e hoje*: percurso linguístico dos pomeranos de Espigão D'Oeste - RO. 1995. 242 f. Dissertação (mestrado em Linguística). Universidade Estadual de Campinas, Campinas, 1995. Disponível em: <http://www.bibliotecadigital.unicamp.br/document/?code=vtls000101925&opt=3>. Acesso em: 07 nov. 2014.

SEIBEL, I. *Imigrante no século de isolamento*: 1870 – 1970. 2010. 350 f. Relatório (Pós-doutorado em Teologia). Escola Superior em Teologia, São Leopoldo, 2010.

SEYFERTH, G. A conflituosa história da formação da etnicidade teutobrasileira. In.: FIORI, Neide Almeida. *Etnia e educação*: a escola "alemã" do Brasil e estudos congêneres. Florianópolis: Editora da UFSC, 2003, p. 21-61.

SUBEXTRACTOR. *Subextractor*. Versão 1031. 2013. Disponível em: <https://subextractor.codeplex.com/releases/view/100949>. Acesso em: 13 set. 2014.

TAGNIN, S. E. O. Glossário de Linguística de *Corpus*. In: VIANA, V.; TAGNIN, S. E. O. (Org.). Corpora *no ensino de línguas estrangeiras*. São Paulo: HUB Editorial, 2010, p. 357- 361.

_____. *O jeito que a gente diz*. São Paulo: Disal, 2005.

TEIXEIRA, E. D. *A Linguística de* Corpus *a serviço do tradutor*: proposta de um dicionário de culinária voltado para a produção textual. 2008. 439 f. Tese (doutorado em Estudos Linguísticos e Literários em Inglês). Universidade de São Paulo, São Paulo, 2008. Disponível em: <http://www.teses.usp.br/teses/disponiveis/8/8147/tde-16022009-141747/pt-br.php>. Acesso em: 02 fev. 2016.

TOGNINI-BONELLI, E. Corpus *Linguistics at Work*. Amsterdam e Philadelphia: John Benjamins, 2001.

TRESSMANN, I. *Dicionário Enciclopédico Pomerano-Português*. Santa Maria de Jetibá: Farese, 2006.

_____. O pomerano: uma língua baixo-saxônica. Educação, cultura e sociedade. *Revista da Farese*, Santa Maria de Jetiba, v. 1, p. 10-21, 2008. Disponível em: <http://www.farese.edu.br/pages/biblioteca/revistafarese.php>. Acesso em 06 ago. 2015.

VOGT, O. P. Germanismo e Nacionalização em Santa Cruz do Sul, RS. *História Política de Santa Cruz do Sul*, Santa Cruz do Sul, v. 7, n. 2, p. 49-92, jul./dez. 2001.

VOLLMER, M. *Zur pommerschen Dialektlexikographie*. Kosegartens Wörterbuch der niederdeutschen Sprache älterer und neuerer Zeit: Jahrbuch des Vereins für niederdeutsche Sprachforschung. Greifswald: [s.n.], 2008.

VON BORSTEL, C. N. Sociolinguística: teoria, método e objeto em pesquisas in loco. *Revista Sociodialeto*, Campo Grande, v. 4, n. 12, p. 504-524, mai. 2014. Disponível em: <www.sociodialeto.com.br>. Acesso em: 09 fev. 2016.

WANGERINER-TREFFEN. Wangeriner Treffen in Bispingen, 2015. Programa. Disponível em: <http://www.kohls-online.de/pommern/26._Wangeriner_Treffen_Programm.pdf>. Acesso em: 03 mar. 2016.

WIESEMANN, U. *Manual Pomerano*. São Leopoldo, 2008. [Material não publicado].

WITTEN, H. *Barther Bibel*. Barth: Druckerei von Herzog Bogislaw XIII, 1588.

ZAVAGLIA, A.; WELKER, H. O que é léxico? O que é lexicologia? [*on-line*] Texto de divulgação científica. GTLEX – ANPOLL. 2008. Disponível em: <http://150.164.100.248/gtlexNovo/CMS/index.asp?pasta=gtlexnovo&path=20101229104440.asp&title=Lexicologia&id=54>. Acesso em: 01 abr. 2015.

Construções de tópico do português brasileiro falado em áreas indígenas em *corpus* especialmente reunido

Topic constructions from Brazilian Portuguese spoken in Indian areas in a specially collected *corpus*

Edivalda Alves Araújo
Wlianna Silva de Araújo

Resumo: Este artigo tem como objetivo identificar a realização de construções de tópico em dados orais do português brasileiro falado em áreas indígenas, com a intenção de levantar os tipos de construções produzidos, a frequência dessas construções e suas características sintáticas. O desenvolvimento deste trabalho amplia os estudos realizados sobre a sintaxe do português brasileiro por trazer dados de uma área pouco explorada nas pesquisas. O *corpus* em estudo foi reunido a partir de entrevistas e depoimentos espontâneos produzidos por falantes indígenas e por eles publicados no YouTube – *site* de partilhamento de vídeos. A distribuição geográfica desses falantes abrange seis Estados brasileiros: Amazonas, Bahia, Pernambuco, Mato Grosso do Sul, Paraná e Pará.

Palavras-chave: Sintaxe. Construções de tópico. Português brasileiro. Áreas indígenas. *Corpus* de fala.

Edivalda Alves Araújo – Professora de Língua Portuguesa na Universidade Federal da Bahia, doutora em Letras pela UFBA, vinculada ao grupo de pesquisa PROHPOR – edivalda@ufba.br.
Wlianna Silva de Araújo – Aluna de graduação em Letras da UFBA, orientanda do PIBIC/UFBA, vinculada ao grupo de pesquisa PROHPOR – wliannasaraujo@outlook.com.

Abstract: This paper aims to identify the realization of topic constructions in oral data from spoken Portuguese in Brazilian Indian areas, intending to raise the types of constructions produced, the frequency of these constructions and their syntactic features. The development of this study enlarges the researches evolved about the Brazilian Portuguese syntax in virtue of bringing data from an area rarely explored in these researches. The *corpus* under analysis was collected from interviews and spontaneous speech produced by Indian speakers and published by themselves in Youtube – video sharing site –. The geographical distribution of these speakers comprises six Brazilian states: Amazonas, Bahia, Pernambuco, Mato Grosso do Sul, Paraná and Pará.

Keywords: Syntax. Topic Constructions. Brazilian Portuguese. Indian Areas. Spoken *Corpus*.

1 Introdução

Os estudos linguísticos no Brasil relacionados à análise da modalidade oral têm privilegiado três grandes áreas de constituição de *corpora*: a) os de língua culta (a exemplo dos trabalhos do NURC); b) os de língua falada em regiões rurais, notadamente os do português falado em comunidades remanescentes de quilombolas ou de agrupamentos afro-brasileiros; e c) os de língua popular urbana, como o C-ORAL-BRASIL (UFMG – http://www.c-oral-brasil.org/) ou da Linguagem na Cidade (UNEB - http://www.linguagemnacidade.com.br). Trabalhos direcionados para a análise do português falado em áreas indígenas são realizados, principalmente, nas universidades do Norte ou Centro-Oeste do país, mas através de programas com a participação de algum pesquisador na área indígena. O *corpus*, levantado para este trabalho, no entanto, apresenta um diferencial por serem vídeos espontaneamente gravados pelos próprios indígenas e colocados à disposição na plataforma do YouTube, sem interferência de algum pesquisador ou pessoa externa à comunidade.

O levantamento do material desse *corpus* faz parte do projeto de pesquisa *O tópico em questão no português brasileiro*[1], cujo objetivo é rastrear as construções de tópico realizadas em várias áreas do país – urbana e rural – ou de várias comunidades – afro, indígena e urbana –, justamente para avaliar a tendência do português brasileiro indicada por alguns estudiosos como uma língua direcionada para o discurso.

Algumas pesquisas sobre a sintaxe do português brasileiro revelam que as construções de tópico se fazem muito presentes na modalidade oral da língua, como apontam as observações sobre o português urbano realizadas por Pontes (1987), que mostraram que tais construções, além de abundantes, são diversificadas entre si.

[1] Projeto de pesquisa registrado no Instituto de Letras da UFBA.

Dados levantados em *corpus* do português rural afro-brasileiro já foram levantados e apresentados no trabalho de Araújo (2009a), a partir das comunidades cadastradas no *Projeto Vertentes* (UFBA)[2] – Helvécia, Cinzento, Rio de Contas e Sapé –, em que também foram observadas diversificadas construções de tópico.

O português falado no *corpus* em análise, relacionado às regiões brasileiras indígenas, revela também a presença de construções de tópico, conforme exemplos abaixo:

(1) "**Rio Seco**, a gente viaja três dias de rabeta pra chegar, dormimos no mato pra chegar até nessas escola." (AM.V1.F1)[3]
(2) "**Barragem**, eu tenho muito medo que *isso* chega acontecer, uma preocupação muito grande pra nós daqui da comunidade, que mora aqui nesse lugar." (PA.V1.F1)

O sintagma nominal destacado em (1) é o tópico da sentença – um elemento locativo posicionado na periferia esquerda da frase, sem retomada interna na oração, e que é utilizado para apresentar o assunto sobre o qual se fará um comentário, como um orientador do discurso. Em (2), a construção de tópico se diferencia da primeira por haver um pronome demonstrativo – *isso* – que o retoma no interior da oração na posição de sujeito da encaixada. Sentenças de tópico como essas e outras, com diferentes características, foram encontradas nos dados orais do português brasileiro falado em áreas indígenas. Desse modo, neste texto, serão apresentadas algumas análises de construções de tópico marcado presentes nos dados coletados.

A análise aqui realizada parte da caracterização teórica do tópico apresentada por Pontes (1987) e Araújo (2006), autoras que discorrem sobre a questão do tópico e sobre sua presença no português. Para a discussão dos dados do português indígena, são aproveitadas as questões apresentadas nos trabalhos de Decat (1989), Galves (1998, 2001) nos quais há a abordagem sobre a sintaxe do português brasileiro e, principalmente, o de Araújo (2009a), que traz análises baseadas em descrições de *corpora* em relação ao comportamento do tópico em área rural.

É preciso ressaltar que, para a análise do tópico, este trabalho segue a perspectiva da teoria gerativa – que prescinde da análise de dados reais, conforme defende Adger (2003, p. 14):

> a *corpus* is of restricted use in another way, since the crucial cases to test our theory may not be available in the *corpus*. Finally, *corpora* have no information in them about the ungrammaticality of sentences, and such information is often crucial in theory development.

[2] Disponível em: <http://www.vertentes.ufba.br/>.
[3] A notação ao final de cada exemplo será explicada na seção 5 – a Metodologia.

Então, à primeira vista, sob uma perspectiva mais ortodoxa, o gerativista não se apoiaria numa análise de *corpora*. Tendências modernas, entretanto, têm enfatizado a busca por dados linguísticos reais, como se observa em Barbiers (2009, p. 1607):

> In the nineties of the past century, there was a growing consensus that syntactic theory had reached a stage in which it had become both possible and necessary to focus on microcomparative syntax, i.e., the study of closely related language varieties such as a family of dialects.

Alguns estudos sob a égide da teoria gerativa, portanto, têm procurado o apoio de material de *corpora* para que se possa afirmar ou fazer algum tipo de generalização sobre um fato ou fenômeno sintático característico de uma língua ou entre línguas. Sendo assim, a análise do *corpus* em dados do português falado em áreas indígenas poderá nos trazer evidência dos fenômenos sintáticos do português brasileiro, corroborando para os estudos dessa variedade do português, principalmente por estar na condição de segunda língua, convivendo com a língua indígena materna.

O presente texto está organizado em sete seções, sendo esta a introdutória. A seguir, na seção 2, há um panorama teórico sobre o tópico e sua caracterização semântico-discursiva, através da perspectiva da estrutura da informação, e sua caracterização sintática através da perspectiva da teoria gerativa; na seção 3, há uma abordagem sobre o direcionamento discursivo do português brasileiro; na seção 4, há uma discussão sobre o processo histórico de glotocídios que levou o português a ser falado em comunidades indígenas brasileiras e sobre a questão da educação bilíngue nessas áreas; em seguida, na seção 5, apresenta-se a metodologia da coleta, caracteriza-se o *corpus* reunido e a metodologia da análise dos dados; na seção 6, são apresentados os resultados das análises; na seção 7, discutem-se os tipos de tópico encontrados. Por fim, na seção 8, são feitas considerações sobre os resultados a que se chegou com essa pesquisa.

2 Sobre o tópico

O tópico, de modo geral, configura-se através da realização de um sintagma nominal à esquerda da oração, podendo estar ou não preposicionado. Em função da possibilidade de ativação de outras posições à esquerda na oração e que promove a ocorrência de outros elementos deslocados à esquerda, como o foco, por exemplo, é necessário estabelecer critérios para a identificação de um tópico para que este não seja confundido com outros elementos. Tal identificação pode ser estabelecida a partir de fatores semânticos e/ou discursivos e sintáticos.

2.1 O tópico e o caráter semântico-discursivo

Pontes (1987), seguindo Chafe (1976) e Li e Thompson (1976), adota a perspectiva discursiva e afirma que o tópico compõe uma rede de referências para as informações novas que serão ditas a seguir, funcionando, então, como um direcionador discursivo. Tal compreensão pode ser observada a partir do exemplo abaixo:

(3) "Então **essa aldeia**, *ela* foi criada por três mulheres por essa necessidade de mostrar pra nossos filho, pra nossos neto, a importância de nos valorizar." (BA.V1.F1)

O contexto discursivo presente em (3) refere-se à *aldeia* na qual os interlocutores se encontram e sobre a qual mantêm diálogo, então, o sintagma nominal em destaque na sentença – *essa aldeia* – consiste em uma informação conhecida pelo falante e pelo ouvinte. Essa informação será retomada com uma cópia pronominal dentro do comentário realizado sobre o tópico, que se manifesta em uma oração completa – *ela foi criada por três mulheres por essa necessidade de mostrar pra nossos filho, pra nossos neto, a importância de nos valorizar*.

Pontes (1987, p. 25) destaca que, no português, "não há restrição quanto ao tipo de elemento da S[entença] que pode ser tópico". Independentemente da função sintática que desempenhe, qualquer sintagma nominal pode figurar como tópico em uma sentença.

Envolvendo as três perspectivas – sintaxe, semântica e discurso –, Araújo (2006) argumenta que a caracterização do tópico não deve ser definida apenas por sua posição na sentença, uma vez que há outros elementos que podem ocupar tal posição, como o foco, por exemplo. Assim sendo, segundo a autora, em razão de o tópico estar localizado na camada discursiva da língua, na interface sintaxe-discurso, para identificá-lo e defini-lo, faz-se necessário observar a função semântica do elemento na cena discursiva, além de considerar as suas propriedades sintáticas.

Abordando a estrutura da informação, Araújo (2006) assume, seguindo a concepção de Lambrecht (1996), que essa está diretamente ligada aos princípios da pragmática do discurso e pode provocar interferências em todo o sistema gramatical. Lambrecht (1996) enfatiza que a estrutura da informação se relaciona diretamente aos princípios da pragmática do discurso – que trata do fato de o mesmo significado ser expresso por duas ou mais formas de sentenças – porque a relação entre uma dada forma da sentença e a função da sentença no discurso é diretamente determinada pelas regras e por princípios da gramática, ambos específicos de uma língua e universais.

Desse modo, em perspectiva pragmático-discursiva, Araújo (2009) afirma que o tópico pode ser caracterizado como um elemento referencial, ativo e identificável, responsável por ativar um elemento pressuposto no discurso. Posto que

a função do tópico relaciona-se à identificação do que está sendo dito entre os interlocutores, a sua caracterização deve considerar a sua inserção em um dado contexto.

Dessa maneira, discursivamente, o tópico pode ser considerado o elemento presente no momento da enunciação, mas cuja informação deve ser partilhada entre os interlocutores.

Assim, Araujo (2009, p. 234) caracteriza o tópico como "um sintagma nominal definido, identificável, ativo e referencial, realizado por um nome ou pronome", além de sustentar a possibilidade de retomada interna na oração ou não, mantendo um laço sintático ou apenas semântico com a sentença, a depender do tipo de construção.

O tópico deve ser definido por ser relacionado a algum referente, e, portanto, referencial. O fato de ser referencial implica que o seu referente deve ser identificável pelo falante/escritor e pelo ouvinte/leitor no processo de interação comunicativa, visto que o tópico faz parte da pressuposição do conhecimento – falante/escritor pressupõe que o ouvinte/leitor tem conhecimento sobre determinado assunto e, por isso, lança mão de construções de tópico. Sendo o tópico identificável, consequentemente, estará acessível no discurso, marcado com os traços de definitude e de especificidade, incluindo-se aí os casos em que o sintagma nominal indefinido esteja ancorado em elementos que o tornem específico. Desse modo, na análise aqui proposta, consideramos o tópico na periferia à esquerda da oração o sintagma nominal portando os seguintes traços semânticos: referencial, identificável, acessível, definido (ou indefinido) e específico (ARAUJO, 2009).

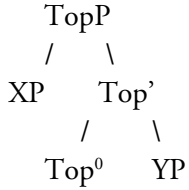
2.2 O tópico e a sintaxe

Para Rizzi (1997), uma das características que define o tópico refere-se à sua posição na sentença. Sendo assim, ele o define sintaticamente como um sintagma nominal, lexical ou pronominal, que se localiza na camada externa da oração, à esquerda da sentença, sendo seguido de um comentário sobre ele.

Rizzi (1997) considera que as construções de tópico envolvem movimento de último recurso para a periferia à esquerda para a satisfação de requerimentos, no caso do sistema A', de traços relacionados a critérios discursivos (Wh, Neg, Top, Foc, ...). Tais traços, em sua opinião, provocam o movimento dos elementos porque eles determinam a interpretação da categoria que lhes dá suporte e de seus constituintes imediatos.

Para dar conta do movimento dos tópicos, ou dos elementos da interface sintático/discursiva, Rizzi (1997) propõe que o sistema CP seja aberto em camadas que possam receber as projeções funcionais para aí movidas, de modo que possa fazer a interface entre um conteúdo proposicional (expresso por um IP) e a

estrutura superordenada (uma oração mais alta ou, possivelmente, a articulação do discurso, se se considerar uma oração raiz), como é o caso do tópico, que, para o autor, recebe a seguinte representação dentro do sistema C:

(4)


O núcleo Top⁰, por ser funcional, pertencente ao sistema do complementizador, projeta seu próprio esquema X-barra com a seguinte interpretação funcional: seu especificador (XP) é o tópico, seu complemento (YP) é o comentário. Top⁰ define um tipo de "predicação mais alta", uma predicação dentro do sistema de Comp.

Com essa representação, Rizzi (1997) considera o Tópico como uma projeção X-barra, e não como uma adjunção, porque assim se podem explicar os efeitos de localidade provocados por sua intervenção.

Rizzi (2004), entretanto, admite a possibilidade de o constituinte movido para a posição de tópico poder ser marcado por um traço, o que explica o seu movimento para satisfazer requerimentos da interface: ou da sintaxe com a morfologia ou da sintaxe com a semântica. Ao que parece, o tópico não é definido por traços sintáticos, mas por traços com propriedades interpretativas/discursivas, como, por exemplo, conexão com o discurso prévio. A existência desse traço com propriedades discursivas também foi proposta por Chomsky (2001), o OCC.

Há duas evidências de que os constituintes portam traços com propriedades discursivas: 1) há línguas que marcam morfologicamente os DPs que têm a função de tópico na oração, como o japonês (cf. KATO, 1989); e 2) as diferentes posições que um constituinte com a mesma função sintática pode ocupar na oração. É o que pode ser visto no exemplo em (5) com o sintagma [casa]. A depender da configuração na frase, esse constituinte pode satisfazer as seguintes propriedades: a argumental – "tema do verbo *comprar*" –, em (5a), e as de escopo/discursivas – interrogativa, tópico, foco, como em (5b-d):

(5) a. João comprou a casa.
b. Que casa João comprou? (interrogativa)
c. A casa, João a comprou. (tópico)
d. A CASA João comprou, não o carro. (foco)

Analisando-se (5), podemos perceber que o movimento é um dispositivo para alcançar a dualidade de interpretações: a do papel temático e depois a de escopo/discursiva. Ou seja: primeiro os elementos são conectados em sua posição-A, depois

são movidos para uma posição-A' para atender aos requerimentos interpretativos/discursivos, ou, como prefere Rizzi (2004), para atender aos requerimentos criteriais. Desse modo, as expressões linguísticas podem receber ambos os tipos de propriedades interpretativas, movendo-se de posições destinadas a propriedades do primeiro tipo (argumentais) para posições destinadas a propriedades do segundo tipo (discursivas). Rizzi (2004), entretanto, considera que o movimento de uma posição para outra é uma operação de último recurso, em que o nível de interface envolvido pode ser o da sintaxe-morfologia, interno ao sistema computacional estreito (como no caso do movimento de núcleo), ou a interface externa com a semântica, como no caso do movimento-A' para a periferia à esquerda – o tópico, por exemplo.

3 A questão do direcionamento discursivo do português brasileiro

A marcação do tópico pode ocorrer de forma diferenciada, a depender da configuração sintática da língua e da motivação pragmático-discursiva. Mas uma mesma língua pode dispor de vários tipos de construção de tópico, como o demonstram, em relação ao tópico deslocado à esquerda: Cinque (1990) e Benincà (2004), para o italiano; Raposo (1996) e Brito, Duarte e Matos (2003), para o português europeu; e Galves (1998), para o português brasileiro. Esses estudos têm também demonstrado que as línguas românicas não apresentam os mesmos tipos de construção de tópico: o que, às vezes, parece ser possível no português europeu não o é em italiano, ou, ainda, algumas construções de tópico que são comuns no português brasileiro não o são no português europeu.

O português brasileiro, por apresentar algumas construções de tópico diferentes das realizadas pelo português europeu, tem sido considerado por alguns autores como uma língua orientada para o discurso (cf. NEGRÃO, 1999) ou de proeminência de tópico (cf. PONTES, 1987; KATO, 1989; GALVES, 1998, 2001)[4]. Tais trabalhos, no entanto, não estão relacionados a algum *corpus* específico. Apresentam-se como reflexões acerca das produções orais do português brasileiro, observadas aleatoriamente, visto serem trabalhos de cunho gerativista.

Kato (1989) observa que as línguas de proeminência de sujeito estabelecem a predicação principal da sentença através da relação sujeito/predicado. Nas de proeminência de tópico, por outro lado, a predicação se dá através da relação entre um constituinte tópico e uma sentença (comentário). As predicações com tópico podem ter ou não um elemento a ele correferente dentro da sentença comentário.

[4] Essas autoras baseiam-se na tipologia apresentada por Li e Thompson (1979). Kato (1989), entretanto, concorda em parte com esses autores porque, para ela, a diferença entre os tipos de língua deve estar calcada em torno do tipo de sujeito que as línguas naturais possam selecionar, que é uma escolha paramétrica, e não em torno do tópico.

O fato de o português brasileiro permitir a ocorrência de objetos nulos possibilita a existência de construções com tópico em que o correferente aparece nulo, como será visto abaixo no exemplo em (7a).

Galves (1998a), dentro da mesma perspectiva, analisa que o português brasileiro apresenta características de línguas orientadas para o tópico porque não tem as mesmas propriedades que as línguas orientadas para o sujeito. Nestas, quando um constituinte é deslocado à esquerda, aparecem marcas que evidenciam a não correspondência entre a estrutura sintática e a estrutura argumental, como, por exemplo: no caso das orações ergativas, que são realizadas ou na voz passiva ou na voz média (cf. 6b, abaixo); ou no caso da topicalização, cuja marca pode ser um pronome resumptivo clítico (cf. 7b, abaixo). Nas línguas orientadas para o tópico, essas marcas não são necessárias. Isso pode ser visto nos seguintes exemplos, comparando-se o português europeu, com proeminência de sujeito, e o português brasileiro, com proeminência de tópico:

(6) a. O vaso partiu. (PB/*PE)[5]
b. O vaso partiu-se (#PB/PE)

(7) a. Os alunos, encontrei na saída da escola. (PB/*PE)
b. Os alunos, encontrei-os na saída da escola. (#PB/PE)

Em (6a), a frase é gramatical no português brasileiro apesar de não ter a marca formal das construções ergativas, o *se*. Tal frase não é gramatical no português europeu, uma vez que neste a presença do *se* é obrigatória, como se pode observar comparando-se (6a) com (6b). Em (7), temos uma construção de tópico, em que o constituinte deslocado à esquerda, *os alunos*, é obrigatoriamente retomado por um clítico resumptivo interno à oração, no português europeu (cf. 7b), mas não no português brasileiro (cf. 7a). Esses fatos, de acordo com Galves (1998), são evidências de que o português brasileiro está-se caracterizando como uma língua de proeminência de tópico, diferente do português europeu, considerada como de proeminência de sujeito.

Além desses casos de construção de tópico do português brasileiro exemplificados em (6a) e (7a), há outros em que se observa a concordância entre o sintagma deslocado à esquerda e o verbo, como o apontam Pontes (1986, 1987), Kato (1989), Galves (1998a, 1998b, 2001), respectivamente:

(8) A Sarinha está nascendo dentes.
(9) O carro furou o pneu.
(10) Essas casas batem sol.

[5] PB – abreviatura para português brasileiro. PE – abreviatura para português europeu.

As construções acima são consideradas por Galves (1998b) como construções de Tópico Sujeito, presentes no português brasileiro, mas não detectadas no português europeu.

São essas construções presentes nos exemplos em (6a), (7a) e (8)-(10) que levam alguns estudiosos a defenderem que o português brasileiro se distingue do português europeu em relação às construções de tópico, sendo, por isso, uma língua com proeminência de tópico e orientada para o discurso.

É necessário que se registre, entretanto, que a possibilidade de produzir construções de tópico é uma propriedade universal nas línguas. Não é uma característica exclusiva do português do Brasil. O que chama a atenção nesta variedade do português são os fatos sintáticos em torno dessas construções (ausência de clítico, por exemplo) ou ainda a alta frequência de determinado tipo de construção, o que leva os estudiosos a definirem a sua orientação para o discurso, e não para a sintaxe, diferenciando-se das outras línguas românicas.

4 O português brasileiro falado em áreas indígenas

Ao longo da história, as línguas autóctones faladas no Brasil foram, gradativamente, desaparecendo de suas comunidades por conta da forte presença e da imposição da língua oficial portuguesa. Rodrigues (2006) estima que no território brasileiro eram faladas cerca de 1.273 línguas autóctones antes da colonização, entretanto, devido ao extermínio das comunidades indígenas – como resultado de ações violentas ou mesmo em decorrência de epidemias –, houve uma série de glotocídios das línguas então faladas no território brasileiro. Dessas línguas indígenas, o último censo do IBGE (2010) aponta que apenas cerca de 280 línguas ainda são faladas no país, além de muitas delas estarem em situação de perigo.

Diante da realidade multilíngue que existia no território brasileiro, Rodrigues (2006, p. 152) afirma que o estabelecimento do português não foi uniforme em função de diversos fatores. Conforme o autor, durante cerca de 250 a 300 anos, duas línguas gerais de base indígena foram utilizadas como língua veicular nos primeiros momentos da colonização portuguesa. Entretanto, diversas ações foram adotadas para a instauração e consolidação da língua portuguesa nesse período. Assim, em favor de uma política colonial de homogeneização, o Marquês de Pombal tomou medidas para a erradicação das línguas gerais em benefício da língua portuguesa. Nesse sentido, Mattos e Silva afirma:

> Em 1757, com o Marquês de Pombal, se define explicitamente para o Brasil uma política linguística e cultural que fez mudar de rumo a trajetória que poderia ter levado o Brasil a ser uma nação de língua majoritária indígena [...] Pombal define o português como língua da colônia, consequentemente obriga o seu uso

> na documentação oficial e implementa o ensino leigo no Brasil, antes restrito à Companhia de Jesus, que foi expulsa do Brasil. (MATTOS; SILVA, 2004, p. 20-21)

A imposição do português não se deu apenas com relação aos falantes de línguas indígenas autóctones, mas também com relação aos africanos escravizados trazidos para o Brasil.

Atualmente, a língua portuguesa é hegemônica no país e, apesar das políticas de conscientização do multilinguismo brasileiro localizado e de cooficialização das línguas indígenas, o poder socioeconômico de que goza a língua oficial faz com que sua entrada e sua consolidação nas comunidades indígenas contemporâneas ainda sejam marcantes, principalmente pela necessidade de comunicação dos indígenas com a sociedade brasileira não indígena.

Desse modo, durante muitos anos sendo priorizado nas escolas indígenas em oposição às línguas autóctones e priorizado também pelos falantes aldeados, o português passou a ter uso predominante nas diversas comunidades indígenas. Tal fato é observado mesmo após as políticas linguísticas posteriores à Constituição de 1988, que asseguram o direito de realização do ensino indígena através das línguas autóctones, que deveria ocorrer através do ensino bilíngue, já que o ensino da língua portuguesa continua a ser obrigatório nos níveis básicos da educação formal.

Existe a consciência, nos dias atuais, da destruição que o português provocou nos valores indígenas e da necessidade de recuperação desses valores através da língua, como se observa no relato apresentado pelas autoras Ferreira e Souza (s/d):

> Esse processo de aculturação determinou a mudança de crenças religiosas, de valores cultivados pelas famílias indígenas, no trabalho na terra para o sustento, na confecção de artesanatos... Enfim, determinou a perda, quase que total, da cultura e dos valores terenas que, consequentemente, acarretou a produção de uma outra identidade, diferente daquela existente até então e que hoje os indígenas buscam resgatar, em especial através do uso e conhecimento de sua língua materna. Para tanto, têm acreditado que a escola (e, dentro dela a proposta de educação escolar indígena bilíngue) é o espaço onde esse resgate poderá acontecer.

A Lei de Diretrizes e Bases (LDB) da Educação Nacional, criada pelo governo federal em 1996, dedica dois capítulos ao ensino voltado para os índios. Conforme Cunha (2008, p. 150), essa Lei asseguraria a educação bilíngue, objetivando "proporcionar a eles a recuperação de suas memórias históricas, a reafirmação de suas identidades étnicas e a valorização de suas línguas e conhecimentos tradicionais". Com as políticas de ensino voltadas para as comunidades indígenas que surgiram nos últimos anos, o ensino do português passou a ser recomendado apenas para crianças que tivessem, no mínimo, 10 anos de idade, para assegurar a aquisição da língua autóctone como língua materna.

Apesar da escolaridade garantida na língua autóctone e, principalmente, com a obrigatoriedade de haver professores oriundos da própria comunidade indígena, em algumas comunidades, o ensino de língua portuguesa ainda se sobrepõe à língua autóctone, em função do seu caráter predominante nas relações externas, como indicam alguns autores:

> embora acreditem que se faz necessário o aprendizado da língua materna, enquanto "forma" e "marca" de uma identidade que se quer resgatar, e façam uso, em situações cotidianas na comunidade e no lar, os pais acreditam e exigem da escola o domínio da leitura e da escrita da língua portuguesa por considerarem-na de maior prestígio. (FERREIRA; SOUZA, s/d)

Ou é preferida pelos próprios professores, como se pode verificar nesse relato em relação às comunidades Sateré-Mawé:

> Constata-se, portanto, uma maior valorização da língua portuguesa pelos professores, pois esta é a língua que dizem gostar mais de trabalhar na escola, em detrimento da língua indígena, a qual praticamente não é trabalhada na escola, já que apenas um professor afirma procurar trabalhar "um pouco" da sua língua materna ao lado do português. Os demais professores justificam usar mais o português por conseguir explicar os conteúdos escolares "com mais precisão" nessa língua; além disso, argumentam que as crianças entendem "melhor" também em português. (FRANCESCHINI; CARNEIRO; SILVA, 2012)

Além disso, de acordo com os estudiosos da área, a educação indígena, no que se refere à língua portuguesa, também enfrenta problemas em relação à adequação do material que deve ser trabalhado nessas comunidades. De modo geral, não apresentam trabalho direcionado para as comunidades e, quando há ajustes, são mal feitos.

Essa realidade linguística "conflitante" em que vivem as comunidades indígenas vai promover construções linguísticas daí resultantes, exibindo a forma como os falantes compreendem e utilizam a língua portuguesa em seus enunciados. É dessa realidade "natural", exposta nos vídeos, que compusemos o *corpus* e extraímos os dados relacionados a esta pesquisa. Ou seja, os dados ora analisados são parte da produção linguística desses falantes de língua portuguesa – como língua materna ou como L2 – de algumas aldeias indígenas brasileiras e refletem o seu processo de produção nessa língua, principalmente em relação às construções de tópico.

5 Metodologia

O *corpus* reunido, e em observação, abarca a produção linguística de seis Estados brasileiros – Amazonas, Bahia, Pernambuco, Mato Grosso do Sul, Paraná

e Pará. É constituído de entrevistas e depoimentos espontaneamente produzidos pelos falantes indígenas e por eles disponibilizados no YouTube, *site* de compartilhamento de vídeos. Para a seleção dos vídeos, impôs-se uma restrição: duração mínima de 4 minutos. Desse modo, poderíamos ter a certeza de uma produção oral robusta, com mais dados linguísticos. Seguindo o critério de corte, foram selecionados, aleatoriamente, 64 vídeos diversificados.

O material coletado ainda está em processo de transcrição e, portanto, ainda sem descrição objetiva específica. Mas, para os objetivos deste trabalho de pesquisa na iniciação científica, foram retiradas e transcritas para a análise somente partes do contexto em que se pudesse detectar uma construção de tópico. Ao todo, foram transcritas 300 sentenças e detectadas 138 construções de tópico no geral. Considerando o tempo de cada vídeo, o *corpus* conta com cerca de oito horas de entrevistas e depoimentos produzidos pelos indígenas (Tabela 1).

As construções de tópico são comumente atribuídas a contextos discursivos em que há menor monitoramento linguístico, desse modo, os vídeos foram escolhidos, aleatoriamente, tendo em vista maior espontaneidade dos falantes no momento da produção linguística e, portanto, não apresentam rigor sociolinguístico. Estima-se, porém, que os falantes entrevistados estejam entre a faixa etária de 20 a 80 anos, conforme alguns revelam em seus depoimentos. Durante a coleta dos dados do *corpus*, foram observadas e registradas as regiões às quais pertencem os falantes entrevistados, conforme distribuição na primeira coluna da Tabela 1:

Tabela 1 – Distribuição dos vídeos por região

REGIÕES	Nº DE VÍDEOS	DURAÇÃO (por minuto)
Amazonas	5	36, 07
Bahia	4	27, 30
Pernambuco	5	47, 28
Mato Grosso do Sul	23	178,11
Paraná	2	19, 04
Pará	5	49, 01
Acre	9	59, 52
Ceará	6	34, 97
Tocantins	5	31, 65
Total	64	482, 95 (+/- 8 horas)

Fonte: Elaborada pelas autoras

Tais registros foram possíveis a partir dos títulos e das descrições que acompanham os vídeos ou mesmo através dos falantes que, como é possível verificar no seguinte exemplo transcrito do *corpus*, em que o falante explicita o local de onde fala:

(11) “Aqui, nós guerreiro, aqui, no Mato Grosso do Sul, nós somos bastante.” (MS.V21.F3)

A partir das descrições abaixo do vídeo ou da própria identificação feita pelos indígenas, construímos a seguinte etiquetagem para os exemplos: sigla do Estado, acompanhada com o número de identificação do vídeo nos dados coletados. Exemplo: MS.V21.F3 – MS: Mato Grosso do Sul; V21: vídeo 21; F3: falante 3.

A análise do *corpus* partiu da identificação dos casos de tópico marcado presentes nos dados, em conformidade com o panorama teórico já apresentado. Posteriormente, realizou-se análise sintática para a identificação de elementos de retomada, caso houvesse, e a classificação dos tipos de construção de tópico, agrupando-as segundo as diversas características sintáticas e semânticas divisadas.

Em função do baixo número de produção de construções de tópico (138), não foi possível usar ferramenta de busca ou de quantificação. Desse modo, todo o processo de análise ocorreu manualmente através da audição e transcrição do que foi identificado em um arquivo do Word. Para contagem, a depender do número de frases, utilizou-se o programa Excel.

6 Resultados

A análise dos dados contabilizou 138 construções de tópico no total, conforme exposto na Tabela 2:

Tabela 2 – Relação entre lugar e quantidade de ocorrências

REGIÓES	Nº DE OCORRÊNCIAS	DURAÇÃO (por minuto)
Mato Grosso do Sul	50	178,11
Pernambuco	18	47, 28
Pará	16	49, 01
Amazonas	15	36, 07
Bahia	11	27, 30
Paraná	11	19, 04
Acre	8	59, 52
Tocantins	6	31, 65
Ceará	3	34, 97
Total	138	482, 95 (+/- 8 horas)

Fonte: Elaborada pelas autoras

Pode-se verificar, na Tabela 2, a ocorrência de construções de tópico contabilizadas de acordo com a região. A região do Mato Grosso do Sul apresenta o maior número de construções de tópico, entretanto estima-se que seja em decorrência da quantidade superior de vídeos. Poderia ser essa a justificativa se as

outras regiões apresentassem homogeneidade ou proporcionalidade entre tempo e produção de construções de tópico. Não é, entretanto, o que se observa. É o que podemos verificar ao serem comparados os dados entre o Acre e o Paraná, por exemplo: o primeiro, com 59 minutos de gravação, exibe pouca produção de tópico; enquanto o segundo, com menos minutos (19,04), evidencia mais construções. Desse modo, em decorrência dos dados coletados, não é possível afirmar com precisão se há diferenças quantitativas na produção de tópicos por região. Além disso, é preciso salientar que as construções de tópico não precisam passar por crivo quantitativo. Tais construções estão subordinadas a dados do contexto e, consequentemente, sua ocorrência depende da pressuposição do conhecimento que o falante imagina ter o ouvinte sobre o assunto, conforme discutido na seção 2. Ressalta-se, portanto, que o quantitativo apresentado é um quadro panorâmico e generalizante dessas ocorrências. A quantificação foi realizada apenas como forma de análise comparativa entre os tipos de construções de tópico realizadas. Além disso, a quantificação servirá como ponto norteador para uma análise e controle mais geral sobre a frequência das construções de tópico no português do Brasil.

Logo abaixo, na Tabela 3, pode-se observar que, dentre as 138 construções de tópico, foram identificados nove tipos diferentes, a saber: Tópico Pronominal com Cópia em posição de sujeito, Tópico Pendente com Retomada, Tópico Pendente, Topicalização de Objeto Direto (TOD), ETop, Topicalização Selvagem, Tópico Cópia, Tópico Locativo e Tópico Nulo. Destaca-se que a identificação das construções de tópico e a análise "tipológica" dessas construções foi realizada a partir da classificação apresentada em Araujo (2006) e (2009a, 2009b).

Tabela 3 – Tipos de tópico

TIPOS DE TÓPICO	QUANTIDADE	%
Tópico com Retomada Pronominal (em posição de sujeito)	52	38
Tópico Pendente com Retomada	27	19
Tópico Pendente	16	12
TOD	12	9
Topicalização Selvagem	11	8
ETop	8	6
Tópico Cópia	6	4
Locativo	4	3
Tópico Nulo	2	1
Total	138	100

Fonte: Elaborada pelas autoras

Conforme apresentado na Tabela 3, as sentenças que apresentam Tópico com Retomada Pronominal em posição de sujeito foram o tipo de construção de tópico mais realizada, com 38% do total. Em sequência, aparece o Tópico Pendente com Retomada, com 19% das construções encontradas e, como o terceiro tipo mais

comum do *corpus*, há o Tópico Pendente (12%). Apenas duas sentenças com o tipo Tópico Nulo foram encontradas entre as 138 construções e, diferentemente dos resultados observados por Araujo (2009a) no português rural afro-brasileiro, não foram encontradas construções de Tópico Sujeito.

Exemplificações respectivas a cada tipo de tópico serão apresentadas na seção a seguir.

7 Discussão: tipos de tópico presentes nos dados

7.1 Tópico com Cópia Pronominal na posição de sujeito

Antes de tecer as considerações relacionadas a esse tipo de construção, é mister apresentar uma explicação sobre a sua análise separada do outro tipo: o Tópico Pendente com Retomada.

Na verdade, toda e qualquer construção de tópico que tenha uma retomada pronominal em alguma posição interna à oração constitui-se como um Tópico Pendente com Retomada. Essa retomada pode ocorrer com várias categorias – pronome pessoal, pronome demonstrativo, pronome nulo, SN lexical, entre outras. Mas as construções de Tópico Pendente com Retomada Pronominal na posição de sujeito constituem-se uma categoria diferenciada porque envolvem outros fatos sintáticos. É consenso entre os pesquisadores da sintaxe do português brasileiro que a frequência com que essas construções aparecem nas diversas ocorrências de oralidade traz evidências de que elas são resultantes de algum outro fenômeno sintático em curso nessa língua, como, por exemplo, a tendência ao preenchimento do sujeito.

As construções de Tópico com Retomada Pronominal na posição de sujeito, as de realização mais numerosa no *corpus*, caracterizam-se, segundo Araujo (2009a), por apresentar um sintagma nominal topicalizado, na periferia esquerda da sentença, e que é retomado com a inserção de um pronome pessoal em posição de sujeito da oração, como nos exemplos em (12), (13) e (14):

(12) "**Os Xapuri que era uma etnia indígena lá do município de Xapuri**, *eles* foram totalmente extinguidos assim [...]" (AC.V5.F1)

(13) "Então **essa aldeia**, *ela* foi criada por três mulheres por essa necessidade de mostrar pra nossos filho, pra nossos neto, a importância de nos valorizar." (BA.V1.F1)

(14) "Pra manter a escola lá é... é difícil, porque **nós**, *a gente* tá numa área muito longe, né?" (AM.V1.F1)

Nas sentenças acima, a retomada dos sintagmas nominais em posição de tópico foi realizada com os pronomes pessoais *eles*, *ela* e *a gente*. Segundo Galves (1998), a necessidade de preencher a posição do sujeito com um pronome se dá em virtude da perda do traço de [pessoa] no verbo no português brasileiro.

Exemplos dessas construções já eram notados em Pontes (1987), entretanto, Decat (1989) evidencia o fato de não ter encontrado construções de tópico com esse tipo de retomada nos dados diacrônicos por ela analisados – datados dos séculos XVIII, XIX e início do XX. Para a autora, a presença dessas sentenças no português do final do século XX seria o resultado de uma mudança no sistema gramatical da língua, uma vez que o aparecimento do pronome lexical correferente ao tópico da sentença reflete a perda de morfologia verbal de pessoa. Segundo Galves (1996, p. 20 *apud* DECAT, 1989, p. 132), houve "a perda do caráter pronominal da flexão que terá que ser substituída sistematicamente pelo pronome lexical". A frequente ocorrência de Tópico com Retomada Pronominal na posição de sujeito mostra que, ao menos nos dados aqui analisados, há certa sistematicidade na presença do pronome para marcar a pessoa na oração.

7.2 Tópico Pendente com Retomada

As sentenças que apresentam Tópico Pendente com Retomada constituem o segundo tipo de construção de tópico mais produzida pelos falantes indígenas. Esse tipo de construção se caracteriza por possuir ligação semântica com a oração e por, no interior dela, apresentar a retomada do tópico. Segundo Araujo (2006, p. 106), "esse tipo de tópico permite qualquer tipo de retomada interna à oração", podendo também ocorrer uma retomada por meio de uma categoria vazia, um elemento nulo foneticamente. Essas construções caracterizam-se também por manter com a oração uma fraca relação sintática, como nos exemplos abaixo:

(15) "**Marçal de Souza**, calaram a boca *dele*." (MS.V22.F1)

(16) "**Essas arma aí** ninguém não pode mais tá esquecendo *dela* e continuar fazendo e depois, mais na frente, pra tá ensinando os mais jovem… mais filho, neto e ninguém, ele falou que ninguém não vai mais esquecer dessas arma para gente tá fazendo." (AC.V9.F8)

(17) "**Esse pedacinho de terra, o qual nós estamos lutando, o qual nós estamos requerendo**, é *isso* que nós estamos querendo." (MS.V16.F1)

(18) "Então onde **o povo**, *a família dele* começou a chegar de novo ali, a ocupar aquele pequeno espaço." (MS.V19.F1)

Nessas construções, a retomada foi realizada com diferentes elementos: pronome possessivo (15) e (16), pronome demonstrativo (17), SN lexical (18).

Conforme Araujo (2009), não há restrições quanto à posição na qual a retomada do elemento topicalizado pode ocorrer. Assim, observa-se que, em (15), o tópico é retomado na posição de adjunto adnominal; em (16), a retomada é de um sintagma com função de objeto indireto; enquanto, em (17), o elemento em posição de tópico foi deslocado da posição de objeto direto da oração subordinada; e, em (18), a expressão que retoma o tópico preenche o sujeito da oração principal e na sentença.

7.3 Tópico Pendente

Nas sentenças com Tópico Pendente, o elemento topicalizado não tem relação sintática com o comentário que é feito sobre ele, pois não apresenta retomada interna à oração (ARAUJO, 2006, p. 99). Desse modo, o tópico mantém relação apenas semântica com a frase que o sucede. Foram contabilizadas 16 ocorrências de Tópico Pendente entre as construções de tópico analisadas, entre elas:

(19) "Aí tudo, então, **essa diferença** hoje a gente vê assim que hoje não é mais, graças a Deus né? hoje a gente é livre, aonde você vai você fala você leva sua cultura." (AM.V1.F1)

(20) "**O rio**, que nem a gente sabemos o que diz a constituição é que o usufruto exclusivo é nosso dentro do nosso território das terras demarcadas." (PE.V1.F3)

(21) "**A organização na aldeia** a gente faz sempre a reunião anual para fazer os planejamentos de dois, três, até cinco anos." (BA.V2.F1)

(22) "[...] **a perseguição**... e já foi matado vários liderança e continua sendo perseguido liderança que está mais principalmente buscando seu direito." (MS.V7.F1)

(23) "**As diretrizes**... os coronéis, na época, anoitecia, eles iam pra aldeia pra não deixar o povo se reunir." (PE.V2.F1)

Os elementos em posição de tópico nas sentenças acima são sintagmas nominais definidos que não possuem um lugar definido ou reservado no interior da oração. Assim, pode-se considerar que essas construções são caracterizadas principalmente por seu caráter discursivo, uma vez que os elementos topicalizados são utilizados pelos falantes como mecanismo para a introdução de um referente.

No exemplo em (23), observa-se o caráter recursivo do tópico: ocorrem dois tipos de tópico: *as diretrizes* (Tópico Pendente) e *os coronéis* (Tópico com Retomada Pronominal na posição de sujeito – *eles*).

7.4 Topicalização Selvagem

Os casos de Topicalização Selvagem constituem cerca de 8% das construções de tópico do *corpus*. Tais construções são identificadas pelo deslocamento de um sintagma preposicionado que, contudo, se apresenta desacompanhado da preposição:

(24) "**Muita ameaça** a gente passa aqui." (MS.V18.F1)

(25) "**Justiça Federal** nós confiávamos, mas enfim... nós fomos traídos por eles." (MS.V20.F4)

(26) "**A maioria de nossas terras indígenas** já saiu a portaria." (MS.V15.F1)

(27) "A gente precisa da mata, né? e **a mata** onde os pistoleiros estão impedindo a entrada né? das pessoas." (MS.V4.F1)

(28) "**A pressão** que a gente tá passando, eu fiquei muito preocupado porque quando eu fui ameaçado de morte… a pessoa parar assim numa via pública e apontar para você 'você vai morrer!' isso para mim foi muito forte." (AC.V7.F1)

Esse tipo de tópico está restrito a ocorrer em contextos de frase-raiz, sendo que o elemento deslocado geralmente se refere ao objeto indireto da sentença, como se pode perceber em (24) e (25). Entretanto, a classificação apresentada por Araujo (2009a, p. 241) abarca elementos com outras funções sintáticas e que não sejam necessariamente selecionados pelo verbo (complemento nominal, agente da passiva ou adjunto adverbial), desde que sejam regidos por preposição sem que esta esteja realizada na sentença, de natureza igual à apresentada em (26) e (27), por exemplo. Geralmente, nesses casos a preposição possui caráter funcional.

7.5 Topicalização de Objeto Direto (TOD)

As ocorrências de objetos diretos topicalizados, como em (29-32), representam 9% das construções de tópico. Em Pontes (1987, p. 18), já é possível observar, no português urbano oral, a presença de construções com o deslocamento do objeto direto, sem a retomada interna à oração. Apoiada em Cyrino (1996),

Araujo (2009a) aponta que há registro escrito de TODs desde o século XIX, estando presentes em todas as modalidades do português brasileiro. A mudança no sistema de clíticos com a perda do clítico acusativo de terceira pessoa estaria na base do surgimento de tais construções.

Para definir a TOD, Araujo (2009a, p. 235) lista quatro particularidades que a caracterizam: apresenta um objeto direto deslocado sem retomada clítica no interior da oração, o sintagma nominal em posição de tópico geralmente apresenta-se acompanhado por determinante definido, não sofre restrições de ilha e há a possibilidade de ocorrência em oração encaixada.

(29) "**Todos esses professores** tiraram daqui." (PA.V4.F2)

(30) "Então, **essa luta**, nós... é... cada vez mais estamos articulando, né?" (MS.V7.F1)

(31) "**A identidade dele** perde." (PE.V2.F1)

(32) "Chegando comigo, sentando é... do meu lado, explicando pra mim, **essa pessoa** eu adoro muito." (PA.V3.F1)

Decat (1989) observou nos dados diacrônicos realizações de pronomes clíticos para retomar o tópico das sentenças, entretanto, a autora registrou uma queda no número de retomadas por pronomes clíticos. Na verdade, a existência de clíticos acusativos de 3ª pessoa no português brasileiro está condicionada à transmissão escolar. Não é um dado espontâneo da língua oral. Em nenhuma das sentenças em observação houve, portanto, a recuperação do sintagma nominal por meio de pronome clítico.

Apesar de não haver retomada com pronome clítico, a posição interna à oração mantém com o elemento deslocado ligação autorizada no contexto discursivo, assim, nos exemplos acima, a partir da cena discursiva, pôde-se identificar os sintagmas nominais deslocados como sendo objeto direto das sentenças. Segundo Galves (2001, p.52), o tópico é sempre acessível no português brasileiro, de modo que não precisa de um clítico para mediar a relação entre o tópico e a posição de objeto, que se dá de forma direta.

7.6 ETop

As construções de tópico do tipo ETop se caracterizam pelo fronteamento de um sintagma nominal que é, necessariamente, um objeto direto e se apresenta por meio de um sintagma nominal nu, sem determinantes, tal como nas seguintes frases:

(33) "**Genocídio** a gente viveu desde a invasão da terra nossa, de todas as terras indígena." (MS.V4.F1)

(34) "**Bicho do mato** quase não tem mais." (MS.V23.F1)

(35) "**Morte** tá tendo quase todo ano né? isso não para realmente." (PA. V3.F1)

(36) "E **trabalho de grupo** eles fazem." (AM.V1.F1)

(37) "**Documentos** já... nós já mandamos várias vezes e nunca foi resolvido." (MS.V7.F1)

Conforme Araujo (2006, p.117), que adota proposta de Raposo (1996), esse tipo de construção não apresenta possibilidade de recuperação do tópico por meio de pronome clítico, mas pode conter retomada através de um elemento quantificador; entretanto, no que se refere aos dados do português indígena, esse tipo de retomada não foi encontrado. Foram identificadas apenas construções com esse tipo de topicalização sem retomada, como nos exemplos em (33-37).

7.7 Tópico Cópia

Houve a ocorrência de apenas seis sentenças com Tópico Cópia nos dados do português indígena, dentre os quais figuram os exemplos abaixo:

(38) "**Teodoro**, quando eu cheguei ainda lá na terra indígena do Poente Novo, mataram *Teodoro*." (MS.V22.F1)

(39) "É essa terra que a gente tá reivindicando. **Essa terra** que a gente hoje chora por *essa terra*." (MS.V6.F1)

(40) "Doce... **aquele milho**... tem *aquele milho* 'mainzêro', aquilo ali nós socava. Fazer a chipa, socava com batata, fica docinho." (PR.V1.F1)

Nas três sentenças, observa-se que o sintagma destacado que figura em posição de tópico é retomado no interior do comentário que o sucede através de uma cópia. Assim, Araújo (2009a, p. 237) caracteriza o Tópico Cópia basicamente por apresentar retomada interna à oração com a repetição do mesmo sintagma nominal topicalizado, com a sua cópia.

7.8 Tópico Locativo

Essas construções se caracterizam por apresentar em posição de tópico "um locativo, que funciona como adjunto ou de verbos existenciais ou de verbos tradicionalmente considerados intransitivos" (ARAUJO, 2009a, p. 242). Esse comportamento pode ser verificado no exemplo em (41):

(41) "**Na beira da estrada** nós temo cinco ano já, cinco ano, muito tempo já." (MS.V1.F1)

Nos dados do português afro-brasileiro, como destaca Araujo (2009a), o elemento deslocado é um sintagma preposicionado e se apresenta acompanhado da preposição, diferenciando-se dos casos de Topicalização Selvagem. Os dados do português falado em áreas indígenas apresentam apenas duas sentenças em que o sintagma topicalizado está acompanhado por preposição, como em (41). Ainda segundo Araujo (2009a), essa baixa realização é resultado de uma aparente tendência para o apagamento da preposição na oralidade.

Nas demais sentenças, há um elemento locativo que aparece sem que a preposição esteja realizada, caracterizando-se como Topicalização Selvagem, segundo classificação de Araujo (2009a).

7.9 Tópico Nulo

Os estudos apresentados por Araujo (2009b) apontam que o português brasileiro é uma língua que permite tanto o sujeito nulo quanto o tópico zero. Em conformidade com Hyams (1992) e Holmberg (2005), a autora evidencia a existência de um *pro*referencial a um tópico nulo. Desse modo, as construções de Tópico Nulo se caracterizam por apresentar em posição de tópico uma categoria vazia que só pode ser recuperada através de informações contextuais.

Dentre as construções de tópico encontradas nos dados do português falado por indígenas brasileiros, há apenas duas sentenças que apresentam um tópico foneticamente nulo, conforme exemplificação a seguir:

(42) "Depois, quando vem pra cá, vem meu marido, passou o carro em cima Ø." (MS.V1.F2)

(43) "**Fazendeiro** se $Ø_i$ vai lá pescar qualquer coisa *ele* já quer matar $Ø_i$."(PR. V1.F1)

Não é possível recuperar a categoria vazia em (42) ou (43) apenas pelos elementos que se apresentam sintaticamente, de modo que recorrer ao contexto

torna-se necessário. A falante se refere em (42) ao atropelamento que provocou a morte do seu marido. O sintagma nominal "meu marido", apesar de realizado na sentença, não é o sujeito do verbo *passou*. *O carro* é o sujeito do verbo passar. Ou seja, em uma construção canônica, teríamos: *o carro passou por cima do meu marido*. *Do meu marido*, complemento verbal, embora não esteja realizado, está acessível para a sua interpretação como tópico nulo. Assim, só é possível identificar que há um tópico onde indicado acima através dos dados do contexto.

Em (43), além de um tópico com retomada pronominal em destaque na sentença, observa-se a presença de um tópico nulo que controla a referência do objeto direto da oração. Sabe-se, pelo contexto do depoimento, que a categoria vazia se refere a qualquer indígena que tente entrar nas terras do fazendeiro para pescar, e não ao fazendeiro. Não há informações na sentença que permitam a recuperação de tal informação para garantir a interpretação do tópico nulo, de modo que o contexto é imprescindível.

Fundamentada em Huang (1984), Araújo (2009b, p. 63) afirma que, em construções desse tipo, há a topicalização de um elemento e, posteriormente, a supressão desse elemento, originando o tópico nulo. Mas a recuperação do conteúdo está subordinada ao contexto, uma vez que "não há elementos internos ao texto que possam preencher a sua referência", como já demonstrado nos exemplos acima.

8 Considerações finais

Perante a análise dos dados presente neste trabalho, pode-se afirmar que o português brasileiro falado em áreas indígenas apresenta, conforme os dados do *corpus* reunido, assim como as outras variedades da língua já estudadas, diversos tipos de construções de tópico, incluindo-se aí o Tópico Nulo, nem sempre observado. Ressalta-se, entretanto, que não foram encontradas construções de Tópico Sujeito, apontado, ao lado da Topicalização Selvagem, como caracterizador do português brasileiro por não se fazer presente em outras línguas românicas.

As construções de Tópico com Retomada Pronominal em posição de sujeito foram, com 38% das sentenças, as que tiveram mais ocorrências, diferenciando-se assim dos dados apresentados por Araújo (2009a), em que o maior número de ocorrências é de construções com Topicalização de Objeto Direto, no português afro-brasileiro. Assim, verifica-se no português indígena, com mais clareza, que a perda de morfologia verbal de pessoa faz com que a retomada do tópico por meio de um pronome lexical seja necessária na posição de sujeito, justamente para demarcar a diferença sintática entre tópico e sujeito.

O Tópico Pendente com Retomada foi o segundo tipo mais encontrado, com 19%, sendo seguido do Tópico Pendente, que constitui 12% das construções contabilizadas. Como suposto, de modo igual ao exposto por Araújo (2009a), os

casos de tópico com retomada não apresentaram nenhuma retomada através de pronome clítico, o que esperado em função da ausência dos clíticos de 3ª pessoa na oralidade espontânea do português brasileiro.

Há, em seguida, Topicalização de Objeto Direto (9%), Topicalização Selvagem (8%), ETop (6%), Tópico Cópia (4%) e, por último, Tópico Locativo (3%) e Tópico Nulo (1%). Todos esses apresentam características iguais ou semelhantes às já analisadas em outros estudos sobre o tópico nas diferentes modalidades e variantes do português brasileiro.

Desse modo, exceto pela ausência de Tópico Sujeito, de modo geral, a ocorrência das construções de tópico no *corpus* em estudo não se mostrou significativamente diferente do exposto nas análises dos dados do português urbano ou do português rural afro-brasileiro, confirmando as características e tendências sintáticas do português brasileiro em relação às construções de tópico, inclusive em relação ao número superior de Tópico com Retomada Pronominal – em posição de sujeito, principalmente, e no interior da oração.

Ao mesmo tempo, o desenvolvimento deste estudo traz à cena das pesquisas linguísticas a língua portuguesa falada nas áreas indígenas, exibindo a sua produção, particularmente, em relação às construções de tópico, cujos dados podem ser somados aos estudos já realizados no português urbano, no português afro-brasileiro e no português do semiárido baiano. O levantamento de dados em áreas geográficas diferentes e diversificadas nos dá suporte para defender e definir as características dessas construções no português brasileiro, contribuindo com dados para a formação da história dessa língua, conforme os estudos desenvolvidos pelo PROHPOR – Programa para a História da Língua Portuguesa.

Quanto ao *corpus*, assim que for realizado todo o trabalho de descrição e de catalogação dos dados, deverá ser disponibilizado na íntegra para consulta no *site* do PROHPOR.

Referências

ADGER, D. *Core syntax*: a minimalist approach. Oxford: Oxford University Press, 2003.

ARAÚJO, E. A. *As Construções de tópico do português nos séculos XVIII e XIX:* uma abordagem sintático-discursiva. Tese (doutorado). Instituto de Letras, Universidade Federal da Bahia, Salvador, 2006.

_____. As construções de tópico. In: LUCCHESI, D.; BAXTER, A.; RIBEIRO, I. (Org.). *O português afro-brasileiro*. Salvador: EDUFBA, 2009a, p. 231-250.

_____. Tópico. In: LOBO, T.; OLIVEIRA, K. (Org.). África à vista. Salvador: EDUFBA, 2009b, p. 50-69.

BARBIERS, S. Locus and limits of syntactic microvariation, *Lingua*, n. 119 (11), p. 1607-1623, 2009.

BENINCÀ, P. *The left periphery of medieval romance*. Disponível em: <http://www.humnet.unipi.it/slifo/2004vol2/Beninca2004.pdf>.

BRITO, A. M.; DUARTE, I.; MATOS, G. Frases com tópicos marcados. In: MIRA MATEUS et al. *Gramática da língua portuguesa*. 5. ed. rev. e aum. Lisboa: Caminho, 2003, p.489-502.

CINQUE, G. *Types of Ā-dependencies*. Linguistic Inquiry Monographs. London, England: MIT Press, 1990.

CYRINO, S. M. L. Observações sobre a mudança diacrônica no português do Brasil: objeto nulo e clíticos. In: KATO, M. A.; ROBERTS, I. (Org.). *Português brasileiro*: uma viagem diacrônica. 2. ed. Campinas: Editora da UNICAMP, 1996, p. 163-184.

CHOMSKY, N. A. Beyond Explanatory Adequacy. *MIT Occasional Papers in Linguistics*, n. 20. Cambridge: MITWPL, 2001.

CUNHA, R. B. Políticas de línguas e educação escolar indígena no Brasil. *Educar*, Curitiba, n. 32, p. 143-159, 2008.

DECAT, M. B. N. Construções de tópico em português: uma abordagem diacrônica à luz do encaixamento no sistema pronominal. In: TARALLO, F. (Org.). *Fotografias Sociolinguísticas*. Campinas: Pontes, 1989, p. 113- 139.

FERREIRA, F. M. N. S.; SOUZA, C. C. de. *Educação escolar indígena*: língua materna x língua portuguesa. Disponível em: <http://www.neppi.org/gera_anexo.php?id=488>.

FRANCESCHINI, D. do C.; CARNEIRO, D. de S.; SILVA, J. de O. dos S. da. O ensino de língua portuguesa em comunidades Sateré-Mawé. *Anais do SIELP*, Uberlândia, v. 2, n. 1, 2012. Disponível em: <http://www.ileel.ufu.br/anaisdosielp/wp-content/uploads/2014/06/volume_2_artigo_099.pdf>.

GALVES, C. Tópicos, sujeitos, pronomes e concordância no português brasileiro. *Cadernos de Estudos Linguísticos*, Campinas, n. 34, p. 19-31, jan./jun. 1998.

_____. A sintaxe do português brasileiro. In: _____. *Ensaios sobre as gramáticas do português*. Campinas: Editora da Unicamp, 2001, p. 43-59.

HOLMBERG, A. Is There a Little Pro? Evidence from Finnish. *Linguistic Inquiry*, n. 36, p. 533-564, 2005.

HUANG, C.-T. J. On the Distribution and Reference of Empty Pronouns. *Linguistic Inquiry*, v. 15, n. 4, p. 531-574, 1984.

HYAMS, N. A Reanalysis of Null Subjects in Child Language. In: WEISSENBORN, J.; GOODLUCK, H.; ROEPER, T. (Ed.). *Theoretical Issues in Language Acquisition*. New Jersey: Lawrence Erlbaum, 1992, p. 249-267

KATO, M. A. Tópico e sujeito: duas categorias na sintaxe? *Cadernos de Estudos Linguísticos*, Campinas, n. 17, p. 109-131, 1989.

KAYNE, R. S. *Comparative Syntax*. New York: New York University, 2012.

LAMBRECHT, K. Information structure and sentence form: topic, focus and the mental representations of discourse referents. *Cambridge Studies in Linguistics*, 71, 1996.

MATTOS E SILVA, R. V. *Ensaios para uma sócio-história do português brasileiro*. São Paulo: Parábola Editorial, 2004.

NEGRÃO, E. V. *O português brasileiro:* uma língua voltada para o discurso. Tese de Livre-Docência. São Paulo: USP, 1999.

PONTES, E. S. L. Da importância do tópico em português. In: _____. *O tópico no português do Brasil*. Campinas: Pontes, 1987, p. 11-40.

RAPOSO, E. *Towards a unification of topic constructions*. UCSB. 1996. Texto inédito. s/r.

RIZZI, L. The fine structure of the left periphery. In: HAEGEMAN, L. (Org.). *Elements of grammar*: handbook of generative syntax. London: Klwer Academic Publishers, 1997, p. 281-337.

RIZZI, L. *On the Form of Chains*: Criterial Positions and ECP Effects. 2004. Disponível em: <http://www.ciscl.unisi.it/doc/doc-pub/rizzi>.

RODRIGUES, A. As outras línguas da colonização do Brasil. In: CARDOSO, S. A.; MOTA, J. A.; MATTOS E SILVA, R. V. (Org.). *Quinhentos anos de história linguística do Brasil*. Salvador: Secretaria da Cultura e do Turismo da Bahia, 2006, p. 143-161.

Para a segmentação automática de fronteira na fala espontânea a partir de parâmetros prosódicos

For automatic boundary segmentation in spontaneous speech from prosodic parameters

Bárbara Helohá Falcão Teixeira
Plínio Almeida Barbosa
Tommaso Raso

Resumo: Este artigo apresenta modelos de um detector automático de fronteiras prosódicas para a fala espontânea baseado em parâmetros acústicos. A ferramenta Praat usa parâmetros acústicos associados à percepção humana como critério de referência para detecção automática das fronteiras prosódicas. Uma amostra de 11 textos de fala espontânea foi segmentada em unidades prosódicas por 14 anotadores. As fronteiras prosódicas percebidas foram marcadas como fronteiras prosódicas terminais e fronteiras não terminais. Um *script* do Praat extraiu então 111 parâmetros acústicos em cada fronteira indicada por pelo menos 7 anotadores. Dois métodos de classificação estatística foram utilizados para gerar modelos com subconjuntos de parâmetros acústicos, que poderiam funcionar como preditores de fronteiras prosódicas. Os resultados iniciais mostram sucesso relativo para detectar automaticamente fronteiras prosódicas percebidas pelos anotadores humanos.

Palavras-chave: Fronteira prosódica. Segmentação automática. C-ORAL-BRASIL.

Bárbara Helohá Falcão Teixeira – Mestranda em Linguística Teórica e Descritiva na Universidade Federal de Minas Gerais – barbaraheloha@gmail.com.
Plínio Almeida Barbosa – Professor na Universidade Estadual de Campinas, doutor pelo Institut de la Communication Parlée – pabarbosa.unicampbr@gmail.com.
Tommaso Raso – Professor na Universidade Federal de Minas Gerais, doutor pela Università Di Napoli Federico II – tommaso.raso@gmail.com.

Abstract: This paper presents models of an automatic prosodic boundary detector for spontaneous speech based on acoustic parameters. The tool uses acoustic parameters associated with human perception as a reference criterion for automatic detection prosodic boundaries. A sample of 11 spontaneous speech texts was segmented into prosodic units by 14 annotators. The perceived prosodic boundaries were tagged as terminal prosodic boundaries and non-terminal boundaries. A Praat script then extracted 111 acoustic parameters at each boundary indicated by at least 7 annotators. Two statistic classification methods were then used to generate models with subsets of acoustic parameters, which could work as predictors for prosodic boundaries. Initial results show relative success to detect automatically prosodic boundaries perceived by human annotators.

Keywords: Prosodic boundary, Automatic segmentation, C-ORAL-BRASIL.

1 Introdução

O fluxo da fala é segmentado em pequenas unidades entonacionais determinadas por fronteiras prosódicas, tanto por motivos cognitivos, quanto por motivos linguísticos. Este trabalho investiga os parâmetros fonético-acústicos associados às fronteiras prosódicas em dados de fala espontânea em português brasileiro (PB). Um trabalho como este tem como objetivo possibilitar a criação de um *script* do Praat[1] (BOERSMA; WEENINK, 2017) para detecção automática ou, pelo menos, semiautomática de fronteiras prosódicas. O *script* contribuirá para o melhor entendimento da segmentação da fala, pois utilizará como critério de referência para a detecção automática das fronteiras a percepção humana em conjunto com os parâmetros fonético-acústicos. O *script* também auxiliará na compilação de *corpora* de fala espontânea, porque poderá tornar o processo de segmentação da fala mais rápido, poupando-se simultaneamente tempo e esforços humanos, o que pode ser visto como uma contribuição para a linguística de *corpus* em geral.

2 Referencial Teórico

A fala tem como característica a junção natural dos segmentos vocálicos e consonantais durante a comunicação. Entretanto, a análise do conteúdo segmental é insuficiente para a análise da diamesia falada, tendo em vista que os fones podem ser agrupados em unidades distintas, de acordo com os propósitos comunicativos. Argumenta-se que uma análise da fala baseada puramente no conteúdo segmental é limitada, porque ela identifica apenas os segmentos, desconsiderando as fronteiras ao longo dos enunciados.

[1] O Praat é um *software* utilizado para análise acústica da fala. Para mais detalhes, veja-se o endereço eletrônico: <http://www.fon.hum.uva.nl/praat/>.

Em geral, a segmentação da fala em unidades é determinada por meio da prosódia, mais especificamente por meio de fronteiras, que assinalam o início e o término do enunciado. Este trabalho tem como foco investigar os parâmetros fonético-acústicos associados às fronteiras prosódicas em dados de fala espontânea monológica em português brasileiro (PB). Para investigar esse tema, a análise aqui desenvolvida utilizou uma metodologia de base acústica em conjunto com a percepção humana de fronteiras para investigar os parâmetros fonético-acústicos relevantes tanto para a produção quanto para a percepção de fronteiras ao longo do fluxo da fala.

É de suma importância segmentar o fluxo da fala em pequenas unidades entonacionais marcadas por fronteiras tanto por motivos cognitivos quanto por motivos linguísticos. Uma das causas que possivelmente explicaria a segmentação da fala em pequenas unidades entonacionais marcadas por fronteiras é a capacidade da memória de trabalho do ser humano. Assim, esse tipo de segmentação é uma necessidade do processamento linguístico devido ao limite da memória de trabalho humana.

A segmentação da fala em unidades entonacionais marcadas por fronteiras também é justificada por razões linguísticas. Uma das razões linguísticas que possivelmente explicaria a divisão da fala em pequenos agrupamentos marcados por fronteiras é a necessidade de identificar adequadamente o domínio das relações linguísticas na fala. Assim, a segmentação do fluxo da fala em unidades discretas é necessária para definir o real domínio das relações linguísticas.

Evidencia-se a importância da segmentação da fala em unidades entonacionais ao analisarmos exemplos simples de português:

Exemplo (1)
Pedro vai para o Rio até amanhã.

Algumas das possibilidades de segmentação e interpretação do exemplo (1) são:

a) Pedro! [*Chamamento*]. Vai para o Rio! [*Ordem*] Até amanhã! [*Despedida*]
b) Pedro vai para o Rio [*Asserção*]. Até amanhã. [*Asserção*]
c) Pedro vai para o Rio até amanhã. [*Asserção*]
d) Pedro, vai para o Rio até amanhã! [*Ordem*]

O exemplo (1) e a suas diversas possibilidades de segmentação evidenciam que, de acordo com a segmentação, as relações linguísticas mudam completamente. Caso o sintagma "*Pedro*" se configure como um chamamento e o sintagma "*vai para o Rio*" como uma ordem, *Pedro* não é sujeito do sintagma "*vai para o Rio*". Assim como, se o sintagma "*até amanhã*" é uma despedida, esse sintagma não se configura como adjunto adverbial do sintagma "*vai para o Rio*". É fundamental

que a segmentação da fala em unidades discretas seja adequada às especificidades da diamesia, pois a interpretação dos enunciados produzidos oralmente só é acessível através do sinal acústico.

O exemplo (2) também evidencia a necessidade de segmentar a fala em unidades entonacionais:

Exemplo (2)
Não comprei a maçã.

O exemplo (2) pode ser segmentado de várias formas, como:

a) Não comprei a maçã ||
b) Não || comprei a maçã ||

As duas possibilidades de segmentação carregam consigo possibilidades de interpretações distintas. O enunciado pode ser interpretado como um enunciado em que o falante não comprou a maçã (opção A) ou como um enunciado em que o falante comprou a maçã (opção B). A ambiguidade de segmentação e interpretação é desfeita pela segmentação prosódica, pois é ela quem diz como a sequência deve ser segmentada e como as unidades segmentadas devem ser interpretadas para identificar adequadamente o real domínio das relações linguísticas.

O exemplo (1) e as duas possibilidades de segmentação (opções A e B) nos mostram que o mesmo conteúdo segmental (fones) pode ser agrupado em duas formas distintas, pelo menos. Observa-se que os fones são iguais, entretanto, os enunciados A e B são muito diferentes entre si. No enunciado A, os fones são agrupados em um único fragmento marcado por uma fronteira no final da sequência, já no enunciado B, os fones são agrupados em dois fragmentos marcados por uma fronteira localizada após a palavra não e outra localizada no final da sequência.

Os exemplos acima ilustram a importância de uma segmentação da fala adequada às especificidades da diamesia. Quando se trata de fala, o real domínio das relações linguísticas é estabelecido por meio de fenômenos de natureza prosódica ao longo do sinal acústico. Os fenômenos que são apontados pela literatura como responsáveis pela identificação adequada das reais relações linguísticas são fronteiras prosódicas ao longo do fluxo da fala.

Se segmentarmos os exemplos (1) e (2) em palavras e/ou *sílabas fonéticas, ambas as segmentações serão inúteis para o estudo das* relações linguísticas relevantes do ponto de vista comunicativo da fala, porque sílabas e palavras não são capazes de identificar relações linguísticas relevantes para a compreensão dos enunciados de fala. Isso se deve ao fato de que, durante a comunicação oral, os falantes produzem segmentos vocálicos e consonantais, mas a identificação adequada das relações linguísticas é determinada por meio de rupturas perceptivelmente

relevantes, realizadas por meio de alguns parâmetros fonético-acústicos. De fato, a segmentação da fala em sílabas ou palavras é muito importante para alguns estudos. Entretanto, as sílabas e palavras mostram-se insuficientes para compreender aspectos interacionais da diamesia falada, pois muitos enunciados de fala considerados simples somente são interpretados adequadamente diante da segmentação prosódica.

É importante enfatizar que as transcrições e os critérios sintáticos mostram-se insuficientes para a segmentação dessa unidade. Embora útil para alguns propósitos específicos, o uso de transcrições mostra-se como uma alternativa muito limitada para segmentar a fala em unidades entonacionais comunicativamente relevantes, tendo em vista que as transcrições não revelam nenhum tipo de informação prosódica ao longo do sinal acústico. As transcrições são apenas representações aproximadas da fala.

A proposta sintática adota a sentença falada como unidade de referência para segmentação da fala. A sentença falada pode ser definida como a máxima projeção do Sintagma Verbal (CHOMSKY, 1970). Todavia, a máxima projeção do SV não inclui a grande quantidade de enunciados de fala sem formas verbais, além daqueles enunciados em que o verbo não é núcleo ou não tem função verbal. No *corpus* C-ORAL-ROM (CRESTI; MONEGLIA, 2005), os resultados mostram que, na fala espontânea, 38,1% do total de enunciados sequer apresentam verbos em sua estrutura. Grande parte dos enunciados é formada somente por sintagmas nominais, sintagmas preposicionais, adjetivos ou interjeições. Nos *corpora* C-ORAL-BRASIL[2] (RASO; MELLO, 2012; Raso e Mello em preparação), 22,1% dos enunciados de textos monológicos e 29,5% dos enunciados de textos dialógicos também não têm formas verbais.

Em diversas línguas, de acordo com Cresti e Moneglia (2005), os enunciados sem verbos constituem cerca de um terço do total. Raso e Mittmann (2012) observam que, em PB, o total de enunciados sem a máxima projeção do SV, sem formas verbais, além dos enunciados em que o verbo não é núcleo ou não possui função verbal supera 50%. Assim, o uso de critérios sintáticos também se mostra como outra alternativa insuficiente para segmentar a fala em unidades, pois os princípios que governam a organização da língua falada e da língua escrita não são iguais devido às características diamésicas distintas.

É comumente reconhecido que, durante a comunicação oral, o fluxo da fala é dividido em pequenos fragmentos, também chamados de unidades entonacionais ou prosódicas, marcados por fronteiras prosódicas (SCHUBIGER, 1958; CHAFE, 1980; SCHUETZE-COBURN, 1994; LADD, 2008; SZCZEPEK REED, 2010). As unidades entonacionais podem ser analisadas segundo

[2] O C-ORAL-BRASIL é um *corpus* de referência que visa a estudar a fala espontânea do português brasileiro. Para mais detalhes, veja-se o endereço eletrônico: <http://www.c-oral-brasil.org/>.

perspectivas teóricas diferentes: sintáticas (COOPER; PACCIA COOPER, 1980; SELKIRK, 2005), pragmáticas (HALLIDAY, 1965; CRESTI, 2000; SZCZEPEK REED, 2012) e cognitivas (CHAFE, 1994; CROFT, 1995; BYBEE, 2010). As fronteiras prosódicas, contudo, podem ser estudadas *per se*, independentemente da perspectiva teórica (BARTH-WEINGARTEN, 2016).

Segundo Zhang (2012, p. 2), "a prosodic break is a perceptible break that marks the grouping of words in an utterance." Blaauw (1994) também define a fronteira prosódica de forma semelhante. De acordo com Blaauw (1994, p. 361) "the term prosodic boundary refers to breaks in the flow of speech, realized by prosodic means". Muitos trabalhos da área adotam as fronteiras de natureza prosódica como critério de orientação para segmentar a fala em unidades relevantes para a comunicação oral entre falantes (CRESTI, 2000; COOPER; PACCIA; LAPOINTE, 1978; SWERTS; COLLIER; TERKEN, 1994). Entre os trabalhos, parece evidente que as fronteiras prosódicas não são sempre iguais. Uma revisão de literatura indica que os tipos de fronteiras são associados à percepção de conclusão ou de continuação do enunciado (PIKE, 1945; PIERREHUMBERT, 1980; SCHEGLOFF, 1998; SZCZEPEK REED, 2004). O primeiro tipo será chamado de fronteira terminal, o segundo de fronteira não terminal.

A revisão de literatura da área também indica que as teorias linguísticas já identificaram uma série de parâmetros fonético-acústicos básicos para detectar e descrever fronteiras prosódicas ao longo do fluxo da fala. Alguns destes parâmetros são recorrentes em diversas teorias, como por exemplo, o *reset* da frequência fundamental (f0), o alongamento pré-fronteiriço, a mudança na taxa de elocução, a pausa silenciosa, o alongamento final, dentre outros (CRUTTENDEN, 1997; AMIR et al., 2004; MO, 2008; BLAAW; 1994). Entretanto, nessas análises, algumas vezes foram utilizados dados de fala gravados em laboratório ou fala lida. Por isso, os resultados obtidos por meio de fala lida ou gravada em laboratório são questionáveis, porque essas situações de comunicação não refletem comunicações reais de fala. Além disso, ainda não há acordo sobre o peso de cada parâmetro para o estabelecimento de uma fronteira (AUER, 2010).

O estabelecimento de uma correlação entre fronteiras prosódicas e parâmetros acústicos para construir uma ferramenta computacional de segmentação automática da fala é extremamente complexo. A complexidade emerge por várias causas. Primeiramente, há uma grande quantidade de parâmetros acústicos, combinados no fluxo da fala, que podem contribuir com o estabelecimento de fronteiras. Ressalta-se também uma maior dificuldade para investigar parâmetros fonético-acústicos, tendo em vista que estes são menos óbvios e mais sutis. Há ainda uma série de dificuldades metodológicas, como por exemplo, a dificuldade de encontrar nos dados trechos de fala absolutamente comparáveis para investigar, as limitações que os dados manipulados em laboratório oferecem, a busca por textos de fala espontânea com alta qualidade acústica etc. Soma-se a esses fatores o fato de que as fronteiras

não constituem uma decisão categórica (BOLINGER, 1972; BIRKNER, 2006; BARTH-WEINGARTEN 2016). Se algumas fronteiras prosódicas são muito salientes e são percebidas por (quase) todas as pessoas, outras têm bem menos acordo quanto à sua percepção. Nesses casos, muitos autores acabam tomando decisões com base em razões teóricas, o que gera um efeito de circularidade (BROWN et al., 1980; PETERS et al., 2005). Por isso, é importante estudar as fronteiras prosódicas independentemente da análise funcional das unidades que elas geram. Por fim, um estudo sobre fronteiras prosódicas *per se* é extremante complexo, porque é difícil investigar um aspecto tão refinado como a percepção humana.

3 Objetivos

O presente trabalho tem como objetivo estabelecer uma correlação entre fronteiras prosódicas e parâmetros fonético-acústicos. O estabelecimento dessa correspondência tem objetivos teóricos e práticos, dentre eles:

- Contribuir ao melhor entendimento teórico sobre a segmentação da fala em unidades entonacionais, tendo em vista que, estas são marcadas por parâmetros fonético-acústicos;
- Investigar os parâmetros fonético-acústicos que orientam a produção e a percepção dos dois tipos de fronteiras prosódicas;
- Possibilitar a criação de um *script* do Praat para segmentação, pelo menos parcial, do fluxo da fala.
- Verificar quais são os parâmetros fonético-acústicos mais relevantes hierarquicamente para a marcação dos dois tipos fronteiras prosódicas.

4 Metodologia

Nesta seção, apresentamos os dados usados, o tratamento recebido pelos dados e as análises estatísticas desenvolvidas.

4.1 Dados

A grande quantidade de material oferecida pelos *corpora* C-ORAL-BRASIL (RASO; MELLO 2012; Raso e Mello em preparação) constitui um conjunto de dados extremamente rico para ser analisado, pois, os dados são representativos do PB na diatopia mineira. Estes *corpora* possibilitam o desenvolvimento de uma grande quantidade de estudos linguísticos, dentre eles, o estudo sobre os

parâmetros fonético-acústicos associados às fronteiras prosódicas, já que, devido às restrições tecnológicas passadas, o estudo da fala baseava-se apenas na transcrição.

O C-ORAL BRASIL prevê a compilação de textos de fala espontânea divididos em uma metade informal e outra formal. A fala espontânea compreende todas as instâncias de comunicação humana em situação natural, sem intervenção do pesquisador, independentemente de ela ser um diálogo, um monólogo ou uma conversação. A metade informal do *corpus* inclui diálogos, monólogos e conversações em contexto natural informal. A metade formal inclui diálogos, monólogos e conversações em contexto natural formal, de mídia e de telefone (em fase de implementação de dados).

Este trabalho usou 11 trechos de fala espontânea monológica, extraídos tanto da metade informal quanto da metade formal do *corpus*. Os textos são relativos a três gêneros de *corpora* de fala espontânea: fala formal em contexto natural, fala informal em contexto natural e fala televisiva. Os trechos são compostos por cerca de 225 palavras e são distribuídos da seguinte forma em duas amostras (I e II):

Tabela 1 – Amostra I

Contexto	Tipologia	Sexo	Texto	Duração	Palavras
Natural informal	Monólogo	Masculino	bfammn11	01'11"	189
Natural informal	Monólogo	Masculino	bfammn24	00'58"	151
Mídia formal	Monólogo	Masculino	bmidmasc01	01'23"	212
Mídia formal	Monólogo	Masculino	bmidmasc02	01'21"	238
Mídia formal	Monólogo	Masculino	bmidmasc03	01'07"	183
Natural formal	Monólogo	Masculino	bnatmasc01	01'30"	205
Natural formal	Monólogo	Masculino	bnatmasc02	01'09"	161
Total			7	08'39"	1.339

Fonte: Elaborada pelos autores

Tabela 2 – Amostra II

Contexto	Tipologia	Sexo	Texto	Duração	Palavras
Natural informal	Monólogo	Feminino	bfammn17	01'31"	217
Natural informal	Monólogo	Feminino	bfammn22	02'01"	369
Mídia formal	Monólogo	Feminino	bmidfem01	01'19"	415
Natural formal	Monólogo	Feminino	bnatfem01	01'18"	140
Total			4	06'09"	1.141

Fonte: Elaborada pelos autores

4.2 Tratamento dos dados

Para o treinamento inicial do *script* de segmentação automática, a Amostra I correspondente à fala monológica masculina foi retida. Optou-se por realizar

o treinamento inicial usando somente os trechos de fala monológica masculina devido às diferenças de frequência fundamental (f0) entre homens e mulheres.

Cada trecho foi segmentado autonomamente por 14 segmentadores *experts*, membros da equipe do Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da Universidade Federal de Minas Gerais. Os segmentadores eram apresentados às gravações e aos textos transcritos, sem nenhuma marcação de fronteiras. A tarefa dos segmentadores consistia em marcar os pontos em que as fronteiras prosódicas eram percebidas, inserindo uma barra simples (/) para fronteira não terminal e uma barra dupla (//) para fronteira terminal. Todos os segmentadores haviam sido treinados e já apresentavam, mesmo que em diferentes medidas, experiência em segmentação prosódica da fala. O treinamento, de duração média de três meses, teve como propósito levar os transcritores a perceber as pistas prosódicas que, consistentemente, estabelecem as fronteiras prosódicas no fluxo da fala e as diferenciam de outras pistas que não estabelecem a marcação de fronteira. O treinamento consistiu em um processo complexo que envolveu várias etapas. Inicialmente, introduziu-se o tema aos anotadores. Em seguida, os anotadores passaram por uma formação que envolveu a realização de exercícios de anotação de percepção de fronteiras, sessões semanais de treinamento e discussão sobre os exercícios.

```
*JAC: basta lembrar que / pra Platão / o mundo / o mundo em que a gente vive / né /
ele / ele é feito pelo demiurgo / que mimetiza / os paradigmas / ou as ideias //
então a gente tem uma concepção de que isso faz parte do processo da natureza / né
/ essa / essa / perspectiva / &he / mimética // e Platão que / &he / critica / em
várias esferas a / mimese / principalmente a mimese dos poetas / ele considera /
que a / a / a mimese é um certo tipo de produção / né "mimesis poiesis tis estin"
// né // tanto é um certo tipo de produção que ela produz / o mundo // na própria
perspectiva dele / não é / inclusive não é um tipo de &pro / de produção / &he /
desclassificado // &he / eu diria que / essa / primeira perspectiva / ela &he / a
mimese / &he / nós poderíamos dizer que ela / é responsável pela forma como o homem
/ tá se apropriando do mundo // &en / enquanto ela implica um reconhecimento / de
que o mundo nu é tocado / diretamente / mas a gente tem / as mediações aí / que é a
nossa ciência / o nosso discurso / e / e e / e essa nossa ciência e o nosso
discurso ele necessita dessas / mediações que o representa / na qualidade dum / de
/ do conhecimento / né / que / a qualidade do conhecimento / dependendo da
qualidade dessas / representações //
```

Figura 1 – Segmentação perceptual de um trecho de fala espontânea monológica
Fonte: Elaborada pelos autores

Os trechos foram segmentados em cinco camadas usando o *software* Praat (BOERSMA; WEENINK, 2017). Foram adotadas as seguintes camadas de anotação do Praat:

1. Segmentação em unidades VV utilizando transcrição fonética larga em caracteres ASCII;
2. Anotação das fronteiras prosódicas não terminais marcadas pelo grupo de segmentadores, informando o número de pessoas que marcaram a fronteira;

3. Anotação das fronteiras prosódicas terminais marcadas pelo grupo de segmentadores, informando o número de pessoas que marcaram a fronteira;
4. Anotação do intervalo referente a pausas silenciosas;
5. Transcrição ortográfica da unidade entonacional.

As unidades VV são calculadas tomando o tempo que vai do *onset* de uma vogal ao *onset* da vogal sucessiva. A unidade VV se diferencia da sílaba convencional, pois ela agrupa a rima de uma sílaba com o ataque da sílaba subsequente. Essa segmentação é mais adequada para o estudo da estrutura rítmica da fala e é consistente tanto em relação à produção da fala como também à sua percepção, conforme vários estudos experimentais realizados em diversas línguas mostram (BARBOSA, 2006).

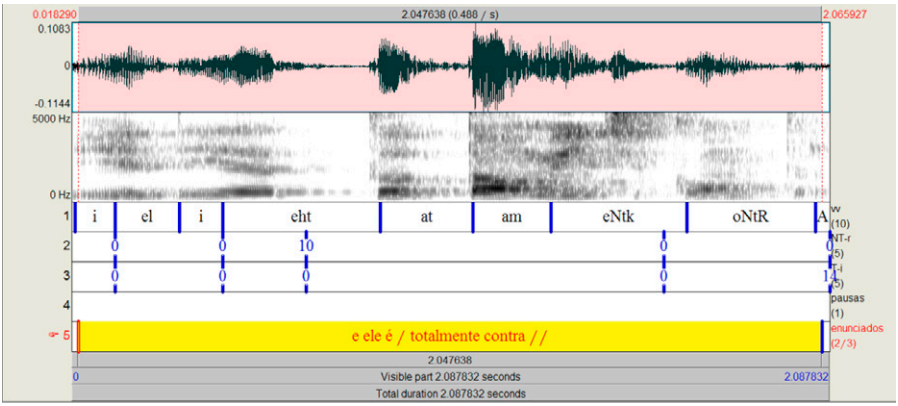


Figura 2 – Exemplo de camadas de anotação do Praat
Fonte: Elaborada pelos autores

Desenvolveu-se uma versão estendida do *script* ProsodyDescriptor (BARBOSA, 2013) para extrair uma série de parâmetros acústicos ao longo do sinal de fala. A versão estendida, denominada BreakDescriptor, extrai os parâmetros acústicos em todas as unidades VV em uma janela centrada em toda fronteira de palavra fonológica, o que inclui as posições percebidas pelos segmentadores como fronteiras e também as não fronteiras. A janela inclui 10 sílabas fonéticas à esquerda e 10 sílabas fonéticas à direita de cada unidade VV. O BreakDescriptor considera como fronteira prosódica as posições em que pelo menos sete segmentadores perceberam uma fronteira, ou seja, pelo menos 50% de acordo. As demais posições, também localizadas em limites de palavras fonológicas, são consideradas pelo *script* como não fronteiras.

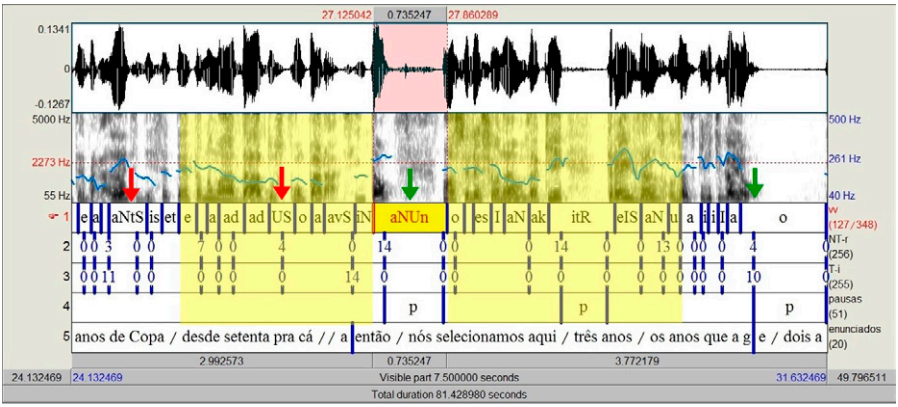


Figura 3 – De cima para baixo: forma de onda, espectrograma de banda larga e camadas de anotação no Praat para trecho de fala espontânea. Destaca-se a janela usada para análise de correlato de fronteira do ponto central
Fonte: Elaborada pelos autores

Na Figura 3, evidencia-se a unidade VV "aNUn" e a janela usada pelo BreakDescriptor para análise de correlato de fronteira prosódica não terminal, porque, na sílaba central destacada, 14 segmentadores marcaram como fronteira não terminal.

Para detalhar melhor o funcionamento do *script*, na figura acima, evidencia-se por meio das setas vermelhas as posições que foram consideradas pelo BreakDescriptor como não fronteiras. As não fronteiras são as posições em que nenhum segmentador indicou fronteira prosódica ou menos de sete pessoas indicaram. Assim, a unidade VV "aNts" e a unidade VV "US" destacadas na Figura 3 são exemplos de posições de não fronteiras.

Observa-se que a unidade VV "o", indicada por uma das setas verdes, obteve desacordo quanto ao tipo de fronteira prosódica. Entre os segmentadores, quatro indicaram fronteira não terminal e dez indicaram fronteira terminal. Por isso, para este caso, a posição foi considerada pelo *script* como uma posição de fronteira terminal, tendo em vista que pelo menos sete pessoas obtiveram acordo quanto ao tipo de fronteira.

No total, para cada unidade VV, são calculadas 111 medidas acústicas globais e locais, possibilitando assim a análise de fenômenos prosódicos ao longo do sinal acústico. As propriedades físicas relativas à produção dos sons extraídos automaticamente pelo *script* compreendem cinco parâmetros fonético-acústicos clássicos:

a. Medidas de taxa de elocução (*speech rate*) e ritmo

A taxa de elocução é um parâmetro que expressa o número de sílabas fonéticas (VVs) produzidas pelo falante a cada segundo. A taxa de elocução é responsável pela percepção de uma velocidade de fala mais acelerada ou lenta. A literatura argumenta que o número de segmentos que o falante produz em um determinado período de tempo também pode ser considerado um parâmetro acústico relevante para o estabelecimento de uma fronteira prosódica, se houver um aumento da taxa de elocução no final da unidade entonacional delimitada por fronteiras (CHAFE, 1994; AMIR; SILVER-VAROD; IZRE'EL, 2004).

b. Medidas de duração normalizada do segmento

A normalização das durações é um procedimento padrão utilizado para atenuar o efeito da duração intrínseca do número de segmentos da unidade silábica e implementação do acento lexical através do cálculo do z-score. O z-score estatístico indica o quanto uma medida se afasta da média em termos de desvios-padrão. Quando o z-score é positivo, isso indica que o dado está acima da média; quando é negativo, significa que o dado está abaixo da média.

O z-score da duração das unidades VVs minimiza os efeitos microprosódicos que não têm função linguística e ressalta a informação prosódica linguisticamente relevante. O valor do z-score assinala o quanto a duração normalizada das unidades VV afasta-se em relação à média da soma das durações dos fones que compõem a unidade VV em português brasileiro (PB). Isto é, o z-score assinala o alongamento ou a compressão da sílaba fonética, independentemente da duração intrínseca dos segmentos que compõem a unidade VV. Neste trabalho, a duração das unidades VVs, segmentadas na primeira camada do Praat, é normalizada automaticamente por meio do BreakDescriptor.

O z-score é calculado em duas etapas. Na primeira, calcula-se a duração normalizada das unidades VVs por meio da fórmula abaixo:

$$z - score = \frac{dur\ - \sum_i \mu_i}{\sqrt{\sum_i var_i}}$$

A variável dur é a duração bruta da sílaba fonética VV em milissegundos. As variáveis μi e $vari$ mostram respectivamente a média e a variância de cada fone que compõe a sílaba fonética. Os valores de média e a variância dos fones são retiradas de uma tabela construída a partir de um *corpus* da língua. Neste trabalho, foram usados os valores de média e variância calculados no *corpus* desenvolvido por Barbosa (2006).

A segunda etapa do cálculo de z-score de duração da unidade VV é feita por meio de um procedimento padrão de média ponderada. Nesta etapa, a curva das durações normalizadas é suavizada por meio da fórmula a seguir:

$$z_{suav}^{l} = \frac{5.z^{i} + 3.z^{i-1} + 3.z^{i+1} + 1.z^{l-2} + 1.z^{l+2}}{13}$$

Em que, z^{i} é o valor de z-score da duração da unidade VV, z^{i-1} é o valor de z-score da duração da unidade VV imediatamente anterior, z^{i+1} é o valor de z-score da duração da unidade VV imediatamente seguinte e assim sucessivamente.

Os descritores estatísticos calculados pelo BreakDescriptor para a duração normalizada da unidade VV compreendem medidas de média, assimetria, desvio padrão e picos de z-score suavizado de duração das unidades VVs. Para as medidas de duração normalizada, também foram calculadas medidas locais em cada unidade de trabalho do *script*.

c. Medidas de frequência fundamental (f0)

A frequência fundamental (f0) é um parâmetro fonético-acústico que expressa o número de ciclos completos de vibração das cordas vocais por segundo. O *pitch* (tom) é o correlato perceptual direto deste parâmetro e sua unidade de medida é o Hertz (Hz). A frequência fundamental é inversamente proporcional à massa das cordas vocais (MARASEK, 1997). Geralmente, as cordas vocais dos homens são mais espessas do que as cordas vocais das mulheres. Como a massa das cordas vocais é mais espessa, o ciclo de vibração das cordas vocais de homens é mais lento, pois, é necessário um tempo maior para realizar um ciclo completo de vibração das cordas vocais. Por isso, a frequência fundamental de homens tende a ter valores menores. Nas mulheres, como a massa das cordas vocais é mais fina, a vibração é mais rápida. Consequentemente, é necessário um menor período de tempo para completar cada ciclo de vibração das cordas vocais. Assim, a frequência fundamental das mulheres costuma ser maior.

Os descritores estatísticos calculados para a frequência fundamental compreendem medidas de mediana, assimetria, desvio padrão e primeira derivada da mediana de f0. Para f0, também foram calculadas medidas locais em cada unidade de trabalho do BreakDescriptor.

d. Medidas de intensidade

A intensidade é um parâmetro fonético-acústico que expressa a quantidade de energia dispersa no sinal sonoro da fala em todo espectro. Geralmente,

a intensidade é medida em dB e tem como correlato perceptual a sensação de volume (*loudness*). Observa-se entre os trabalhos da área que a intensidade não é muito citada na literatura que investiga parâmetros fonético-acústicos relevantes para o estabelecimento de fronteiras prosódicas. Uma das razões é o fato de que a intensidade é um tipo de medida muito afetada pela distância do falante em relação ao microfone de gravação da fala. Naturalmente, se a distância do microfone em relação ao aparelho fonador do falante for muito próxima, a intensidade medida será um valor alto, o que pode não ser necessariamente verdade.

Uma forma de contornar esse problema é usar uma medida de intensidade relativa de bandas, denominada "ênfase espectral". Esta medida é capaz de atenuar os efeitos da posição do microfone sob as medidas de intensidade e fornecer uma medida indireta do esforço vocal. Quanto maior o valor de ênfase espectral, maior o esforço vocal produzido pelo falante.

Traunmüller e Eriksson (2000) observaram que as vogais prosodicamente proeminentes são produzidas com um maior esforço vocal, o que faz com que as faixas de alta frequência do espectro apresentem maior energia. A ênfase espectral é uma medida de intensidade relativa de bandas que permite analisar a produção de vogais proeminentes por meio da presença de mais energia em faixas de frequências mais altas, como atestado anteriormente por Traunmüller e Eriksson (2000). A ênfase espectral é uma medida que indica a contribuição relativa das altas frequências para a intensidade total do espectrograma.

Traunmüller e Eriksson (2000) definem ênfase espectral como a diferença entre a intensidade acústica total do sinal original e a intensidade do sinal após ser submetido a um filtro passa-baixas com frequência de corte em 400 Hz.

Neste trabalho, optou-se pelo cálculo da ênfase espectral considerando a energia total 22050 Hz. Fixou-se 22050 Hz como intensidade total, porque esta é a taxa de amostragem padrão dos textos do *corpus* C-ORAL BRASIL. Os descritores estatísticos calculados para a ênfase espectral compreendem apenas medidas de média.

e. Medidas de pausa

Frequentemente, distingue-se "pausa silenciosa" e "pausa preenchida" (ZELLNER, 1994; SORIANELO, 2006; MERLO; BARBOSA, 2012). A pausa silenciosa é composta por uma interrupção do sinal acústico, sendo assim, um silêncio no fluxo da fala. As pausas preenchidas "are disfluencies that consist generally of voiced material that can correspond to elongated single vowels like "uh" in English, of portions of syllables" (FLETCHER, p. 573, 2010). A pausa preenchida é considerada uma interrupção cognitiva que não interrompe o sinal acústico, pois esse tipo de pausa é realizado por meio de alguma vocalização.

Para este trabalho, as medidas de pausa calculadas pelo *script* compreendem apenas medidas de pausa silenciosa.

4.3 Análise estatística

As medidas extraídas automaticamente foram submetidas a dois modelos estatísticos de classificação, Random Forest (RF) e Linear Discriminant Analysis (LDA), para identificar a combinação de medidas que melhor explicam a segmentação realizada pelos segmentadores perceptualmente. Consideramos, para ambos os modelos de classificação, a presença de fronteira e a ausência de fronteira, tanto para as fronteiras terminais, quanto para as fronteiras não terminais. Consideramos também o poder de predição dos dois modelos obtidos. A predição compara a classificação perceptual prévia dos grupos com a classificação automática feita pelo modelo obtido, mostrando acertos e falsos alarmes para um conjunto de dados. Nesta etapa, utilizou-se 70% da Amostra I, independentemente se os 70% correspondiam às fronteiras ou às não fronteiras. Os 70% dos dados usados foram escolhidos aleatoriamente. Optou-se por utilizar 70% para que em 30% restante dos dados seja avaliado posteriormente o poder de generalização alcançado pelos classificadores estatísticos usados. Nos resultados aqui apresentados, apenas uma parte dos parâmetros acústicos foi usada para o treinamento.

5 Resultados parciais

Apresentamos abaixo as fronteiras prosódicas identificadas perceptualmente pelos segmentadores e o poder de predição alcançado pelos dois modelos de classificação usados na análise.

5.1 Resultados – Random Forest

O Random Forest utilizou em seus cálculos as ocorrências de presença e ausência de fronteiras discriminadas na Tabela 3.

Tabela 3 – Treinamento inicial – Frequência bruta de fronteiras terminais e não terminais retidas

Fronteira	Presença	Ausência
Terminal	47	785
Não terminal	185	646

Fonte: Elaborada pelos autores

A Tabela 4 mostra os resultados alcançados pelo classificador Random Forest:

Tabela 4 – Predição – RF – Treinamento inicial

Fronteira	Acertos Presença	Erros Presença	Acertos Ausência	Erros Ausência
Terminal	28%	72%	99%	1%
Não terminal	19%	81%	94%	6%

Fonte: Elaborada pelos autores

O modelo RF acertou 28% das fronteiras terminais. Em outras palavras, o modelo RF apresentou uma convergência de 28% com as fronteiras terminais marcadas perceptualmente pelos segmentadores. Consequentemente, 72% das fronteiras terminais marcadas pelos segmentadores não foram identificadas.

O modelo RF apresentou uma convergência de 99% com os pontos em que os segmentadores não perceberam fronteiras terminais. Com isso, o modelo RF obteve apenas 1% de falsos alarmes em relação à ausência de fronteiras terminais. Os falsos alarmes são erros relativos as posições em que os segmentadores não perceberam fronteira, entretanto, o modelo indicou a presença de uma fronteira.

O modelo RF acertou 19% das fronteiras não terminais. Em outras palavras, o modelo RF apresentou uma convergência de 19% com as fronteiras não terminais marcadas perceptualmente pelos segmentadores. Consequentemente, 81% das fronteiras não terminais marcadas pelos segmentadores deixaram de ser identificadas.

O modelo RF apresentou uma convergência de 94% com os pontos em que os segmentadores não perceberam fronteiras não terminais. Assim, o modelo RF obteve 6% de falsos alarmes em relação à ausência de fronteiras não terminais.

5.2 Resultados – Linear Discriminant Analysis

O LDA utilizou em seus cálculos as ocorrências de presença e ausência de fronteiras discriminadas na Tabela 5:

Tabela 5 – LDA – Treinamento inicial – Frequência bruta de fronteiras terminais e não terminais retidas

Fronteira	Presença	Ausência
Terminal	75	1076
Não terminal	142	1010

Fonte: Elaborada pelos autores

A Tabela 6 mostra os resultados alcançados pelo classificador Linear Discriminant Analysis:

Tabela 6 – Predição – LDA – Treinamento inicial

Fronteira	Acertos Presença	Erros Presença	Acertos Ausência	Erros Ausência
Terminal	57%	43%	98%	2%
Não terminal	38%	62%	95%	5%

Fonte: Elaborada pelos autores

O modelo LDA acertou 57% das fronteiras terminais, ou seja, o modelo LDA apresentou uma convergência de 57% com as fronteiras terminais marcadas perceptualmente pelos segmentadores. Por outro lado, 43% das fronteiras terminais marcadas pelos segmentadores não foram identificadas.

O modelo LDA apresentou uma convergência de 98% com os pontos em que os segmentadores não perceberam fronteiras terminais. Por isso, o modelo LDA obteve 2% de falsos alarmes em relação à ausência de fronteiras terminais.

O modelo LDA acertou 38% das fronteiras não terminais, ou seja, o modelo LDA apresentou uma convergência de 38% com as fronteiras não terminais marcadas perceptualmente pelo grupo de segmentadores. Desse modo, 62% das fronteiras não terminais marcadas pelos segmentadores não foram identificadas.

O modelo LDA apresentou uma convergência de 95% com os pontos em que os segmentadores não perceberam fronteiras não terminais. Assim sendo, o modelo LDA obteve 5% de falsos alarmes em relação à ausência de fronteiras não terminais.

5.3 Parâmetros fonético-acústicos de marcação de fronteiras prosódicas

Apresentamos abaixo os parâmetros fonético-acústicos mais relevantes hierarquicamente para o estabelecimento de fronteiras prosódicas na fala espontânea, de acordo com o modelo LDA, porque este modelo apresentou um maior índice de acerto para detectar automaticamente fronteiras terminais e não terminais.

No caso das fronteiras terminais, os resultados indicam que os parâmetros mais relevantes para a percepção deste tipo de fronteira prosódica são:

Tabela 7 – Parâmetros fonético-acústicos de marcação de fronteiras terminais

Relevância	Parâmetro fonético-acústico	Peso
1º	Duração de pausa silenciosa	2,84
2º	Presença de pausa silenciosa	1,61
3º	Taxa de saliência duracional (picos z-score suavizado de duração das unidades VVs por segundo) à esquerda da fronteira	0,37
4º	Mediana da frequência fundamental em semitons à esquerda da fronteira	0,18
5º	Taxa de unidades VV não salientes por segundo à esquerda da fronteira	0,17
6º	Assimetria da frequência fundamental - Diferença entre assimetria à direita e à esquerda da fronteira	0,14

Fonte: Elaborada pelos autores

De acordo com o modelo obtido pelo LDA, no caso das fronteiras não terminais, os parâmetros fonético-acústicos que parecem ser mais relevantes para a percepção desta fronteira são:

Tabela 8 – Parâmetros fonético-acústicos de marcação de fronteiras não terminais

Relevância	Parâmetro fonético-acústico	Peso
1º	Presença de pausa silenciosa	3,77
2º	Duração de pausa silenciosa	2,74
3º	Taxa de saliência duracional (picos z-score suavizado de duração das unidades VVs por segundo) à esquerda da fronteira	0,37
4º	Primeira derivada da mediana de f0 à esquerda da fronteira	0,33
5º	Assimetria de z-score suavizado de duração das unidades VVs por segundo – diferença entre assimetria à direita e à esquerda da fronteira	0,17
6º	*Reset* de f0	0,15

Fonte: Elaborada pelos autores

6 Discussão dos resultados

Dois modelos de classificação estatística foram utilizados, Random Forest e Linear Discriminant Analysis, para gerar modelos de combinações de parâmetros fonético-acústico capazes de predizer a realização das fronteiras prosódicas percebidas pelos segmentadores. Os resultados indicam sucesso relativo de ambos os modelos na identificação automática de fronteiras terminais e não terminais.

O modelo LDA apresentou um maior índice de acerto na previsão de fronteiras terminais e não terminais do que o RF, porém, observa-se que o LDA apresentou uma taxa de falsos alarmes para a ausência de fronteira maior. Em termos gerais, observamos que o modelo estatístico LDA aproximou-se mais às decisões tomadas pela maioria dos segmentadores. Também há uma maior dificuldade para detectar automaticamente fronteiras não terminais. Contudo, a maior dificuldade pode ser justificada pela percepção humana de fronteiras, já que se observa um maior desacordo inclusive entre anotadores humanos para perceber fronteiras não conclusivas.

De acordo com os resultados obtidos por meio do classificador LDA, os três primeiros parâmetros fonético-acústicos mais relevantes hierarquicamente para o estabelecimento de fronteiras terminais e não terminais são a presença de pausa silenciosa, a duração da pausa e a taxa de saliência duracional (picos z-score suavizado de duração das unidades VVs por segundo) à esquerda da fronteira. Esses resultados sugerem que os parâmetros relativos à pausa e à taxa de proeminência são os parâmetros em comum para a marcação de fronteira sem distinguir o tipo de fronteira, tendo em vista que, o correlato perceptual direto da taxa de saliência duracional é a taxa de proeminência.

Os resultados também indicam que a mediana da frequência fundamental em semitons à esquerda da fronteira, a taxa de unidades VV não salientes por segundo à esquerda da fronteira e a diferença de assimetria de f0 à direita e à esquerda da fronteira estabelecem a terminalidade de uma fronteira prosódica. As medidas supracitadas indicam respectivamente a taxa de articulação à esquerda da fronteira, o *pitch* à esquerda da fronteira e o desvio de f0 em relação à média.

A primeira derivada da mediana de f0 à esquerda da fronteira, a diferença de assimetria de z-score suavizado de duração das unidades VVs por segundo à direita e à esquerda da fronteira, e, o *reset* de f0 estabelecem a continuidade de uma fronteira prosódica. Estas medidas indicam respectivamente mudanças abruptas no contorno entonacional de f0 e no *pitch*, a maior ocorrência de alongamentos ou encurtamentos na duração das unidades VVs à esquerda da fronteira e o reajuste de f0 para um valor mais alto.

Contudo, observa-se que o índice de acerto para identificar automaticamente as fronteiras terminais e não terminais ainda é relativamente baixo. O LDA identificou 57% das fronteiras terminais e 38% das fronteiras não terminais. Como consequência disso, 43% das fronteiras terminais e 62% não terminais não foram identificadas. Por isso, é importante enfatizar que possivelmente sejam necessários outros parâmetros fonético-acústicos para distinguir a terminalidade e a continuidade de fronteiras prosódicas.

As próximas fases da pesquisa têm como enfoque o refinamento do treinamento do modelo LDA, porque observou-se que ele apresentou melhores resultados para a identificar a presença dos dois tipos de fronteira prosódica e foi

capaz de reproduzir melhor o comportamento dos segmentadores obtido perceptualmente. Nas próximas fases, os parâmetros serão eliminados de duas formas distintas. Primeiramente, serão eliminados progressivamente os parâmetros que não se mostram relevantes hierarquicamente para a classificação. Em seguida, os parâmetros também serão eliminados com base nos fenômenos fonético-acústicos que eles representam.

Esse duplo processo de treinamento tem como objetivo reduzir o "ruído" no modelo gerado por parâmetros não relevantes para o estabelecimento de fronteiras, aumentar a porcentagem de acertos para a presença de fronteira e reduzir aquela de falsos alarmes para a ausência de fronteira. Como perspectiva futura, o estudo partirá da percepção de fronteiras prosódicas em um sistema binário categorizado entre terminal e não terminal, mas buscará uma categorização mais refinada sobre subtipos de fronteiras prosódicas além de terminal e não terminal.

Agradecimentos

Agradecemos à equipe do Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da Universidade Federal de Minas Gerais pelo trabalho de segmentação manual das amostras de fala espontânea em unidades VV e anotação da percepção de fronteiras prosódicas.

Referências

AMIR, N.; SILVER-VAROD, V.; IZRE´EL, S. Characteristics of intonation unit boundaries in spontaneous spoken Hebrew – perception and acoustic correlates. ISCA

Archive, *Speech Prosody 2004.* Nara: Japan, 2004.

AUER, P. Zum Segmentierungsproblem in der gesprochenen Sprache. *InLiSt – Interaction and Linguistic Structures*, 49, 2010. Disponível em: <http://www.inlist.uni-bayreuth.de/issues/49/InList49.pdf>. Acesso em: 18 jan. 2017.

BARBOSA, P. A. *Incursões em torno do ritmo da fala*. São Paulo: Pontes Editores, 2006.

_____. Semi-automatic and automatic tools for generating prosodic descriptors for prosody research. 2013. In: BIGI, Brigitte; HIRST, Daniel (Ed.). *Proceedings of the Tools and Resources for the Analysis of Speech Prosody*, v. 13, p. 86-89. Aix-en-Provence: Laboratoire Parole et Language. Disponível em: <http://www.lpl-aix.fr/~trasp/Proceedings/19874-trasp2013.pdf>. Acesso em: 19 abr. 2017.

BARTH-WEINGARTEN, D. *Intonation Units Revised*: Cesuras in talk-in-interaction. Amsterdam: John Benjamins Publishing Company, 2016.

BIRKNER, K. Relative Konstruktionen zur Personenattribuierung. In: GÜNTHNER, S.; WOLFGANG, I. *Konstruktionen in der Interaktion.* Berlim: Mouton de Gruyter, 2006, p. 205-238.

BLAAUW, E. The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication*, 14, p. 359-375, 1994.

BOERSMA, P.; WEENINK, D. *Praat: doing phonetics by computer*. 2017. Disponível em: <http://www.fon.hum.uva.nl/praat/p0p>. Acesso em: 16 jan. 2017.

BOLINGER, D. Around the edges of language. In: BOLINGER, D. (Ed.). *Intonation*: Selected Readings. Harmondsworth: Penguin, 197, p. 19-292.

BROWN, G.; KAREN, C.; KENWORTHY, J. *Questions of Intonation.* Londres: Croom Helm, 1980.

BYBEE, J. *Language, Usage and Cognition*. Cambridge: CUP, 2010.

CHAFE, W. The Deployment of Consciousness in the production of a Narrative. In: _____. (Org.). *The pear stories*: Cognitive, cultural, and linguistic aspects of narrative production. Norwood: Ablex, 1980, p. 9-50.

_____. *Discourse, consciousness and time*: The Flow and displacement of Conscious Experience in Speaking and writing. Chicago: University of Chicago Press, 1994.

CHOMSKY, N. Remarks on nominalization. In: JACOBS, R.; ROSENBAUM, P. (Ed.). *Reading in English Transformational Grammar*. Waltham: Ginn, 1970, p. 184-221.

COOPER, W.; PACCIA COOPER, J. *Syntax and Speech*. Cambridge: Harvard University Press, 1980.

COOPER, W.; PACCIA, J.; LAPOINTE, G. Hierarchical coding in speech timing. *Cognitive psychology*, 10, p. 154-177, 1978.

CRESTI, E. Corpus *di Italiano parlato*. v. 1. Firenze: Accademia dela Crusca, 2000.

_____. Enunciato e frase: teoria e verifiche empiriche. In: BIFFI, M.; CALABRESE, O.; SALIBRA, L. (Ed.). *Italia linguistica*: discorsi di scritto e di parlato – nuovi studi di linguistica italiana per Giovanni Nencioni. Siena: Protagon, 2005, p. 249-260.

CRESTI, E.; MONEGLIA, M. (Ed.). *C-ORAL-ROM*: integrated reference *corpora* for spoken Romance languages. Amsterdam: John Benjamins, 2005.

CROFT, W. Intonation Units and Grammatical Structure. *Linguistics*, v. 33, p. 839-882, 1995.

CRUTTENDEN, A. *Intonation*. 2. ed. Cambridge Textbook in Linguistics. Cambridge:

Cambridge University Press, 1997.

FLETCHER, J. The prosody of speech: Timing and rhythm. In: HARDCASTLE, W. J.; LAVER, J.; GIBBON, F. E. *The handbook of phonetic sciences*, 2. ed. West Sussex: WileyBlackwell, 2010, p. 523-602.

HALLIDAY, M. A. K. *Speech and Situation.* Londres: University College, 1965.

LADD, R. *Intonational phonology*. 2. ed. Cambridge: CUP, 2008.

MARASEK, K. *EGG e Voice quality*. 1997. Disponível em: <http://www.ims.uni-stuttgart.de/institut/arbeitsgruppen/phonetik/EGG/page1.htm>. Acesso em: 18 fev. 2014.

MERLO, S.; BARBOSA, P. A. Séries temporais de pausas e de hesitações na fala espontânea. *Caderno de Estudos Linguísticos*, v. 1, n. 54, p. 11-24, 2012. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/cel>. Acesso em: 20 set. 2015.

MO, Y. Duration and intensity as perceptual cues for naïve listeners' prominence and

boundary perception. *Proceedings of Speech Prosody.* Campinas: Brasil, 2008, p.739- 742.

PETERS, B. et al. Phonetische Merkmale prosodischer Phrasierung in deutshcher Spontansprache. In: KOHLER, K. et al. *Prosodic Structures in German Spontaneous Speech*. Arbeitsberichte des Institut

für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), v. 35a. Kiel: IPdS, 2005, p. 143-184.

PIERREHUMBERT, J. *The phonology and phonetics of English intonation*. Thesis (Ph.D.). Dept. of Linguistics and Philosophy, Massachusetts Institute of Technology, Massachusetts, 1980.

PIKE, L. *The intonation of American English*. Ann Arbor: University of Michigan Press, 1945.

RASO, T.; MELLO, H. (Org.). *C-ORAL-BRASIL I*: *corpus* de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012.

_____. (Org.). *C-ORAL-BRASIL II*: *corpus* de referência do português brasileiro falado formal (em preparação).

RASO, T.; MITTMANN, M. M. As principais medidas da fala. In: RASO, T.; MELLO, H. (Org.). *C-ORAL-BRASIL I*: *Corpus* de referência do português brasileiro falado informal. 1. ed. Belo Horizonte: UFMG, 2012, p.178-221.

SCHEGLOFF, E. Reflections on Studying Prosody in Talk-in-interaction. *Language and Speech*, v. 41, p. 235-263, 1998.

SCHUBIGER, M. *English Intonation*: Its Form and Function. Tübingen: Niemeyer, 1958.

SCHUETZE-COBURN, S. *Prosody, syntax, and discourse pragmatics*: Assessing information flow in German conversation. Theses (PhD.). University of California, Los Angeles, 1994.

SELKIRK, E. Comments on Intonational Phrasing in English. In: FROTA, S.; VIGÁRIO, M.; FREITAS, M. *Prosodies*: With Special Reference to Iberian Languages. Berlim: Mouton de Gruyter, 2005, p. 11-58.

SORIANELLO, P. *Prosodia*. Roma: Carocci, 2006.

SWERTS, M.; COLLIER, R.; TERKEN, J. Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication*, v. 15, n. 1-2, p. 79-90, 1994.

SZCZEPEK REED, B. Turn-final intonation in English. In: COUPER-KUHLEN, E.; FORD, C. *Sound Patterns in Interaction.* Amsterdam: John Benjamins Publishing Company, 2004, p. 97-117.

_____. Intonation Phrases in Natural Conversation: A Participants Category? In: BARTH-WEINGARTEN, D. et al. *Prosody in Interaction*. Amsterdam: John Benjamins Publishing Company, 2010, p. 191-212.

_____. Prosody, Syntax and Action Formation: Intonation Phrases and Action Components. In: BERGMANN, P. et al. *Prosody and Embodiment in Interactional Grammar*. Berlim: Mouton de Gruyter, 2012, p. 142-169.

TRAUNMÜLLER, H.; ERIKSSON, A. Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, v. 107, n. 6, p. 3438-3451, 2000.

ZELLNER, B. Pauses and the temporal structure of speech. In: KELLER, E. (Org.). *Fundamentals of speech synthesis and speech recognition*. Chichester: John Wiley, 1994, p. 41-62.

ZHANG, X. *A comparison of cue-weighting in the perception of prosodic phrase boundaries in English and Chinese*. Theses (PhD.). University of Michigan, Ann Arbor, 2012.

Fluência e interação no inglês aeronáutico: uma análise baseada em pragmática e Linguística de *Corpus*

Fluency and interaction in aviation English: an analysis based on pragmatics and *Corpus* Linguistics

Malila Carvalho de Almeida Prado

Resumo: A proficiência em língua inglesa de pilotos e controladores de tráfego aéreo é exigida para operações internacionais, respeitando-se parâmetros propostos em uma escala de proficiência linguística dividida em seis áreas, dentre as quais, fluência e interação. Na tentativa de compreendermos como essas duas áreas são materializadas em comunicações autênticas e, assim, buscarmos uma aproximação entre o que é ensinado e o real, propusemos uma investigação, por meio da correlação entre a Pragmática e a Linguística de *Corpus*, em um *corpus* com 130 textos transcritos de situações anormais na comunicação entre esses dois profissionais (ou radiotelefônica). Apresentamos a análise de alguns agrupamentos lexicais que possuem funções pragmáticas no contexto de estudo. Concluímos com amostras do *corpus* que justifiquem uma perspectiva mais ampla do ensino e avaliação desse tipo de comunicação utilizada por uma comunidade profissional.

Palavras-chave: Fluência. Interação. Inglês Aeronáutico. *Corpus* oral.

Malila Carvalho de Almeida Prado – Doutoranda pela Universidade de São Paulo – malilaprado@usp.br.

Abstract: Pilots and Air Traffic Controllers are required to have a minimum level of English proficiency for international operations, respecting parameters proposed in a proficiency scale divided in six areas, among which, fluency and interaction. In an attempt to understand how such areas are materialized in authentic communications and, therefore, search for a closer bridge between what is taught and what is real, we propose an investigation by means of Pragmatics and *Corpus* Linguistics, in a *corpus* of 130 transcribed texts from abnormal situations in pilot-controller communication/ radiotelephony. We present the analysis of some clusters, which carry pragmatic functions in the context of study. We conclude with *corpus* samples that justify a wider perspective in the teaching and testing of this communication used by a professional community.

Keywords: Fluency. Interaction. Aviation English. Spoken *corpus*.

1 Introdução

As comunicações via rádio entre pilotos e controladores de tráfego aéreo (ATCO, do inglês *Air Traffic Control Operator*) ocorrem por meio de uma linguagem roteirizada denominada Fraseologia Aeronáutica (ICAO, 2001). Essa linguagem é composta por palavras e frases que descrevem instruções e solicitações durante todo um voo e deve ser clara, objetiva e isenta de itens que possam causar ambiguidades, tais como preposições, determinantes, auxiliares, pronomes, entre outros. A motivação para essa orientação é evitar acidentes como, por exemplo, o ocorrido em Kuala Lumpur, Malásia, no ano de 1989, no qual o número *two* foi interpretado como a preposição *to* na instrução *descend two four zero zero* [desça para dois quatro zero zero], gerando uma diferença de 2.000 pés de altitude e resultando em um final trágico[1].

Outros acidentes aéreos também decorreram de equívocos na comunicação. Para prevenir outros eventos fatais, a *International Civil Aviation Organization* (ICAO) impôs o uso da Fraseologia Aeronáutica como obrigatória, de forma que todas as situações previstas em um voo tenham sua expressão formalmente documentada. Porém, imprevistos podem ocorrer.

Nos momentos em que algum problema é detectado, pilotos e ATCOs devem recorrer ao que a ICAO denomina *plain English*, ou seja, um tipo de linguagem que deve seguir os princípios da Fraseologia Aeronáutica (concisa, objetiva, clara), porém mais espontânea, e que só deve ser utilizada quando a Fraseologia Aeronáutica não for suficiente. Com a finalidade de regulamentar tal uso da língua, a ICAO estabeleceu que os governos signatários averbassem aos seus profissionais uma licença de proficiência linguística para operações internacionais.

[1] Relatório disponível em: <https://aviation-safety.net/database/record.php?id=19890219-0>. Acesso em 04 out. 2017.

Como parâmetro para a avaliação, a ICAO propôs uma Escala de Proficiência Linguística (2004, 2010) dividida em seis níveis (de 1 a 6) e em seis áreas linguísticas (pronúncia, estrutura, vocabulário, fluência, compreensão e interação) a partir da divulgação do Doc 9835, o Manual de Implementação de Proficiência Linguística (ICAO, 2004, 2010). A fim de obter o licenciamento, o piloto ou o ATCO deveria atingir no mínimo o nível quatro (4) em todas as áreas linguísticas. Contudo, não houve consenso em relação ao que seria ou ao que deveria contemplar o *plain English* envolvido nas comunicações, o que levantou dúvidas sobre o material didático e os testes elaborados para ensino e avaliação desses profissionais (ALDERSON, 2009).

Devido à especificidade da linguagem aeronáutica, assim como de outras áreas de inglês para propósitos específicos, há necessidade de maior compreensão sobre a linguagem a ser ensinada e avaliada, especialmente se considerarmos que o escopo de trabalho é a língua falada por toda a comunidade internacional. Essa transição entre culturas distintas, cada vez mais frequente, promove diversos questionamentos sobre que tipo de linguagem deve ser contemplada em sala de aula e, consequentemente, avaliada.[2]

Para entendermos um pouco mais sobre o *plain English* utilizado nas comunicações radiotelefônicas, propomos um estudo pelo viés da Linguística de *Corpus* (LC), mais especificamente sobre fluência e interação, neste trabalho. Como ponto de partida, tomamos por base o conceito da própria ICAO, explicitado no DOC 9835, a saber:

> Para nossas propostas, o objetivo da fluência é a naturalidade do fluxo de produção do discurso, o grau em que a compreensão é impossibilitada por **qualquer hesitação não natural ou incomum, partidas e paradas distrativas, marcadores distrativos (em... huh... er...) ou silêncio inapropriado**. Os níveis de fluência são mais aparentes durante enunciados mais longos em uma interação. Também são afetados pelo grau de expectativa do insumo precedente que é dependente da familiaridade com o roteiro ou esquemas[3]. (ICAO, 2010, p. 4-12, grifo nosso)

Naturalidade do discurso pode remeter à ideia do falante nativo como modelo; é, em si, um termo vago, pois não há como saber o que é um fluxo natural. Um indício mais concreto do que poderia ser considerado na avaliação do construto

[2] Muitas vezes o caminho percorrido é inverso, por isso é denominado "efeito retroativo". Para mais informações, recomendamos a leitura de Scaramucci (2004).

[3] No original: "*For our purposes, fluency is intended to refer to the naturalness of the flow of speech production, the degree to which comprehension is hindered by any unnatural or unusual hesitancy, distracting starts and stops, distracting fillers (em … huh … er …) or inappropriate silence. Levels of fluency will be most apparent during longer utterances in an interaction. They will also be affected by the degree of expectedness of the preceding input which is dependent on familiarity with scripts or schemata"* [todas as traduções a partir daqui são nossas, salvo quando indicado].

de fluência é a disfluência (falsas partidas, pausas longas, hesitações), marcada em negrito na citação antecedente. Há breve menção sobre a importância do interlocutor colocando-o como "insumo anterior", e também sobre a relevância do domínio do conteúdo ("familiaridade com o roteiro ou esquemas").

Definido fluência para a ICAO, buscamos a explicação, no mesmo documento, sobre interação:

> Como as comunicações radiotelefônicas acontecem em um ambiente congestionado, as comunicações entre pilotos e ATCOs devem ser não somente claras, **concisas** e não ambíguas, mas respostas apropriadas devem ser emitidas eficientemente e espera-se a resposta em um tempo curto. A habilidade Interações refere-se a essa competência, assim como à competência de se iniciar turnos e identificar e esclarecer quaisquer mal-entendidos[4] (ICAO, 2010, p. 4-14, grifo nosso).

Percebe-se ênfase na prontidão e adequação da resposta como primordial à comunicação radiotelefônica, principalmente devido ao alto número de falantes em uma mesma frequência. Destacamos a palavra "concisa" nessa última citação, que contradiz o exposto previamente sobre fluência que, segundo o Doc 9835, seria evidenciada em turnos maiores. Há, ainda, duas outras estratégias referidas, aqui denominadas competências: a de abrir turnos e a de resolução de equívocos.

Nas próximas seções, explicaremos a construção do *corpus* que nos forneceu o insumo das análises e, em seguida, a intersecção entre a Linguística de *Corpus* e a Pragmática Linguística, que possibilitou a investigação, descrita posteriormente. Encerraremos com algumas sugestões para a continuação desta pesquisa.

2 O *corpus*

O *corpus* utilizado neste estudo é composto por comunicações via rádio entre pilotos e ATCOs em situações anormais, ou seja, cenários em que haja algum problema técnico que exija dos profissionais um tipo de linguagem que não o uso da Fraseologia Aeronáutica. O áudio é transcrito a partir do momento em que o problema é abordado, e encerrado na sua resolução ou quando o diálogo não envolve mais os profissionais supracitados. Almejamos, dessa forma, selecionar apenas o contexto de produção do denominado *plain English* pela ICAO (2004, 2010), a saber, o inglês utilizado por pilotos e ATCOs quando a Fraseologia Aeronáutica não for suficiente.

[4] No original: *"Because radiotelephony communications take place in a busy environment, the communications of air traffic controllers and pilots must not only be clear, concise and unambiguous, but appropriate responses must be delivered efficiently, and a rapid response time is expected. The interactions skill refers to this ability, as well as to the ability to initiate exchanges and to identify and clear up misunderstandings*".

Compilamos 130 áudios de diversos sítios eletrônicos; o mais utilizado foi o *liveatc.net*, um repositório de comunicações aeronáuticas. Para ampliarmos o repertório de problemas, utilizamos uma taxonomia da ICAO, cujo propósito é uniformizar os relatórios de acidentes e/ou incidentes aéreos entre todos os governos signatários (ICAO, 2006). Totalizando 33 categorias, das quais duas foram descartadas (por não atenderem às exigências de compilação do *corpus*, uma vez que não envolviam os profissionais indicados), coletamos pelo menos quatro áudios para cada ocorrência, sendo que um deles deveria obrigatoriamente envolver tráfego internacional, ou seja, um dos falantes deveria ser estrangeiro no território ou espaço aéreo onde o evento aconteceu. Essa proposta nos permite analisar o material linguístico a partir de uma perspectiva do uso do inglês como *língua franca*.

Para as transcrições, adotamos a Teoria da Língua em Ato (TLA) (CRESTI, 2000). O embasamento da TLA está no entendimento de que a língua falada tem uma unidade de referência distinta daquela da língua escrita, construída pragmaticamente e identificada pelos usuários por meio de um paradigma entonacional (RASO, 2012). Para a TLA, a unidade tonal é percebida pelo interlocutor como um ato de fala, ou uma ação linguística. Na transcrição, a unidade tonal é representada por barras. Ao distinguirmos as quebras prosódicas, marcadas por essas barras, podemos analisar blocos que apresentam unidade de sentido. Como exemplo, apresentamos abaixo dois enunciados extraídos do *corpus* de estudo:

Quadro 1 – Enunciados retirados do *corpus* de estudo

a. basically you know both engines are running so we just got a vibration on the number two engine
b. continue inbound on the approach follow the approach and I will let you know

Fonte: Elaborado pela autora

No primeiro enunciado, é difícil distinguir se *you know* inicia uma oração subordinada ou se a expressão se trata de um marcador de discurso. Já *so* não parece cumprir o papel de conjunção conclusiva dentro do enunciado. No segundo enunciado, nota-se uma possível repetição de instrução (*continue inbound on the approach* [mantenha o curso na aproximação] e *follow the approach* [siga na aproximação]), mas tal hipótese não pode ser confirmada sem se recorrer à gravação. A última parte do enunciado não identifica o que será informado (*I will let you know*), e nem mesmo se o enunciado foi encerrado ali.

Ao adicionarmos as barras representativas de quebras prosódicas, terminal (representada por duas barras; sinaliza o fechamento do enunciado) ou não

terminal (representada por uma barra; sinaliza a unidade tonal dentro de um enunciado) aos enunciados aqui ilustrados, é possível sanar as dúvidas:

Quadro 2 – Enunciados sinalizados com quebras prosódicas

a. basically / you know / both engines are running // so we just got a vibration on the number two engine //
b. continue inbound on the approach / follow the approach / and I will let you know //

Fonte: Elaborado pela autora

No primeiro enunciado, reconhece-se pelas barras que *you know* [você sabe] funciona, nesse segmento, como marcador de discurso, não como introdução à oração subordinada. Denotando uma atenuação do problema (*so we just got a vibration on the number two engine* [então só tivemos uma vibração no motor número 2]), a conjunção *so* [então] abre novo enunciado, identificado por meio das barras duplas que a antecedem. No segundo enunciado, as barras simples confirmam a repetição da instrução emitida pelo ATCO, e a conjunção *and* aponta a função de uma sequência lógica: "continue a aproximação e informar-lhe-ei" (algo solicitado pelo piloto). As barras duplas evidenciam o término do enunciado, exigindo que busquemos as partes precedentes no diálogo para entendermos o que foi solicitado pelo piloto.

O *corpus* soma 110.737 palavras. Apesar de não ser de grandes proporções se considerarmos outros *corpora* de língua falada (tais como o Michigan Corpus of Academic Spoken English[5], com quase 2 milhões de palavras, ou o recém-lançado Spoken BNC 2014[6], com 10 milhões de palavras), nosso *corpus* está em conformidade com outros estudos que preveem pequenas quantidades de textos para a compilação de *corpora* técnicos (GAVIOLI, 2005). Além disso, durante a compilação foram realizados testes para a verificação da riqueza lexical, que jamais ultrapassou 4%, atestando que o acréscimo de textos não amplia o conteúdo significativamente. Atualmente, o *corpus* está encerrado com 2,96% de riqueza lexical.

3 A Pragmática Linguística e a Linguística de *Corpus*

Elementos típicos da linguagem oral têm sido objeto de estudo da Pragmática Linguística há décadas. Adotamos a seguinte concepção de Pragmática: "o estudo da linguagem do ponto de vista de seu usuário, especialmente das escolhas que

[5] Disponível em: <https://quod.lib.umich.edu/m/micase/>. Acesso em: 07 out. 2017.
[6] Disponível em: <http://cass.lancs.ac.uk/?page_id=1386>. Acesso em: 07 out. 2017.

fazem, dos limites que esbarram ao utilizar a língua na interação social e dos efeitos que a língua tem em outros participantes no ato da comunicação"[7] (CRYSTAL, 1997, p. 240). A partir dessa definição, temos a possibilidade de buscar respostas para questões que permeiam esta pesquisa desde a confecção do *corpus*: sendo o inglês aeronáutico uma linguagem tão específica, por que os profissionais se utilizam de grande quantidade de itens de cortesia? Por que dêixis figuram entre as palavras mais frequentemente usadas se os profissionais devem se dirigir uns aos outros por matrículas de voo e/ou estação?

Os itens mencionados estão presentes na língua em uso, interesse de pesquisa das duas teorias aqui utilizadas – a LC e a Pragmática. A LC necessita de *corpora* para a investigação linguística, pois lhe interessa o estudo da convencionalidade, do uso recorrente e do fenômeno social. Já a Pragmática se ocupa do efeito que determinados enunciados têm sobre o usuário, com o estudo do contexto de produção e da interação entre os falantes. Assim, a Pragmática aborda alguns temas importantes para esta pesquisa, tais como:

– atos de fala: a fala não é meramente um sistema abstraído cujo objetivo é a comunicação, mas sim a execução de ações dentro de contextos sociais específicos (AUSTIN, 1962);

– dêiticos: são nomes cujos significados referenciais dependem do contexto de produção. Podem ser relacionados ao tempo, ao lugar ou à pessoa (FILLMORE, 1966, 1971). Para ilustrar, "são demonstrativos, pronomes pessoais, tempos verbais, certos advérbios de tempo e lugar e alguns verbos como *come* [vir] e *go* [ir]"[8] (LENZ, 2003, p. vii), ou seja, são classes gramaticais ou itens lexicais que necessitam do aqui e agora da enunciação para o entendimento de seu significado;

– cortesia: segundo a teoria de *face work* (GOFFMAN, 1967), ou trabalho de face, os participantes de uma interação atenuam ou suavizam a linguagem para não ameaçar a si mesmos, ou seja, a sua própria face, ou a face do interlocutor. Esse ritual representa um jogo de imagens, tanto de si quanto do outro, em que os interagentes evitam, por questões culturais e/ou sociais, expor um ao outro, forçando-os, entre outras estratégias, a se mostrarem polidos uns com os outros. Brown e Levinson (1987), no entanto, afirmam que não há necessidade de elementos indicadores de cortesia em situações de urgência, como as estudadas na presente pesquisa.

A congruência entre as duas linhas teóricas surgiu, inicialmente, da análise da lista de palavras do *corpus* de estudo, que apontou elementos como pronomes pessoais, advérbios de lugar e tempo, o honorífico *sir* ([senhor] dêitico social), e

[7] No original: "[...] *the study of language from the point of view of users, especially of the choices they make, the constraints they encounter in using language in social interaction and the effects their language has on other participants in the act of communication".*

[8] No original: "*Typical examples are demonstratives, personal pronouns, tenses, certain place and time adverbials and some verbs such as come and go".*

verbos modais. Porém, a maior ocorrência de tais itens está no levantamento de *clusters*, ou blocos de linguagem de duas, três ou mais palavras que frequentemente aparecem juntas. E está nos blocos de linguagem a principal diferença entre as duas perspectivas, mais especificamente na conceituação de significado (ADOLPHS, 2008). Pesquisas voltadas à Pragmática Linguística frequentemente aceitam a distinção entre forma e significado, em contraste aos estudos voltados à LC, para os quais não há distinção entre os dois (ADOLPHS, 2008). O princípio idiomático, que prevê o significado na palavra e seu entorno, não somente no item lexical isolado (SINCLAIR, 1991), é reconhecidamente uma das maiores contribuições da LC, comprovando que forma e significado são indivisíveis.

Os blocos de linguagem, também denominados agrupamentos lexicais, pacotes lexicais, sequências formulaicas, linguagem pré-fabricada, ou, ainda, em inglês, *clusters* ou *chunks*, são encadeamentos de palavras comumente adjacentes, e "têm mais significado determinado idiomaticamente do que linguagem que é construída a todo momento"[9] (NATTINGER; DECARRICO, 1992, p. 57). Schmitt e Carter (2000, p. 6) afirmam que as cadeias de palavras que "recorrem frequentemente [...] estão quase sempre conectadas ao uso funcional da linguagem"[10]. Enfim, há uma junção "que é fortemente associada com uma função"[11] (ADOLPHS, 2008, p. 27) entre a unidade de sentido que se manifesta no construto léxico-gramatical e o ato de fala de Austin (1962). Essa função tem característica pragmática porque reside no efeito que determinada escolha lexical tem no uso da linguagem (ADOLPHS, 2008; AIJMER; RÜHLEMANN, 2015).

A função pode também abrigar elementos conversacionais, ou itens que são necessários à organização textual, oral ou escrita. Sinclair (2004) defende que o léxico vazio (aquele que aparentemente não tem significado transparente) pode servir ao propósito de organizar o texto falado, atestando a característica interacional da fala. Dessa forma, é interessante analisar o "perfil funcional"[12] (ADOLPHS, 2008) de determinadas expressões que podem parecer "vazias".

A utilização dos blocos de linguagem é, também, indicação de fluência (WOOD, 2006), esta comumente vista apenas pelas variáveis temporais (velocidade da fala, número de palavras por frase/enunciado, pausas etc.) (GÖTZ, 2013; MCCARTHY, 2005). Considerando a fluência como "os processos psicolinguísticos de planejamento e produção de fala que funcionam fácil e eficientemente" (LENNON, 1990, p. 391), procuramos itens investigáveis desse planejamento e produção denominados fluencemas, ou seja, "característica abstrata e idealizada da

[9] No original: "[...] *have more idiomatically determined meaning than language that is put together each time".*
[10] No original: "[...] *recur frequently and are often connected with the functional usage of language".*
[11] No original: "[...] *which is closely associated with a specific function".*
[12] No original: "[...] *functional profile".*

fala que contribui para a produção ou percepção da fluência, independentemente de sua realização concreta" (GÖTZ, 2013, p. 8).

Os fluencemas podem ser de produção (foco no falante, por meio de variáveis temporais como velocidade e pausas, sequências formulaicas e estratégia de aprimoramento de fluência), de percepção (foco no ouvinte, por meio de sotaque, precisão, aspectos pragmáticos, diversidade lexical, registro) e de elementos não verbais (gestos, expressões faciais, linguagem corporal) (GÖTZ, 2013). Dedicar-nos-emos aos de produção, que são detectáveis na pesquisa com *corpora*, por meio das sequências formulaicas (WOOD, 2006) e das estratégias de aprimoramento de fluência (repetições, pausas preenchidas e marcadores de discurso). Nessa perspectiva, as estratégias têm uma função mais comunicativa, mais interacional, em que a concepção de fluência não está relacionada ao desempenho individual do falante, mas é coconstruída (MCCARTHY, 2005). Hesitações, pausas preenchidas, falsas partidas podem identificar elementos como tomada de turno, planejamento sobre o conteúdo (não somente linguístico), ou pedido de contribuição ou auxílio dos participantes da interação e, assim como proposto por Götz (2013), consideramos essas disfluências como palavras, assumindo que o falante as produziu intencionalmente.

4 A análise

A primeira pesquisa com o *corpus* de estudo tratou das palavras-chave e seu entorno, contrastando os dados às áreas linguísticas de estrutura e vocabulário da Escala de Proficiência da ICAO (PRADO, 2015). Porém, os dêiticos, o honorífico *sir*, verbos modais e auxiliares no topo da lista de palavras-chave inquietava-nos. Esses mesmos itens são frequentes também nas listas de blocos de linguagem de duas, três e quatro palavras, geradas a partir da lista de palavras-chave, obtida com a ferramenta Word Smith Tools 7.0 (SCOTT, 2016). Se comparadas umas às outras, percebe-se que são partes do mesmo agrupamento (*and uh we*, por exemplo, ocorre 33 vezes, *and uh we're* ocorre 12 vezes, *and uh we'll* ocorre 6 vezes). Para este artigo, selecionamos a primeira parte da lista de blocos de três palavras. Apresentamos, na Tabela 1, os 36 blocos mais frequentes do *corpus*:

Tabela 1 – Os 36 blocos de 3 palavras mais frequentes do *corpus* de estudo

N	Word	Freq	N	Word	Freq	N	Word	Freq
1	We re gonna	106	13	And uh we	33	25	You want to	27
2	Hold short of	71	14	Do you want	33	26	To the ramp	26
3	On the runway	56	15	I m sorry	33	27	Don t know	25
4	I don t	47	16	Souls on board	33	28	Okay we re	24
5	D like to	42	17	At this time	32	29	We ve got	24
6	Let me know	41	18	And we ll	31	30	Be able to	23
7	Uh we re	40	19	To the gate	31	31	Declaring an emergency	23
8	We d like	39	20	Do you need	29	32	That s fine	23
9	Do you have	38	21	If you can	29	33	Hold your position	22
10	I m gonna	35	22	We need to	29	34	If you need	22
11	Thank you very	35	23	You re gonna	29	35	Let you know	22
12	You need to	35	24	Okay thank you	28	36	Uh we ll	22

Fonte: Elaborado pela autora

Ocupamo-nos do estudo dos blocos de linguagem do *corpus* de estudo a fim de investigar se apresentam o perfil funcional sugerido pelas teorias explicitadas na seção anterior.

Foram removidos da lista todos os blocos que são exclusivos da Fraseologia Aeronáutica, que contenham números, ou expressões mandatórias, como *cleared for takeoff* [autorizada decolagem] ou *cleared to land* [pouso autorizado], para citarmos algumas. Nota-se que alguns dos blocos apresentados têm significado semântico mais transparentes, e são típicos da linguagem da aviação, como *hold short of* [mantenha posição no], *on the runway* [na pista], *souls on board* [pessoas a bordo], *to the gate* [para o portão], *to the ramp* [para o pátio], *declaring an emergency* [declarando emergência] e *hold your position* [mantenha posição]. Apesar de serem características da linguagem da aviação, não são expressões formalmente documentadas, com exceção de *hold short of*. O restante da lista, no entanto, pode ser encontrado em quaisquer outros *corpora* orais. São os blocos de linguagem que interessam a esta pesquisa, aqueles que Sinclair (2004) denomina "vazios".

Analisados um a um tanto nas linhas de concordância quanto no contexto expandido, buscamos seu perfil funcional dentro da produção e os agrupamos em quatro principais funções, descritas nas próximas subseções; podem atuar como iniciadores de turno ou enunciado, atenuar a mensagem, descrever ações futuras (a intenção do falante) e apontar a necessidade de troca de informações.

4.1 Iniciadores de turno

O iniciador de turno é "a primeiríssima forma com a qual o falante começa um novo turno na conversa"[13] (TAO, 2003, p. 189). As formas lexicais, podendo ser inclusive marcadores de discurso, são predominantes nos iniciadores (SCHEGLOFF, 1996), e Tao (2003), em um estudo sobre iniciadores em *corpora*, identifica a pausa preenchida como um lexema independente.

Em nosso estudo, representada principalmente por *uh*[14] (salvo em momentos que a hesitação é explicitamente distinta, como demasiadamente prolongada ou nasalizada), a pausa preenchida raramente quebra o bloco de linguagem, como previsto em outras teorias (GÖTZ, 2013). Também a consideramos como palavra, pois o falante optou por utilizá-la em vez de silenciar-se. Essa escolha salienta a importância do interlocutor, uma vez que, na comunicação via rádio, a pausa preenchida demonstra que (i) o falante está segurando o turno e (ii) o falante está sinalizando algum tipo de dificuldade. Na linguagem roteirizada da Fraseologia Aeronáutica, a pausa preenchida aparece especialmente na leitura da matrícula de voo (majoritariamente entre o nome da empresa aérea e o número do voo), ou entre instruções e números (como em *turn right uh two three zero*), como uma indicação de pausa para o raciocínio. Mas quando há uma mudança da linguagem roteirizada para a espontânea, a pausa preenchida ganha um caráter distinto, como pode ser visto no exemplo a seguir:

Quadro 3 – Exemplo com a pausa preenchida

P: **Uh** AIRCRAFT 73E //
C: Go ahead //
P: **Uh** we have some **uh** unruly passengers on board so we would like to return to the gate //

Fonte: Elaborado pela autora

A pausa preenchida no início do turno parece assinalar o anúncio do problema ao ATCO, que simplesmente responde com *go ahead*, sem nomear a si mesmo ou à aeronave, aparentemente compreendendo a intenção não verbalizada do piloto (a entonação também pode ter importância nessa identificação). O piloto, ao receber o turno de volta, inicia novamente com outra pausa preenchida.

Na Tabela 1, *uh* aparece em três blocos: *and uh we*, *uh we´re* e *uh we´ll*. Durante a investigação desses itens, detectamos que são, na sua maioria, iniciadores

[13] No original: "[...] *the very first form with which a speaker starts a new turn in conversation".*
[14] A fim de facilitar a investigação por meio das ferramentas computacionais.

de turno e/ou enunciados, sempre utilizados durante a explanação da anormalidade. Portanto, a pausa preenchida é uma forma de marcação da transição para o uso do inglês comum, ou seja, para o apontamento à menção do problema.

Ao investigar a comunicação entre o piloto e o ATCO do voo que terminou com o pouso no rio Hudson, em Nova Iorque, Garcia (2016) detectou migração entre a linguagem roteirizada da Fraseologia Aeronáutica e o inglês mais coloquial; essa transição é reconhecida pelas barreiras do enunciado, ou seja, pela identificação da pausa preenchida, pelo uso de dêiticos pessoais (ainda que acompanhando matrículas ou nomes das estações), e por expressões como *okay yeah* iniciando turnos. A própria pesquisadora, no entanto, reforça que é necessário verificar se o mesmo ocorre em outras ocasiões. Nossa investigação corrobora a de Garcia (2016). Como em sua pesquisa, outro iniciador comumente utilizado nos nossos dados é *okay*, na Tabela 1, nos blocos *okay thank you* e *okay we´re*. Na Fraseologia Aeronáutica, a afirmação deve ser feita por meio da palavra *affirm* (ICAO, 2001, p. 5-6), e a ciência da informação deve ser dada usando-se a palavra *roger* (ICAO, 2001, p. 5-7). Apesar disso, a palavra *okay* parece confirmar o entendimento de que a comunicação adquiriu um tom mais conversacional após a abordagem do problema. A continuação do exemplo dado anteriormente confirma esse ponto, agora visualizado por completo:

Quadro 4 – Extrato ampliado retirado do *corpus* de estudo

P: **Uh** AIRCRAFT 73E //
C: Go ahead //
P: **Uh** we have some uh unruly passengers on board so we would like to return to the gate //
C: **Okay** / AIRCRAFT73E / umm what would like <interruption> would you like to return to the gate? / that's fine / taxi down the runway and escape via S4 //
P: Vacate via S4 / AIRCRAFT73E / thank you // and Tower / AIRCRAFT73E / uh we would like to request uh the police at the gate //
C: AIRCRAFT73E / that's copied / we're working on that //
P: Thank you //
C: AIRCRAFT73E / we've called the police for you // contact Ground on 121.705//
P: 121.705 / thank you / AIRCRAFT 73E//

Fonte: Elaborado pela autora

Nesse último excerto de comunicação, percebe-se a dificuldade em separar a linguagem roteirizada da Fraseologia Aeronáutica e o *plain English*, pois são mesclados naturalmente pelos profissionais envolvidos. Julgamos importante mencionar que o diálogo aqui descrito ocorreu em um aeroporto situado em um país onde o inglês não é considerado língua local, com uma aeronave que compartilha a nacionalidade da estação de solo, nesse caso, a Torre. Mesmo assim, a conversa continua em inglês.

Para fins de estudo, insistindo em um filtro que separe a linguagem roteirizada do inglês comum, o que se aproximaria mais do modelo descrito pela ICAO (2001) são dois enunciados: (i) *taxi down the runway and escape via S4* [taxie pela pista e livre via S4] e seu cotejamento (a resposta do piloto); e (ii) *contact Ground on 121.705* [contate o Solo na frequência 121.705] e seu cotejamento. O restante compõe o que denominamos *plain English*, que se dá em torno do problema com o passageiro e da necessidade de autoridade a bordo da aeronave. Percebe-se que, dentro dos turnos, os enunciados que se referem ao problema são assinalados por hesitação, por *okay* ou pelo dêitico de pessoa *we*. Dêiticos não devem ser utilizados segundo a Fraseologia Aeronáutica padronizada; argumenta-se que são vagos e dependentes de um contexto que deve ser altamente preciso, e deve-se utilizar a matrícula de voo para a identificação da aeronave e o nome da estação de solo para o seu reconhecimento. Contudo, tanto pelo diálogo aqui exposto quanto pela lista de blocos de linguagem, pode-se observar que o uso de dêiticos é altamente recorrente.

That's fine é outro bloco que abre turnos nas comunicações aeronáuticas. Aparece como iniciador dezessete vezes, algumas dessas após *yes* ou *okay*, complementando o bloco, e nas outras oito vezes como uma aceitação à solicitação feita pelo interlocutor, como no extrato mencionado aqui (*would you like to return to the gate? / that's fine / taxi down the runway and escape via sierra four //*). Nesses casos, *that's fine* está no próprio enunciado, mas sempre delimitado por quebras prosódicas.

Confirmamos, então, que os blocos até aqui descritos funcionam como organizadores textuais, como previsto por Sinclair (2004), principalmente nas aberturas de turnos ou de enunciados.

4.2 Ações futuras

Cinco dos blocos de linguagem exibidos na Tabela 1 referem-se a ações futuras: *we´re gonna*, *I´m gonna*, *and we´ll*, *you´re gonna* e *uh we´ll*. Optamos por não agregar as formas não contraídas a esses últimos blocos (por exemplo, *we are going to*, *and we will*) por dois motivos: são pouco frequentes (a expressão *we are going to* ocorre seis vezes, sendo que em cinco delas é referente ao futuro; *I am*

going to aparece três vezes; *and we will*, seis vezes; *you are going to* nenhuma vez; e *uh we will* somente uma vez) e decidimos analisar as expressões do modo como foram evidenciadas pelas ferramentas da LC. A contração, portanto, é preferida pelos interagentes do *corpus* de estudo.

Na Fraseologia Aeronáutica, ações futuras devem ser emitidas com o auxiliar *will*. Por exemplo, a instrução *climb to flight level 350, report passing flight level 280* [suba para o nível 350, reporte passando nível 280] deve ser verbalizada pelo piloto como *climbing to flight level 350, will report passing flight level 280* [subindo para nível 350, reportará passando nível 280]. No *corpus* de estudo, o uso recorrente dos blocos referentes a ações futuras confirma o seu contexto de produção, já que o piloto deve informar ao ATCO suas intenções ao relatar um problema técnico (ICAO, 2001). Observam-se algumas linhas de concordância com o agrupamento *we're gonna* na Tabela 2:

Tabela 2 – Linhas de concordância com o bloco *we're gonna*

N	Concordance
1	re declaring emergency / we're gonna land three one right /
2	declaring an emergency / we're gonna plan on landing on the
3	d // Yes sir / we figure we're gonna need about thirty minut
4	ird seven two six golf / we're gonna head zero one zero degr
5	e zero two three heavy / we're gonna have the emergency vehi
6	ta two sixty-six heavy / we're gonna have to wait here for a
7	in the number 2 here // We're gonna check it out / Wait / p
8	ve two gear indication / we're gonna level at three thousand
9	ing it is the indication / we're gonna stop on the runway / ha
10	d / advise intentions // We're gonna return and land // Alri
11	ot / just let me know // We're gonna do some checks / we'll
12	e number three to land / we're gonna keep all the supers tog
13	e airport and looks like we're gonna have to dump fuel uh /
14	e / it doesn't look like we're gonna be able to fix the light
15	equency for a minute? // we're gonna try to uh contact Dispa

Fonte: Elaborado pela autora

Mesmo extraídas de seu contexto de produção, as linhas de concordância na Tabela 2 demonstram claramente, com exceção da linha 4, que pilotos e ATCOs estão reiterando suas ações futuras diante de um evento anormal. A linha 4 manifesta uso de inglês comum, preenchendo espaços em que seria possível ater-se à Fraseologia Aeronáutica; mas quando essa linha foi buscada no contexto de produção, percebe-se que é uma tentativa do piloto de compadecer-se com o ATCO após presenciarem uma incursão de pista (e uma possível colisão entre aeronaves).

4.3 Polidez

Na subseção anterior, abordamos a questão sobre o alto uso de itens de cortesia identificados durante a confecção do *corpus* e no levantamento dos agrupamentos de duas, três e quatro palavras. Na Tabela 1, encontram-se alguns deles, a saber: *´d like to*, *we´d like*, *if you can*, *we need to*, *be able to*, *if you need*, e os atos de fala de agradecimento *thank you very* [*much*], *okay thank you* e de pedido de desculpa *I´m sorry*. Conforme mencionado anteriormente, Brown e Levinson (1987) alegam que há abdicação de cortesia em situações de urgência, fato que nossa análise não corrobora. Independentemente do grau de severidade em que a aeronave se encontra, pilotos e ATCOs mantêm a polidez, contrariamente também ao que é pregado pela Fraseologia Aeronáutica, que categoricamente explicita sobre esse fato afirmando que "palavras e expressões que não sejam essenciais, como expressões de polidez, não devem ser usadas"[15] (ICAO, 2001, p.4-2). Para ilustrar, apresentamos os seguintes exemplos na Tabela 3:

Tabela 3 – Linhas de concordância com itens de cortesia centralizados e em negrito

N	Concordance
1	ight right / just advise **if you need** any other assistance //
2	el niner zero and advise **if you need** lower than that // Okay
3	t area // just to advise **if you need** any assistance please /
4	position and be advised **we need to** return to Singapore / ke
5	ect? That's affirmative / **we'd like to** bring the airplane ba
6	o the alerting aircraft / **we'd like to** proceed to whiskey an
7	n one seven center / and **if you can** notify the company I'd a
8	´re just dumping fuel and **we´d like** your help if you uh if y
9	landing on the runway and **we'd like** the uh fire chief to uh
10	do the low pass yet / and **we'd like** the equipment standing b
11	e uh brakes overheat and **we need to** stop to cool down our br
12	lco // Affirmative / and **if you need** any further assistance
13	ot smoke in the cockpit / **we'd like** you to roll trucks to ou
14	y one zero // Disregard / **we'd like** to land on two eight //
15	old short of papa echo / **if you can** do that // Aircraft just

Fonte: Elaborado pela autora

No trabalho em sala de aula[16], ao mostrarmos tais situações os alunos são unânimes em responder que o problema vivenciado é compartilhado entre o piloto e o ATCO, o que poderia explicar o excesso de modalidade nesse contexto.

[15] No original: "[...] *Words and phrases which are not essential, such as expressions of politeness, shall not be used"*.

[16] Até a data da escrita deste artigo, acumulamos nove anos de experiência de ensino de inglês aeronáutico a pilotos civis.

Outra explicação é o fato de o ATCO não saber precisamente o que pode fazer, tornando-se solícito e compreensivo ao problema do piloto. Isso também pode ser verificado em outros agrupamentos da Tabela 1: *do you want*, *do you need* e *you want to*, todos emitidos por ATCOs em oferecimento de auxílio aos pilotos, como visto nas linhas de concordância na Tabela 4:

Tabela 4 – Linhas de concordância com blocos identificando ofertas em negrito

N	Concordance
1	´m sorry / say again? // **Do you need** any detail from us or a
2	is <name>airport // **Do you want to** try and go to Airport
3	two thirty heading / and **do you want to** stay stopped on the
4	ay six is still closed / **do you want** uh the localizer one ei
5	// Three Charlie delta / **do you want to** talk to Airport appr
6	titude your discretion / **do you want to** maintain two thousan
7	ou want to / uh / how **do you want to** proceed from this point
8	// At how many miles **do you want to** intercept localizer? //
9	ct? // Correct // And do **do you want** uh the crash rescue sta
10	check for any flaw / **do you want to** wait for them to inspec
11	and which the taxiway **do you want to** the north side // Come
12	y-five / which runway **do you want to** land on at Airport? //
13	// Aircraft one eighty / **do you need** any assistance? // Nega
14	Aircraft three eight five / **do you want** us to send emergency ve
15	you // Niner zero four / **do you need** to talk to the port veh

Fonte: Elaborado pela autora

A grande maioria das ocorrências está em perguntas, demonstrando o caráter de oferta ao piloto. São todas oriundas de ATCOs, o que pode torná-los prolixos nos momentos em que o piloto necessita de atenção para a resolução do problema. Em muitas transcrições, verifica-se que o piloto comunica o problema, mas não sua intenção (acreditamos que esteja cumprindo os procedimentos que lhe cabem antes de determinar o que fazer), fato que prejudica as ações do ATCO, que passa a oferecer alternativas. Essa questão é também percebida nos próximos blocos analisados, em que pilotos e ATCOs levantam a necessidade de partilha de informações sobre o problema.

4.4 Conhecimento (não) compartilhado

Os últimos três blocos de linguagem da Tabela 1 a serem analisados são *let me know*, *don´t know* e *let you know*. Na Fraseologia Aeronáutica, a palavra a ser usada ao solicitar informação é *advise*, utilizada apenas 87 vezes no *corpus* (poucas vezes quando comparada às 46 vezes que a expressão *let me know* foi usada).

Quando o problema ocorre, cabe ao ATCO oferecer alternativas para a aeronave em questão. Para tanto, o ATCO necessita de informações sobre o problema para poder agir sobre ele, acomodando a aeronave em posição segura (no ar ou em solo); também precisa saber quando agir, principalmente em relação aos vetores (direcionamentos em relação à navegação) a serem instruídos à aeronave. Nesse mesmo instante, os pilotos estão ocupados, envolvidos com uma possível resolução do problema ou com a adequação à nova situação, não podendo priorizar a comunicação, fato que pode ser observado no excerto seguinte:

Quadro 5 – Extrato retirado do *corpus* de estudo (grifo nosso)

C: **Just for your information** runway 16 **is an option // it's available if you need it //**
P: Okay / copied / AIRCRAFT 411 / thank you //
C: Okay / **just keep me advised once you're ready //**
P: AIRCRAFT 411 / thank you //
C: **Recalling** runway 16 / runway 6 is the longest runway available // **you can have** 14 or 16 / **just let me know //**
P: Roger / it'll be uh runway 16 / AIRCRAFT 411 //
C: **Do you prefer** 16? / **confirm**? //
P: Affirm / 16 / AIRCRAFT 411 //

Fonte: Elaborado pela autora

A necessidade de oferecimento de pistas ao piloto torna o ATCO mais verboso, o que pode ser constatado também pela modalização da linguagem (*just*, *if you need it*, *once you´re ready*, *you can have*, *do you prefer*). O piloto percebe a intenção de cooperação e agradece ao ATCO duas vezes, mas o uso de uma linguagem concisa e objetiva indica a falta de disponibilidade para o gerenciamento da comunicação. Mesmo assim, não deixa o ATCO sem resposta, demonstrando o entendimento da importância de sua participação – ainda que mínima – na interação.

O estudo proposto aqui demonstra a dificuldade em discernirem-se as duas áreas linguísticas que propomos investigar: fluência e interação. Na busca por elementos que justifiquem a avaliação da interação no construto linguístico, adotamos uma perspectiva mais pragmática, que possibilita identificar o perfil funcional dos blocos de linguagem considerados semanticamente vazios. Essa mesma função auxilia na investigação dos fluencemas de produção, a saber, hesitação e iniciadores de turno, que por sua vez atestam que a fluência é coconstruída. Assim, propomos a junção dessas áreas para a nossa investigação.

5 Conclusões e encaminhamentos

Ao propormos uma descrição de linguagem dirigida pelo *corpus*, e que toma por parâmetro a Escala de Proficiência Linguística da ICAO, percebemos que a comunicação via rádio é mais interacional do que é sugerido pela escala. A escala desvenda as áreas de fluência e interação, apresentando uma visão mais voltada ao desempenho do candidato, mas isolada do contexto de produção. Verificamos que é necessário considerar o construto linguístico dentro do contexto de produção e que as duas áreas linguísticas devem ser unificadas, uma vez que elementos que se relacionam tradicionalmente com a falta de fluência parecem indicar organização conversacional útil no *face work* de pilotos e ATCOs, inclusive na abertura de turnos ou enunciados. Além disso, os blocos de linguagem – itens de aprimoramento de fluência – frequentemente utilizados por esses dois profissionais estão conectados a uma função pragmática e, assim, interacional. Percebe-se, também, na alta recorrência de itens de cortesia, que há uma certa empatia ou até mesmo cooperação entre pilotos e ATCOs, que parecem assumir a responsabilidade do problema conjuntamente. Outros elementos atenuadores de linguagem atestam essa suposição.

O *plain English* é identificado por barreiras definidas linguística ou prosodicamente, e é mais conversacional do que a linguagem roteirizada da Fraseologia Aeronáutica, considerando a hibridez em que os dois tipos de linguagem ocorrem. A transição entre os tipos de linguagem pode ser reconhecida por meio de pausas preenchidas, alguns iniciadores de turnos e atenuação da mensagem. Essa alternância entre a linguagem roteirizada e uma linguagem mais coloquial ocorre quando os falantes lidam com a situação anormal. A alta presença de atenuadores de linguagem parece identificar a relação profissional entre os interagentes, que compartilham a responsabilidade pelo problema em que estão envolvidos.

A relação forma e significado carrega um perfil funcional, inclusive nos blocos que contêm meramente palavras gramaticais, ou seja, que podem ser considerados lexicalmente vazios. Tais blocos organizam a fala, principalmente ao assinalar ao interlocutor um evento inesperado. Além disso, facilitam o entendimento da transição entre a Fraseologia Aeronáutica e a linguagem espontânea.

Identificamos, até o presente momento, quatro principais funções – ou perfis funcionais – nos blocos de linguagem: iniciadores de turno, itens relacionados a ações futuras, elementos de cortesia e referências a conhecimento (não) compartilhado. Essas funções corroboram um estudo resumidamente descrito no próprio Doc 9835 sobre funções comunicativas (MELL, 2004); os quatro perfis funcionais aqui investigados participam do elenco sugerido.

Assim como em outros projetos que contam com o ensino por meio de *corpora*, os blocos de linguagem aqui analisados podem aprimorar o trabalho didático, ao fornecerem para o professor ferramentas que exemplificam funções

a serem ensinadas. Também podem auxiliar a atividade do avaliador ao prover-lhe um construto mais próximo ao real, menos abstrato e sugestivo. No futuro, temos como objetivo aplicar essa investigação descritiva na sala de aula para verificarmos essa hipótese.

Referências

ADOLPHS, S. Corpus *and context:* investigating pragmatic functions in spoken discourse. Amsterdam: John Benjamins, 2008.

AIJMER, K. *Conversational routines in English:* convention and creativity. London: Longman, 1996.

AIJMER, K.; RÜHLEMANN, C. Corpus *Pragmatics:* a handbook. Cambridge: Cambridge University Press, 2015.

ALDERSON, C. Air safety, language assessment policy, and policy implementation: The case of Aviation English. *Annual Review of Applied Linguistics*, v. 29, p. 168-187, 2009.

AUSTIN, J. *How to do things with words.* Cambridge: Harvard University Press, 1962.

BROWN, P.; LEVINSON, S. *Politeness*: some universals in language use. Cambridge: Cambridge University Press, 1987.

CRESTI, E. Corpus *di italiano parlato*. Vol. I. Firenze: Accademia della Crusca, 2000.

CRYSTAL, D. *A dictionary of linguistics and phonetics*. Cambridge: Blackwell, 1997.

FILLMORE, C. Deictic categories in the semantics of 'come'. *Foundations of language*, v. 2, p. 219-227, 1966.

_____. *Towards a theory of deixis*. The PCCLLU Papers. University of Hawaii, 1971, p. 219-241.

GARCIA, A. Air traffic communications in routine and emergency contexts: A case study of Flight 1549 'miracle on the Hudson'. *Journal of Pragmatics*, v. 106, p. 57-71, 2016. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378216616305689>. Acesso em: 3 out. 2017.

GAVIOLI, L. *Exploring* corpora *for ESP learning*. Amsterdam: John Benjamins, 2005.

GOFFMAN, E. *Interaction Ritual:* essays on face to face behavior. New York: Doubleday, 1967.

GÖTZ, S. *Fluency in Native and Nonnative English Speech*. Amsterdam: John Benjamins, 2013.

INTERNATIONAL CIVIL AVIATION ORGANIZATION (ICAO). *Annex 10 to the Convention on International Civil Aviation*: Aeronautical Telecommunications. Montreal: International Civil Aviation Organization, 2001.

_____. *Manual of implementation of the language proficiency requirements (DOC9835-AN/453)*. Montreal: International Civil Aviation Organization, 2004.

_____. *Manual of implementation of the language proficiency requirements (DOC9835-AN/453)*. 2nd. ed. Montreal: International Civil Aviation Organization, 2010.

_____. *Aviation Occurrence Categories:* Definitions and Usage Notes. Montreal: International Civil Aviation Organization, 2006.

LENNON, P. Investigating Fluency in EFL: A Quantitative Approach. *Language Learning*, v. 40, n. 3, p. 387-417, 1990.

LENZ, F. *Deictic Conceptualisation of Space, Time and Person*. Amsterdam: John Benjamins, 2003.

MCCARTHY, M. Fluency and confluence: what fluent speakers do. *The Language Teacher*, v. 29, p. 26-28, 2005.

MELL, J. Specific purpose language teaching and aviation language competencies. In: ICAO AVIATION LANGUAGE SYMPOSIUM. *Proceedings*... Montreal: [s/n], 2004.

NATTINGER, J.; DECARRICO, J. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press, 1992.

PRADO, M. *Levantamento dos padrões léxico-gramaticais do inglês para aviação:* um estudo vetorado pela Linguística de *Corpus*. 2015, 135f. Dissertação (mestrado em Estudos Linguísticos e Literários em Inglês). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2015. Disponível em: <http://www.teses.usp.br/teses/disponiveis/8/8147/tde-16062015-131340/pt-br.php>. Acesso em: 8 out. 2017.

RASO, T. O *corpus* C-ORAL-BRASIL. In: RASO, T.; MELLO, H. *C-ORAL-BRASIL: corpus* de referência do português brasileiro falado informal. Belo Horizonte: UFMG, 2012, p. 55-90.

SCARAMUCCI, M. Efeito retroativo da avaliação no ensino/aprendizagem de línguas: o estado da arte. *Trabalhos em Lingüística Aplicada*, Campinas, v. 2, p. 203-226, 2004.

SCHEGLOFF, E. Turn Organization: One Intersection of Grammar and Interaction. In: OCHS, E.; SCHEGLOFF, E.; THOMPSON, S. *Interaction and Grammar.* Cambridge: Cambridge University Press, 1996, p. 52-133.

SCHMITT, N.; CARTER, R. Lexical Phrases in Language Learning. *The Language Teacher*, v. 24, p. 1-7, 2000.

SCOTT, M. *WordSmith Tools version 7*. Stroud: Lexical Analysis Software, 2016.

SINCLAIR, J. Corpus*, concordance, collocation:* describing English language. Oxford: Oxford University Press, 1991.

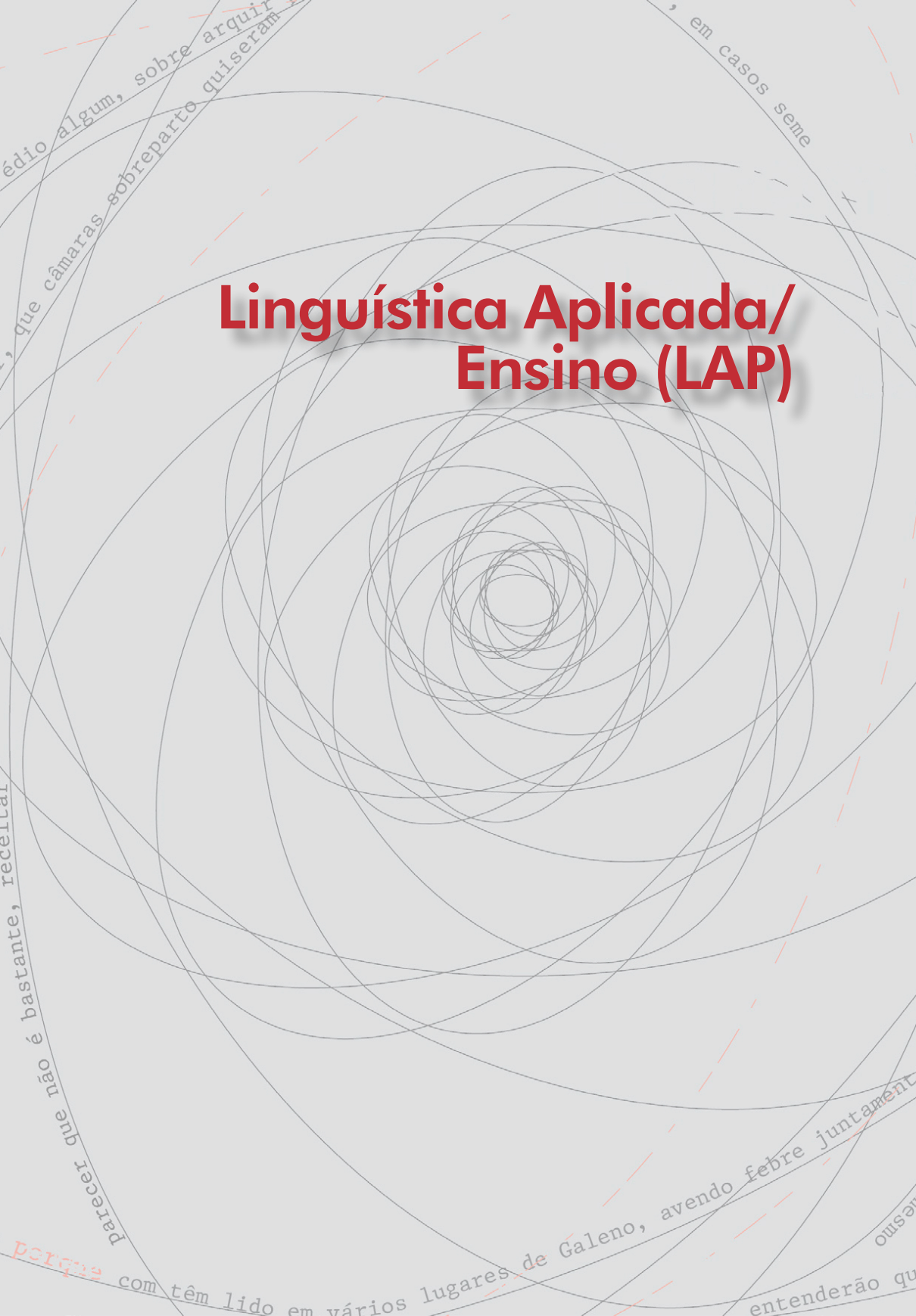
_____. *Trust the text:* language, *corpus* and discourse. London: Routledge, 2004.

TAO, H. Turn initiators in spoken English: a *corpus*-based approach to interaction and grammar. In: LEISTYNA, P.; MEYER, C. Corpus *Analysis:* language structure and language use. Amsterdam: Rodopi, 2003, p. 187-207.

WALSH, S. *Investigating classroom discourse*. London: Routledge, 2006.

WOOD, D. Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review*, v. 63, n. 1, p. 13-33, 2006.

Linguística Aplicada/ Ensino (LAP)



Aos professores, as colocações

To the teachers, the collocations

Andréa Geroldo dos Santos

Resumo: O objetivo deste trabalho é demonstrar a elaboração de oficinas para professores de inglês, com vistas a sensibilizar esses profissionais em relação a quão importante é ensinar colocações, assim como chamar a atenção deles em relação ao fato de que é fundamental se valer de textos autênticos e de ferramentas da Linguística de *Corpus* para essa tarefa. Tais oficinas são necessárias porque o grupo-alvo trabalha com um livro didático, produzido pelo Sistema Mackenzie de Ensino, baseado em *corpus* e textos autênticos – ou seja, o material apresenta uma abordagem diferente da tradicional no mercado de ensino de inglês.

Palavras-chave: Linguística de *Corpus*. Convencionalidade. Livro didático de inglês. Oficinas para professores.

Abstract: The aim of this paper is to show how we have planned workshops for teachers of English, in order to call their attention to the importance of teaching collocations, as well as to use authentic texts and *Corpus* Linguistics tools for such a task. These workshops are necessary because the target audience works with an English textbook, by 'Sistema Mackenzie de Ensino', based on *corpus* and authentic texts – a different approach from what is traditional in the English Language Teaching market.

Keywords: *Corpus* Linguistics. Conventionality. English textbook. Workshop for teachers.

Andréa Geroldo dos Santos – Editora de conteúdo bilíngue na International School, mestre pela Universidade de São Paulo, doutoranda pela mesma universidade – andrea.geroldo@gmail.com.

1 Introdução

Este trabalho, um recorte de nossa pesquisa de doutorado, tem como objetivo demonstrar como temos desenvolvido oficinas para professores de Inglês do Ensino Fundamental II e Ensino Médio, visando a sensibilizá-los quanto à importância de se ensinar colocações e à utilidade das ferramentas baseadas em Linguística de *Corpus* (doravante LC) para realizar tal tarefa. Para isso, valemo-nos do gênero discursivo "livro didático", já que esse tipo de material se constitui em fonte de conhecimento no processo de ensino-aprendizagem no Brasil e, muitas vezes, é a única fonte de apoio pedagógico para os professores (CORACINI, 1999; SOUZA, 2002).

Além disso, desenvolvemos essas oficinas como parte de nosso trabalho como editora de livros didáticos para o ensino de inglês para brasileiros, no Sistema Mackenzie de Ensino (doravante SME) – patrocinado pela Universidade Presbiteriana Mackenzie. Esse sistema de ensino elabora o material didático não só para os colégios Mackenzie, mas também para mais de 200 escolas no Brasil. O livro didático de inglês é constituído de unidades temáticas que apresentam textos autênticos adaptados (em geral, textos jornalísticos) e informado por *corpus*, valendo-se, principalmente, de ferramentas disponíveis no COCA (*Corpus of Contemporary American English*)[1]: KWIC, Key Word in Context – para a elaboração de exercícios lexicais e gramaticais; e Word and Phrase – para a análise dos textos autênticos utilizados, no que se refere às colocações e palavras mais frequentes.

Contudo, tal abordagem não é a que se encontra normalmente no mercado editorial de ELT (*English Language Teaching*) em geral (BURTON, 2012; TOMLINSON, 2003), muito menos no Brasil, onde vários professores de inglês ainda tendem a escolher materiais que contêm textos pedagógicos curtos (criados pelos autores), listas de palavras e apresentação de regras. Tal fato, aliado ao desafio da pesquisa e adaptação constantes quando se trabalha com textos autênticos, parecem explicar por que muitos professores acabam por rejeitar materiais assim elaborados (HANNA, 2012). Outro fator também seria a formação muitas vezes insuficiente dos docentes.

Naturalmente, esses fatores têm influenciado o *feedback* que recebemos em relação ao material do SME. Por essa razão, elaboramos oficinas de curta duração a fim de: a) verificar o modo como os docentes abordam o texto autêntico; b) chamar a atenção quanto à importância de se trabalhar a convencionalidade, em especial, as colocações; c) introduzir ou revisar os elementos básicos da LC; d) propor que os docentes elaborem material didático com vistas ao uso de textos autênticos e ferramentas baseadas em *corpus*.

[1] Disponível em: <http://corpus.byu.edu>. Acesso em: 2 out. 2017.

Dessas oficinas, realizamos dois pilotos relacionados aos itens "a" e "b" supramencionados, os quais descreveremos neste trabalho após discorrermos sobre as premissas teóricas de nossa pesquisa: o livro didático como gênero (seção 2); por que usar texto autêntico (seção 3); a importância de se ensinar a convencionalidade de uma língua estrangeira (seção 4); e a LC e o ensino de línguas (seção 5).

2 O livro didático como gênero

O livro didático (doravante LD) é amplamente adotado no contexto escolar brasileiro, aspecto corroborado pela Câmara Brasileira do Livro, cujos dados atestam que esse segmento do mercado editorial é responsável por 35% do faturamento do setor[2].

Mesmo sob críticas – por exemplo, o LD seria monótono, redundante e limitador, conforme depoimentos de professores (SOUZA, 2002) – esse material constitui-se na principal (e às vezes única) ferramenta e fonte de conhecimento no processo de ensino-aprendizagem do país. Muitas vezes, o LD é a única fonte de apoio pedagógico para os professores (CORACINI, 1999; SOUZA, 2002). Por exemplo, em pesquisa realizada pela Secretaria de Educação Municipal de São Paulo, no início de 2017, constatou-se que 86% dos professores da rede municipal da cidade usam o LD como principal material de apoio em sala de aula[3].

Inegavelmente, o LD possui um caráter de **autoridade**, na medida em que funciona como depositário de um saber estável, de verdade(s) a ser(em) transmitida(s) e compartilhada(s), cuja interpretação é fornecida *a priori* (SOUZA, 2002). Segundo Grigoletto, esse discurso de verdade do LD, discurso "que ilusoriamente se estabelece como um lugar de completude dos sentidos" (GRIGOLETTO, 1999, p. 67-68)[4], pode ser reconhecido no seu modo de funcionamento:

a) por seu **caráter homogeneizante** (todos devem fazer a mesma leitura e chegar às mesmas conclusões);
b) pela **repetição** de estruturas, seções e/ou tipos de exercícios;
c) pela **apresentação** de formas e conteúdo como elementos "naturais".

[2] Disponível em: <http://www1.folha.uol.com.br/mercado/2014/03/1427097-amazon-da-de-graca-ao-brasil-tecnologia-para-converter-livro-didatico-em-digital.shtml>. Acesso em: 18 mar. 2014.
[3] Disponível em <http://www1.folha.uol.com.br/educacao/2017/07/1903081-em-sp-alunos-do-municipio-querem-tecnologia-mas-rejeitam-reforco.shtml >. Acesso em: 21 jul. 2017.
[4] Possibilidade "dada por Foucault, (1979) a partir de sua formulação de que existe um 'como' do poder, uma certa maneira de o poder se disseminar em nossa sociedade, que produz efeitos de verdade" (GRIGOLETTO, 1999, p. 67).

Kramsch (2017) também discorre sobre as características do LD como gênero textual. Segundo o pesquisador, o LD é:

a) orientado por princípios – o LD oferece princípios básicos do conhecimento;
b) metódico – a organização e a progressão do LD compreendem a noção de que o conhecimento pode ser organizado em itens e classificado, e de que a aprendizagem é sequencial e cumulativa;
c) literal – o objetivo de um LD é o de ser compreendido literalmente, ou seja, toda a informação ali descrita não pode dar margem a interpretações diversas;
d) caráter de autoridade – o LD é fonte de autoridade, acima de qualquer crítica; de essência normativa, apresenta a língua como deve ser falada e escrita pelos falantes, assim como a cultura que deve ser lida e interpretada; não há margem para erros de qualquer natureza.

Revestido dessa aparência de autoridade, como discurso da verdade, o LD chega às mãos do professor como produto completo, ora embasado em modelos já usados à exaustão (com nova roupagem), ora como representante de novas abordagens de ensino-aprendizagem – ou mesmo apostando em uma miscelânea desses dois lados. Ao professor, cabe seguir essas verdades, agindo mais como consumidor do que construtor, conforme Grigoletto (1999).

3 Por que usar texto autêntico no ensino de línguas?

A utilização de textos autênticos para o ensino de línguas não é novidade: Titone (1968) e Mishan (2005) observam que, em teoria, se pensarmos no aspecto comunicativo, essa estratégia já era utilizada na Grécia e Roma antigas; como abordagem pedagógica, nota-se seu uso desde o ensino de Latim na Inglaterra do século XVI ao ensino comunicativo no século XX. Contudo, o que tem se alterado é o modo como esses textos têm sido utilizados, de acordo com a abordagem predominante nas diferentes épocas.

Os textos (orais e escritos) autênticos permitem o contato com a língua-alvo em uso e com a cultura dos povos que a usam para se comunicar. Tomlinson ainda destaca que:

> [...] um pré-requisito para a aquisição de uma língua é ter uma experiência rica com a língua em uso.
> [...]
> Isso significa que materiais para aprendizes de todos os níveis devem oferecer contato com o uso autêntico de inglês através de textos orais e escritos que possuam o

> potencial de atrair esses aprendizes cognitiva e afetivamente. Se eles não oferecem tais textos e não estimulam os alunos a pensar e sentir enquanto os experimentam, há pouca chance de esses materiais facilitarem a aquisição de uma língua de maneira duradoura.
> [...]
> Acredito que ajudar os aprendizes a notar os padrões da língua autêntica a que são expostos pode facilitar e acelerar o processo de aquisição dessa língua.[5] (TOMLINSON, 2008, p. 4)

Claro é, conforme observa Tomlinson (2003 e 2008), que muito já se discutiu (e ainda se discute) sobre os possíveis aspectos positivos (por exemplo, o contato com a língua "real") e negativos (como a dificuldade de compreensão desses textos por parte dos alunos iniciantes). Em relação aos aspectos negativos, Mishan (2005) aponta que haveria controvérsias se o uso de textos autênticos:

a) promoveria autonomia, desenvolvimento do senso crítico e aprendizagem efetiva da língua-alvo, ou rejeição por parte principalmente do aluno iniciante;
b) ocorreria realmente quando os textos são adaptados;
c) manter-se-ia quando um texto é retirado de seu contexto de produção e usado em material didático, ou, ainda, quando disponível em um *corpus*.

Controvérsias à parte, Mishan destaca a importância do uso de textos autênticos, pois com base em pesquisas na área de aquisição de segunda língua (*Second Language Acquisition* – SLA), ela afirma que os textos autênticos

> oferecem a melhor fonte de informações ricas e variadas para os aprendizes de língua estrangeira. [...] têm efeito sobre fatores afetivos essenciais à aprendizagem, como a motivação, a empatia e o envolvimento emocional. [...] se prestam a uma abordagem naturalista e de conscientização em relação à aprendizagem da gramática da língua-alvo. [...] estimulam o processamento do cérebro como um todo o que pode resultar em uma aprendizagem duradoura.[6] (MISHAN, 2005, p. 41-42)

[5] "[...] a pre-requisite for language acquisition is a rich experience of language in use. [...] This means that materials for learners at all levels must provide exposure to authentic use of English through spoken and written texts with the potential to engage the learners cognitively and affectively. If they don't provide such texts and they don't stimulate the learners to think and feel whilst experiencing them there is very little chance of the materials facilitating any durable language acquisition at all. [...] It is my belief that helping learners to notice features of the authentic language they are exposed to can facilitate and accelerate language acquisition" (tradução minha).

[6] "provide the best source of rich and varied […] input for language learners. […] impact on affective factors essential to learning, such as motivation, empathy and emotional involvement. […] are suited to naturalistic, consciousness-raising approach to learning TL grammar. […] and stimulate 'whole-brain' processing which can result in more durable learning" (tradução minha).

No tocante aos argumentos pedagógicos para o uso de textos autênticos, Mishan defende que eles podem ser expressos por 3 Cs, em inglês: *Culture*, *Currency* e *Challenge* – Cultura, Circulação e Desafio[7]. Assim, para a pesquisadora, o uso de textos autênticos em material didático é fundamental pois:

a) contêm e representam a(s) cultura(s) dos países da língua-alvo;
b) oferecem temas e a língua correntes;
c) são mais complexos, mas podem ser usados em todos os níveis.

Os textos autênticos constituem-se, assim, em material fundamental para o ensino de línguas, como registro da língua em uso e da cultura de determinada língua. O uso adequado desse material inclui uma abordagem que vá além do trabalho com textos como mero "pretexto" para ensinar gramática, por exemplo.

4 A importância de se ensinar convencionalidade na aula de língua estrangeira

Só o conhecimento das regras gramaticais não garante que o aprendiz escolherá corretamente os elementos lexicais que soem naturais, quando aprendemos/ensinamos uma língua estrangeira. Essa escolha, na verdade, depende de **convenções linguísticas**, as quais são os "jeitos" aceitos pela comunidade que fala determinada língua – a **convencionalidade**. Tagnin (2013, p. 153) define o termo como "o aspecto que caracteriza a forma peculiar de expressão numa dada língua ou comunidade linguística".

A pesquisadora ainda classifica o termo em três níveis, conforme indicados a seguir (TAGNIN, 2013, p. 25-28):

a) pragmático: situações de interação entre os falantes, por exemplo, a necessidade de agradecer por algo recebido, o que exigiria o uso de uma expressão linguística (como "obrigado");
b) semântico: a convencionalidade é observada na relação não motivada entre uma expressão e seu significado (como "bater as botas" para indicar a ideia "morrer"), ou no significado de uma imagem, as chamadas "metáforas visuais" (na cultura ocidental, por exemplo, tudo o que é bom é "para cima", como em "cabeça erguida"; e tudo o que é mau, "para baixo", como em "estar na fossa");
c) sintático: esse nível compreende como os elementos se combinam (associação consagrada pelo uso, como "varinha mágica"), sua ordem ("cama e

[7] Tradução minha.

mesa") e gramaticalidade (o uso consagrou expressões que fogem às regras gramaticais de determinada língua, por exemplo, "de vez em quando")[8].

Dos três níveis, interessa-nos o sintático, pois nele se encontram associações consagradas pelo uso, caracterizadas pela combinabilidade de seus elementos, e formadas por: **base**, a palavra mais conhecida, de forte conteúdo semântico e que determina a ocorrência da outra. Exemplo: vinho (**vinho** tinto); e **colocado**, a palavra que não conhecemos ou de que não nos lembramos, determinada pela base. Exemplo: tinto (vinho **tinto**).

Tais associações são denominadas de coligações (combinações gramaticais) e colocações (combinações lexicais). Como há divergências sobre qual seria a diferença entre esses termos na literatura, usaremos as definições adotadas por Tagnin (2013, p.53-54), a saber:

- **coligações de regência**: verbo + preposição, substantivo + preposição, adjetivo + preposição, advérbio + preposição; locuções prepositivas e verbos frasais[9].
- **colocações**: nominais, verbais, adjetivas e adverbiais.

Ainda em relação às colocações, Firth (1957) introduziu a definição clássica do termo ("You shall know a word by the company it keeps"), ao se referir ao fato de algumas palavras coocorrerem com certa frequência, em associações consagradas pelo uso, e, desta forma, contribuírem para o significado de uma palavra.

5 A LC e o ensino de línguas

A descrição da língua de falantes nativos é a área da LC que mais concentra pesquisas. De acordo com McCarthy (O'KEEFE; McCARTHY; CARTER, 2007), a contribuição da LC para a descrição da língua representa uma mudança sem precedentes em termos de uso de métodos e técnicas científicos para o ensino de línguas. Para o acadêmico, a LC provavelmente prenuncia profundas alterações tecnológicas que vão confrontar noções clássicas de ensino, do papel dos professores, do contexto cultural onde as aulas são dadas e da mediação entre teoria e técnica.

Anteriormente a McCarthy, Sinclair (1988) já apontava para o potencial da LC de alterar os rumos do ensino de línguas, já que as evidências provenientes do *corpus* desafiam certos mitos de que, por exemplo, no ensino de inglês, não existem mais dúvidas quanto aos fatos relativos à sua estrutura. Tal posicionamento dos

[8] Todos os exemplos foram extraídos de Tagnin (2013).
[9] Para saber mais sobre coligações, ver Tagnin (2013).

que acreditam nesses mitos fez e ainda faz com que as metodologias de ensino praticamente ignorem a descrição da língua, segundo resume Berber Sardinha (2004, p.260), perpetuando a crença de que:

- Há dois níveis independentes de organização da linguagem, a sintaxe e o léxico, que justificam o ensino de línguas por meio de currículos e abordagens firmadas na separação entre *gramática* e *vocabulário*.
- A sintaxe tem precedência sobre o léxico, isto é, o vocabulário é subserviente à sintaxe, servindo como *preenchimento de lacunas sintáticas*.
- A fluência nativa ou quase nativa é algo subjetivo que reside na mente dos falantes nativos e que, portanto, não pode ser observada, retratada e descrita objetivamente.
- A frequência dos traços linguísticos como reveladora da padronização e convencionalidade do uso da língua é irrelevante, porque o mais importante na linguagem é seu caráter criativo; portanto, os alunos não precisam aprender sobre modos típicos de expressão em contextos específicos.

Ao confrontar esses mitos, a LC contribui para a:

a) diminuição da separação entre léxico e gramática (tal separação foi introduzida no ensino de línguas pelas gramáticas para consulta e, posteriormente, reproduzida por gramáticas e materiais pedagógicos), já que postula a existência de um nível do sistema linguístico que une o vocabulário às regras gramaticais (léxico-gramática) (BERBER SARDINHA, 2004; SINCLAIR, 2000; XIAO; McENERY, 2005);
b) análise crítica da língua utilizada no material didático para o ensino de línguas, principalmente, livros didáticos e gramáticas de referência, demonstrando que não só a língua abordada por essas publicações difere consideravelmente da língua falada fora das salas de aula, como atestam Mindt (apud XIAO; McENERY, 2005) e O'Keefe, McCarthy e Carter, entre outros, mas também que as observações feitas são baseadas na intuição sobre como usar a língua, do que na evidência do seu uso (O'KEEFE; McCARTHY; CARTER, 2007);
c) utilização de *corpora* eletrônicos para a elaboração de materiais de referência, como gramáticas pedagógicas, dicionários e livros didáticos – postura assumida, por exemplo, pela equipe do projeto Cobuild.

Além da descrição da língua, a utilização de metodologias de pesquisa acadêmica na sala de aula constitui-se em uma das áreas majoritárias da LC. Nela observamos o uso do instrumental analítico da LC em sala de aula, tais como listas de palavras e concordâncias – o instrumento mais utilizado. As linhas de

concordância servem de base para materiais de ensino com propósito variado, como o esclarecimento de dúvidas quanto ao uso de determinadas palavras. Abordaremos as linhas de concordância na subseção 5.1.

Outra área majoritária é a de elaboração de métodos, currículos e material didático. Essa área vale-se dos conceitos da LC ou da exploração de *corpora* para criar metodologias ou abordagens de ensino, por exemplo, a Abordagem DDL, cujo maior expoente é Tim Johns (1991), e sobre a qual discorreremos na subseção 5.2.

5.1 Linhas de concordância e ensino de línguas

O principal instrumento da LC utilizado no ensino de línguas é a concordância. No sentido original, concordância é "um livro de consulta que contém todas as palavras usadas em um texto em particular ou na obra de determinado autor [...], junto com uma lista de contextos nos quais cada uma ocorre"[10] (TRIBBLE; JONES, 1997, p. 1).

As concordâncias não foram inventadas para o ensino de línguas. Na verdade, as primeiras registravam o uso de palavras da Bíblia e teriam sido feitas por cerca de quinhentos monges, liderados por Hugo de San Charo, no século XIII (TRIBBLE; JONES, 1997).

As concordâncias se popularizaram com o advento do computador, já que um programa específico pode, facilmente e de modo confiável, realizar todas as tarefas relacionadas à sua compilação, ou seja, localizar todas as ocorrências de determinada palavra e listar os contextos em que ela ocorre. Esses programas de computador são conhecidos como "geradores de concordância" ou "concordanciadores". Os concordanciadores possibilitam uma fácil visualização de padrões existentes na língua.

Um dos tipos mais comuns é o concordanciador KWIC (*Key Word In Context* – palavra-chave em contexto). Como podemos ver na Figura 1, ele gera resultados na forma de linhas de concordância, as quais mostram a palavra ou expressão que se pretende analisar (*bake*) – o nódulo ou palavra de busca, e seu contexto.

[10] "a reference book containing all the words used in a particular text or in the works of a particular author [...], together with a list of the contexts in which each occurs" (tradução minha).

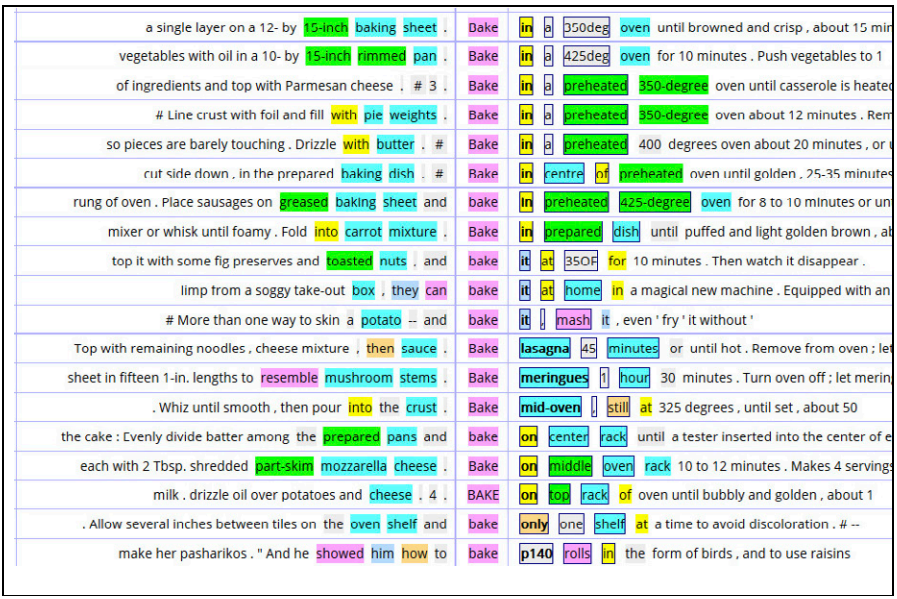


Figura 1 – Linhas de concordância de *bake*, extraídas do COCA (alinhamento à direita)
Fonte: Elaborada pela autora

Nesse concordanciador do COCA, verificamos também que as palavras diretamente relacionadas à de busca também são classificadas morfologicamente. Na Figura 1, como o *span* estava configurado para três palavras à esquerda e à direita, são essas as classificadas: palavras em azul (*oven*), são substantivos; em rosa (*bake*), verbos; em verde (*preheated*), adjetivos; em amarelo (*in*), preposições; e em laranja (*then*; *only*), advérbios e conjunções.

Para facilitar a observação das concordâncias, principalmente se há muitos dados a serem analisados, Berber Sardinha (2004) aconselha fazer uma classificação alfabética das linhas, ordenando-as também, se necessário, pelas palavras à direita e à esquerda da palavra de busca. No COCA, por exemplo, isso pode ser feito na opção KWIC, clicando em R (*right* – direita) ou L (*left* – esquerda), conforme a Figura 2.

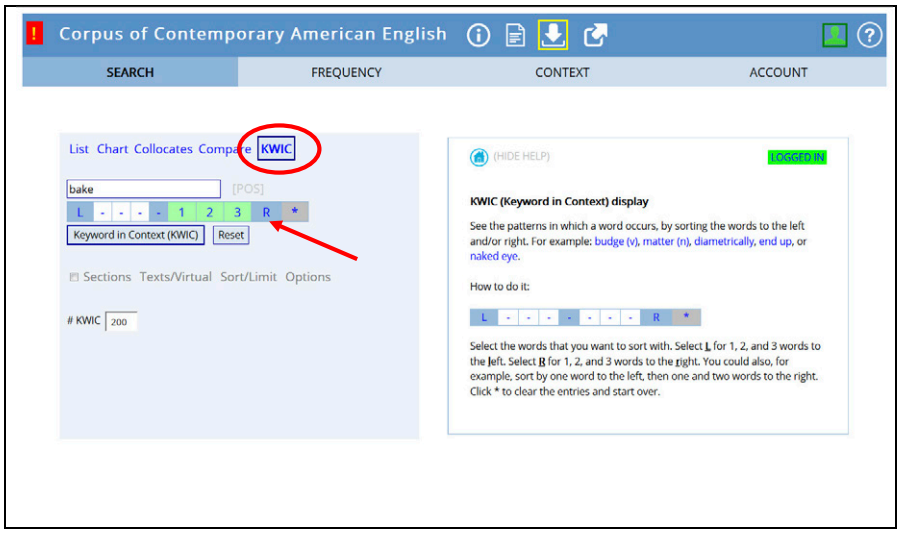


Figura 2 – Concordanciador do COCA
Fonte: Elaborada pela autora

Durante a pesquisa com as linhas de concordância, devemos observar (TOGNINI-BONELLI, 2001):

– todas as linhas, analisando tanto o eixo horizontal, quanto o vertical;
– a repetição de palavras que coocorrem à direita e à esquerda da palavra de busca;
– a repetição de padrões gramaticais, semânticos e funcionais.

A fim de exemplificar, vejamos as linhas de concordância com o verbo *bake*, alinhadas à direita da palavra de busca (Figura 1). Podemos observar que, à direita da colocação, o verbo coocorre oito vezes com a preposição *in* – em sete delas, ocorrendo com o substantivo *oven*, com algumas variações, mas sempre indicando que "o forno deve estar preaquecido":

– *~ in a 350deg / 450deg oven*;
– *~ in a preheated 350-degree / 450-degree oven*;
– *~ in centre of a preheated oven*;
– *~ in preheated 425-degree oven*.

Com as linhas de concordância alinhadas à esquerda de *bake* (Figura 3), notamos a recorrência do verbo *cover* com a conjunção aditiva *and*, indicando quais as ações que precisam ser feitas nessa parte da receita culinária.

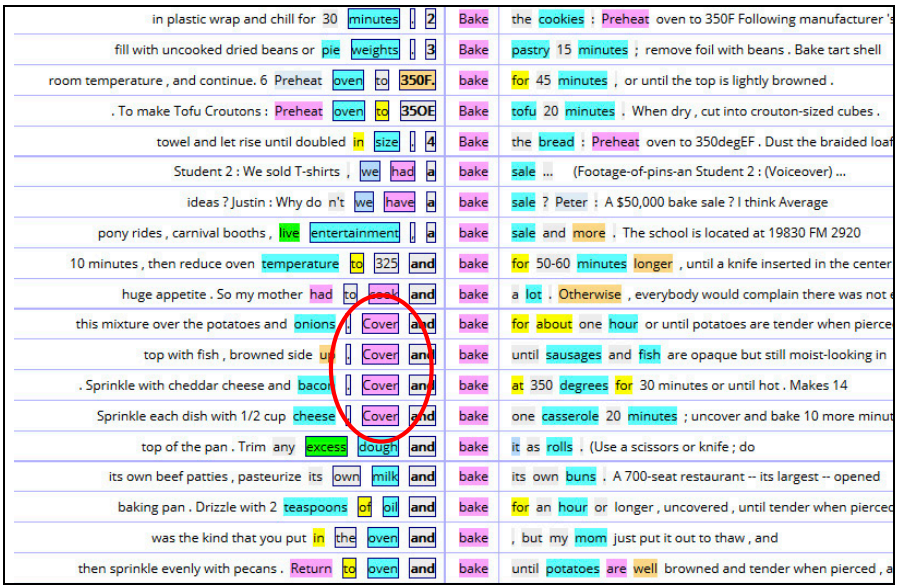


Figura 3 – Linhas de concordância de *bake*, extraídas do COCA (alinhamento à esquerda)
Fonte: Elaborada pela autora

Há duas vantagens gerais quanto ao emprego das linhas de concordância no ensino, segundo Tribble e Jones (1997). A primeira é que elas favorecem o aprendizado por descoberta, já que os alunos podem sozinhos encontrar a resposta para suas dúvidas. A segunda refere-se ao fato de as concordâncias fornecerem exemplos autênticos da língua em uso. Quanto ao modo de se usar as concordâncias no ensino, há várias possiblidades, como a investigação da convencionalidade da língua (BERBER SARDINHA, 2004).

As concordâncias também podem ser usadas para avaliação. Para isso, Berber Sardinha (2004) sugere usá-las em exercícios de preenchimento de lacuna (*gap-filling*), apagando-se algum item lexical de determinadas concordâncias. Outra possibilidade consiste em ligar as colunas (*matching-up exercise*) – para isso, deve-se dividir a concordância entre os lados direito e esquerdo da palavra de busca e misturar as linhas de cada lado.

Contudo, alguns pesquisadores apontam que o uso pedagógico das concordâncias pode enfrentar dois problemas básicos: de treinamento e de conscientização (AIJMER, 2009; BERBER SARDINHA, 2004), afetando tanto o corpo docente quanto o discente. No que se refere ao treinamento dos alunos, se eles não forem instruídos quanto à leitura apropriada das concordâncias – ou seja, diferente da convencional (da esquerda para direita, de cima para baixo, linha por linha), pois ela é guiada pela palavra de busca e sua relação com os outros itens ao redor, poderão se confundir, se frustrar e até rejeitar o trabalho com elas.

No caso dos professores, mais do que interesse e motivação, é preciso saber analisar as concordâncias antes de levá-las para a sala de aula. Caso contrário, o professor poderá introduzir uma concordância que complicará mais ainda o entendimento de determinado aspecto, não ver novas evidências, e/ou restringir a identificação dos padrões por analisar as concordâncias valendo-se apenas de categorias preexistentes.

Torna-se necessário, então, conscientizar alunos e professores quanto aos benefícios do uso da concordância como:

a) instrumento (ou seja, as concordâncias permitem uma compreensão melhor do funcionamento da língua, pois auxiliam na visualização de grande quantidade de palavras e/ou estruturas ao mesmo tempo);
b) instrumento de exploração (o aluno pode ir além da intuição e das regras preestabelecidas e checar o uso dos aspectos linguísticos que lhe interessam e/ou cujas respostas pode não encontrar em obras de referência);
c) possibilidade de independência em relação a professor (no caso dos alunos), a materiais didáticos e de referência.

Entretanto, tais benefícios podem não ser bem-aceitos em contextos mais tradicionais de ensino, em que se privilegia um papel mais passivo do aluno, pois os professores poderiam ser mal interpretados, como se não estivessem cumprindo "sua função". Mesmo em ambientes menos tradicionais, o professor pode ter dificuldade para implantar práticas que necessitam de maior independência e pesquisa por parte do aluno (BERBER SARDINHA, 2004).

Quanto ao aspecto pedagógico, o uso das concordâncias é criticado porque sua prática estaria em desacordo com o ensino comunicativo de línguas; confundiria os meios (o instrumental da LC) com os fins (ensino da língua autêntica representada em um *corpus*) e descontextualizaria a língua (ASTON, 1995; WIDDOWSON, 2000). Em relação à última crítica, a descontextualização ocorreria porque as concordâncias apresentam os padrões em pequenos trechos, oriundos de vários textos, fora do contexto de produção e sem serem empregados em um novo contexto (WIDDOWSON, 2000). Ainda segundo Widdowson (2000), não importa se os exemplos venham da intuição ou sejam autênticos, eles devem ser recontextualizados para uso em sala de aula a fim de torná-los reais para os alunos e para que a aprendizagem seja eficaz.

Embora essas observações de Widdowson já datassem dez anos quando da publicação do artigo em 2000, frutos de uma controvérsia iniciada com Sinclair[11] em 1991, elas ainda se mantinham válidas, conforme atestam McEnery, Xiao e

[11] Para mais informações sobre essa controvérsia, ver Seidlhofer (2003) e McEnery, Xiao e Tono (2006).

Tono (2006) e não foram contestadas por outros teóricos que criticaram as posições de Widdowson. Na verdade, esses teóricos, tais como Stubbs (2001) e de Beaugrande (2001 apud MCENERY; XIAO; TONO, 2006), ao responderem às críticas de Widdowson, ativeram-se àquelas quanto ao uso de *corpora* como linguagem autêntica, à ausência de intuição e à preocupação com dados referentes à frequência, mas ignoraram seu aspecto pedagógico.

Para o nosso trabalho, é inegável a importância dos dados obtidos num *corpus* assim como o uso desses dados e das concordâncias para o ensino da convencionalidade. Porém, não podemos ignorar as críticas de Widdowson (2000) no tocante à necessidade de se contextualizar as concordâncias (aquelas provenientes de artigos de jornal podem ser apresentadas como manchetes, por exemplo) e, muitas vezes, de recontextualizá-las a fim de auxiliar o entendimento dessas linhas e a aprendizagem.

Vistos o uso pedagógico das concordâncias no ensino de línguas e as críticas a esses usos, trataremos de uma das mais populares abordagens a se valer das linhas de concordância, a Abordagem DDL.

5.2 Abordagem DDL

A Abordagem DDL (*Data Driven Learning* – Aprendizagem Movida por Dados), postulada por Tim Johns, constitui-se em uma das propostas mais sólidas a utilizar o *corpus* em sala de aula (BERBER SARDINHA, 2004; GAVIOLI, 2005; McENERY; XIAO; TONO, 2006; O'KEEFE; McCARTHY; CARTER, 2007; TRIBBLE; JONES, 1997). Essa abordagem foi inicialmente criada para ensinar a gramática do inglês, mas se expandiu para outras áreas e línguas. Para Johns (1991), o aluno deve ser encorajado a se tornar um pesquisador, cujo aprendizado é movido pelo acesso a dados linguísticos disponíveis no computador (o aluno tem acesso a *corpora* disponíveis na internet) ou a concordâncias impressas pelo professor.

Essa abordagem teria sido elaborada com base na longa experiência de Johns como educador, segundo Mike Scott (2010). Tal experiência permitiu a Johns observar que a aprendizagem se dá quando o aluno tem de se esforçar para "perceber, raciocinar, inferir e associar" (SCOTT, 2010, p.7). Assim, Johns (1991) propôs o trabalho com uso do *corpus*, por parte dos alunos, dividido em três etapas de cunho indutivo, intituladas *Identify – Clarify – Generalise* (JOHNS, 1991, p.4), "Identificar – Elucidar – Generalizar". A primeira refere-se à observação dos dados na concordância. Na segunda etapa, ocorre a classificação dos padrões evidentes. Na terceira e última etapa, chega-se à generalização das regras.

Em outras palavras, a produção do conhecimento por parte dos alunos é feita de modo ascendente (*bottom-up*), contrária à abordagem dos "três P" (*Presentation, Practice* e *Production* – Apresentação, Prática e Produção), ainda em voga nos LDs,

que é descendente (*top-down)*. A abordagem de Johns alinha-se à proposta por Carter e McCarthy para ser usada com o *corpus*, os "três I" (*Illustration, Interaction* e *Induction* – Ilustração, Interação e Indução), conforme Xiao & McEnery (2005).

Há várias vantagens na Abordagem DDL. Entre as principais, está a possibilidade de o aluno desenvolver as habilidades de identificar padrões e generalizar as regras. Além disso, o ensino é centrado no aluno, com o professor assumindo o papel de orientador. Outra vantagem é que, como as regras não são transmitidas pelo professor, mas inferidas pelos alunos por meio da observação das linhas de concordâncias, muda-se o foco tradicional do ensino da gramática, em que as regras são apresentadas prontas aos alunos (BERBER SARDINHA, 2004; XIAO; McENERY, 2005).

Com base no que foi observado nessa subseção, decidimos adotar a abordagem DDL na elaboração de LDs, no que se refere ao trabalho indutivo de três etapas com as concordâncias. Porém, porque postulamos que os professores precisam ter total conhecimento do material que vão apresentar aos alunos, não seguimos o que Johns (1991) propõe quanto ao professor não saber de antemão os padrões que os alunos encontrarão em determinado *corpus*.

Após discorrermos sobre a base teórica de nossa pesquisa, demonstraremos na próxima seção a metodologia das oficinas que desenvolvemos.

6 Metodologia

Nesta seção, discorreremos sobre a metodologia de duas partes distintas de nosso trabalho, mas complementares: elaboração de material didático baseado em *corpus*; elaboração e condução de oficinas.

6.1 Metodologia para a elaboração de LD

Para a elaboração dos LDs, privilegiamos o uso de textos autênticos ao longo das quatro unidades temáticas de cada livro, com três capítulos cada. Os exercícios propostos para a compreensão linguística desses textos autênticos são:

a) baseados na análise obtida na ferramenta Word and Phrase.Info[12] (doravante WPI), no tocante à frequência das palavras e às colocações;
b) elaborados levando-se em consideração o seguinte:
   - a Abordagem DDL, que propõe que o aluno infira regras e usos de fenômenos linguísticos a partir da análise de *corpus* (JOHNS, 1991);

[12] Disponível em: <https://www.wordandphrase.info>. Acesso em: 2 out. 2017.

- o uso de linhas de concordância para o ensino (BERBER SARDINHA, 2004; GAVIOLI, 2005; TRIBBLE; JONES 1997);
- os "três I's" – *Illustration, Interaction and Induction* (CARTER; McCARTHY, cf. XIAO; McENERY, 2005), enfatizando a produção de conhecimento de modo ascendente, com orientação indutiva;
- a Modelagem (*Modeling*), que postula modelar os dados a serem ensinados com base em padrões autênticos – ou seja, *corpora* (CARTER, 1998).

A ferramenta WPI, disponível no *site* do COCA, é um concordanciador *on-line* que analisa a frequência das palavras de qualquer texto em inglês providenciado pelo pesquisador, usando o COCA como *corpus* de referência. Essa ferramenta permite identificar as palavras e colocações mais frequentes, de modo mais fácil. A interface do programa é simples e *user-friendly*, como podemos notar na Figura 4.

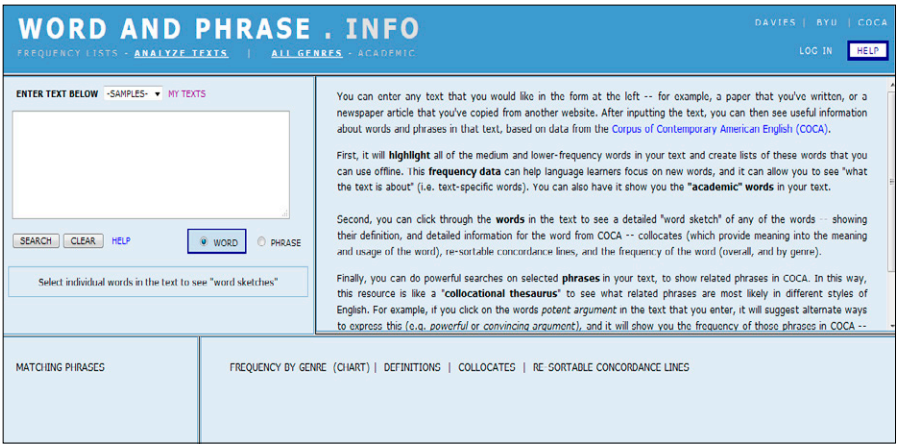


Figura 4 – Interface do programa Word and Phrase.Info
Fonte: Elaborada pela autora

Para iniciar a pesquisa, basta digitar ou colar o texto que desejamos analisar no espaço em branco e escolher se queremos que a análise seja feita por palavras (*Word*) ou por frases (*Phrase*). Após isso, basta clicar no botão *search* (busca), conforme Figura 5.

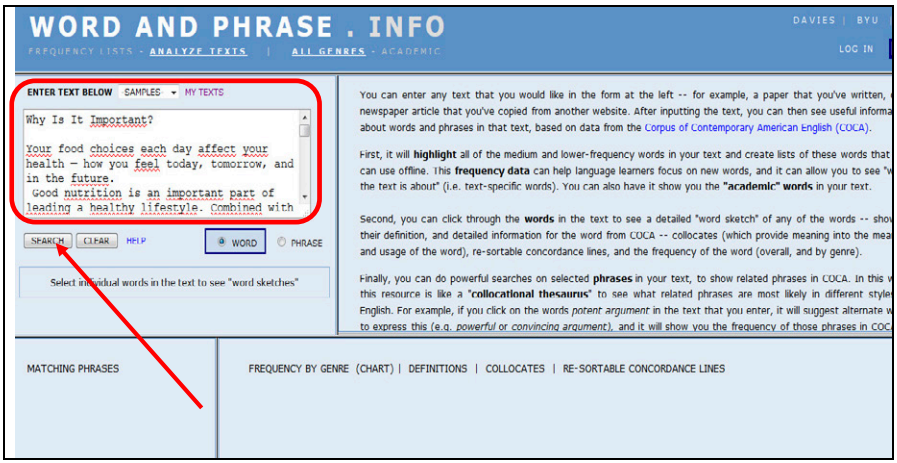


Figura 5 – Interface do programa WPI – iniciando a análise do texto
Fonte: Elaborada pela autora

Em poucos segundos, a ferramenta apresenta, em um espaço com fundo azul escuro à direita, as palavras do texto analisadas em três diferentes escalas (*ranges*), classificando-as em cores diferentes, conforme podemos notar nas Figuras 6 e 7:

- a Escala 1 (palavras em azul) compreende as palavras mais comuns da língua, tais como *the*, *with* e *take*;
- a Escala 2 (palavras em verde) contém as palavras do texto que são frequentes no COCA;
- a Escala 3 (palavras em amarelo – atualmente, essas palavras também podem ser convertidas para o vermelho, para auxiliar a visualização, se necessário) apresenta as palavras frequentes no texto, mas que muitas vezes possuem baixa frequência no COCA. Tais palavras remetem ao tema do texto analisado.

Figura 6 – Interface do programa WPI – análise do texto concluída
Fonte: Elaborada pela autora

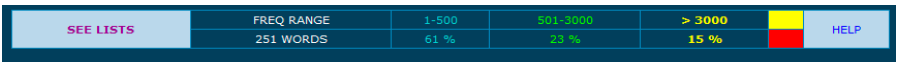


Figura 7 – Interface do programa WPI – Escala de frequência (*Freq range*)
Fonte: Elaborada pela autora

A fim de facilitar o trabalho com o texto analisado, o espaço em azul escuro é expansível, conforme a Figura 8.

Figura 8 – Interface do programa WPI – Expansão da análise do texto
Fonte: Elaborada pela autora

Com a ferramenta apresentada, passemos agora a um exemplo de elaboração de um LDI.

6.1.1 A elaboração de LDI

Para ilustrar o que temos desenvolvido, vejamos por exemplo, a Unidade 3 – Capítulo 7 do LD de inglês para o 1º Ano do EM, do SME, cujo tema é "alimentação saudável". Os capítulos sempre se iniciam com a seção intitulada *Reading*, para a leitura, a interpretação e o trabalho com o gênero de um texto autêntico, ligado à temática das unidades e dos capítulos. Após a escolha de um texto adequado à temática do capítulo, ele é analisado na ferramenta WPI.

No caso do Capítulo 7, o texto de leitura é *Food Revolution Day: exclusive interview with Jamie Oliver*, do qual obtivemos os seguintes resultados (Figura 9), após a análise com a ferramenta do COCA:

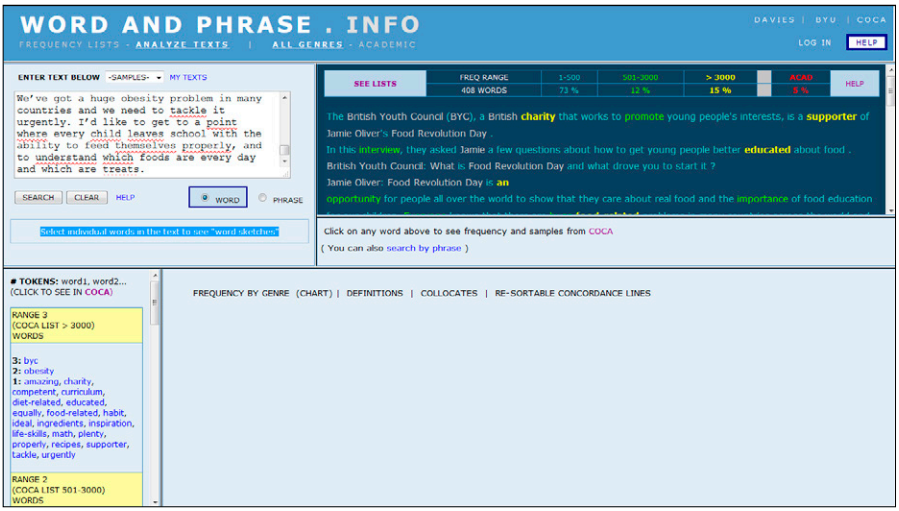


Figura 9 – Interface do programa WPI – análise do texto *Food Revolution Day: exclusive interview with Jamie Oliver*
Fonte: Elaborada pela autora

A fim de facilitar a visualização, expandimos a lista das palavras mais frequentes do texto (*Tokens*), na Figura 10. A lista é descendente e de acordo com o número de vezes que uma palavra ocorre. Por exemplo, na Escala 3, apenas uma palavra (na verdade, a abreviatura BYC) ocorreu três vezes, enquanto a maioria ocorreu uma vez; na Escala 1, notamos que a preposição *to* ocorreu 21 vezes, enquanto a segunda mais frequente, o artigo *the*, 13 vezes.

Figura 10 – Interface do programa WPI – listas das palavras mais frequentes do texto *Food Revolution Day: exclusive interview with Jamie Oliver*
Fonte: Elaborada pela autora

Com a lista das palavras mais frequentes em mãos, fazemos uma análise, a fim de verificar qual o desafio em relação ao nível linguístico e à maturidade dos alunos que lerão esse texto. Por exemplo, consideramos o seguinte:

a) se as palavras da Escala 3, ligadas ao tema, são restritas a esse tema e de difícil compreensão; em caso afirmativo, vão para o boxe intitulado *Glossary*, sem que seja inclusa em algum exercício específico. Por exemplo, *charity*.
b) ainda em relação à Escala 3, notamos que a palavra mais frequente, *byc*, é a abreviação de *British Youth Council*, organização que entrevistou Jamie Oliver para o artigo; desse modo, ela também não é inclusa nesse exercício. Outra mais frequente, *obesity*, por ser cognata, também não precisa de um exercício específico.
c) em relação à Escala 2, observamos que as palavras mais frequentes *cook*, *cooking* e *teach* já são conhecidas do público-alvo; assim, preferimos focar em outras que estejam relacionadas ao tema e precisem ser praticadas.
d) quanto à Escala 1, como já havíamos explicado na seção anterior, verificamos que há a concentração de palavras gramaticais. No caso desse texto, *to*, *the*, *a* e *and* são as mais recorrentes.

Após fazermos essa avaliação do que a ferramenta selecionou, chegamos à seguinte divisão:

– palavras da Escala 2, trabalhadas nos exercícios de compreensão lexical: *host, rise, disease, treat* e *recipe*;
– palavras da Escala 3, itens do glossário: *charity, supporter, life-skill, plenty, to achieve* e *to tackle (with).*

Acreditamos que a explicação do modo de seleção do léxico a ser trabalhado no capítulo demonstre a importância do trabalho de pesquisa feita, derrubando mitos de que trabalhar com LC significa apenas extrair listas de palavras mais frequentes de um concordanciador e passá-las aos alunos. Pelo contrário, o pesquisador / linguista / professor / autor de LD precisa estar atento a características linguísticas, metodológicas e pedagógicas, assim como a seu público-alvo, a fim de elaborar exercícios que respondam às necessidades de todos os elementos envolvidos.

Por fim, seguidos os passos supramencionados, os exercícios de compreensão lexical para a seção *Reading* do Capítulo 7 compreendem:

1. Relacionar as palavras selecionadas da Escala 2 à sua definição.
2. Preencher lacunas (*gap-filling*), em que as palavras do exercício 1 são praticadas em outro contexto, mas com o mesmo significado e função pragmática. As frases para essa prática são retiradas do COCA, em vez de serem inventadas pelo autor do LD.
3. Praticar por escrito e/ou oralmente questões que usam o léxico estudado, inclusive, aproximando-o do contexto cultural do aluno.

Ainda no Capítulo 7, a seção lexical do LD (intitulada *Vocabulary*) trata de *cooking verbs* (verbos ligados ao ato de cozinhar) e contém exercícios que convidam os alunos a prestarem atenção a linhas de concordância do COCA, a fim de encontrar colocações para cada grupo de verbos apresentado. Na Figura 11, apresentamos um exemplo de concordâncias com o verbo *bake*.

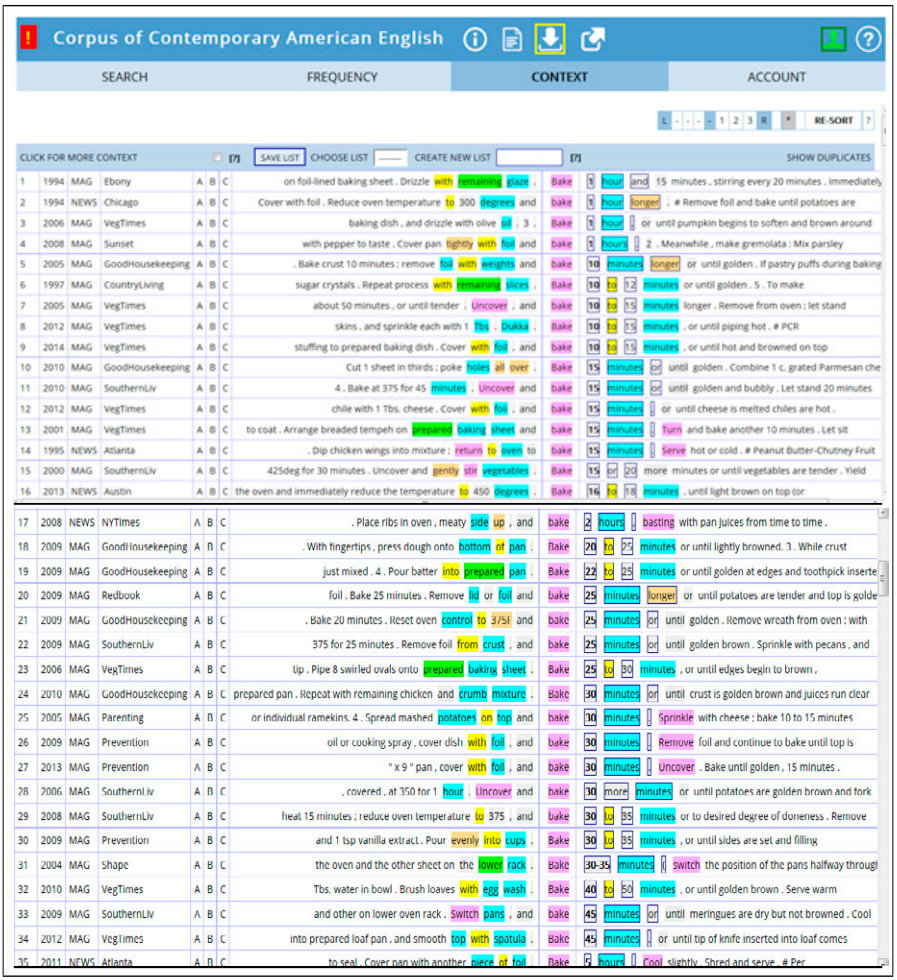


Figura 11 – Interface do COCA – linhas de concordância do verbo *bake*, alinhadas à direita
Fonte: Elaborada pela autora

Na diagramação do LD, por dispormos de pouco espaço, escolhemos as linhas de concordância relacionadas ao tema, apresentando-as de um modo mais estilizado, como podemos notar na Figura 12.

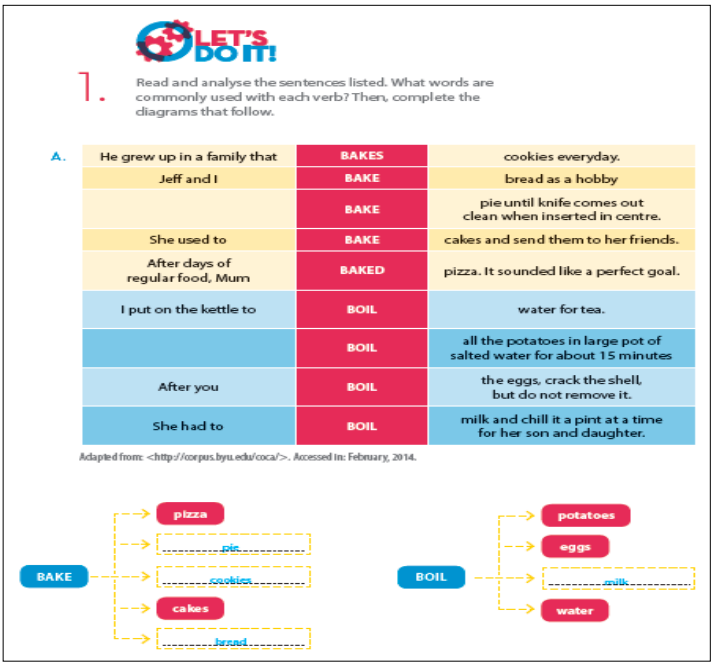


Figura 12 – Linhas de concordância estilizadas dos verbos *bake* e *boil*
Fonte: Elaborada pela autora

Em um primeiro momento, os exercícios convidam os alunos a ler as concordâncias e identificar quais palavras acompanham cada *cooking verb*. Por exemplo, *bake cookies* / *pie* / *bread* / *cakes*; *boil water* / *potatoes* / *eggs*; *chop onions* / *vegetables* / *tomatoes* / *garlic*. Assim, a partir dessas inferências, os alunos brasileiros poderão notar de modo indutivo que, embora em português digamos "assar bolo" e "assar (um) frango", em inglês usamos *bake (a) cake*, porém *roast chicken*.

Posteriormente, o nível de desafio é elevado, pois pede-se que os alunos retornem às concordâncias e: a) encontrem o verbo que significa "*to fill meat or vegetables with small pieces of another type of food*" – *to stuff* ("rechear"); b) onde e como as pessoas "*fry food*" ("fritam alimentos") – *on the griddle, in hot oil or without oil*. Há outras sugestões no Manual do Professor, caso o professor disponha de tempo para aprofundar o estudo das concordâncias com os *cooking verbs*. A seção caminha, então, para a prática desses verbos, em exercícios de *gap-filling* que usam receitas culinárias, assim como questões relacionadas à realidade do aluno.

A seção gramatical desse capítulo do LD (intitulada *Working on gramar*) também usa linhas de concordância do COCA para apresentar o uso do Imperativo. No caso, a maioria dos exemplos é proveniente de receitas de fontes variadas

disponíveis no *corpus*. A seleção das concordâncias seguiu o mesmo critério empregado para a seção *Vocabulary*.

Pelo que foi exposto, demonstramos que é possível nos valermos de textos autênticos e dos pressupostos da LC para elaborar um LD. Procuramos, ainda, nos enquadrar nos 3 Cs para o uso de texto autêntico, postulados por Mishan, já que o LD elaborado apresenta:

a) elementos culturais da língua-alvo (o chefe inglês Jamie Oliver discorrendo sobre sua campanha contra a obesidade no Reino Unido; receitas originais) e da cultura brasileira, pois a todo o momento o aluno é convidado a refletir sobre sua própria realidade.

b) elementos atuais, com a linguagem corrente, já que os exercícios põem os alunos em contato com textos jornalísticos, receitas e textos oriundos do COCA.

c) elementos de desafio, providenciados pelo contato dos alunos com textos autênticos e pelo trabalho indutivo desenvolvido com a observação das concordâncias, já que os alunos podem tirar suas próprias conclusões acerca do fenômeno estudado.

A elaboração de um LD é trabalhosa, mas mais trabalhoso ainda é fazer com que esse material seja aceito por seu público consumidor, professores e alunos. Daí a importância de se trabalhar com os professores que utilizam esse LD para que possam entender as implicações metodológicas e pedagógicas levadas em consideração na elaboração desse material. Para tanto, descreveremos na próxima seção como temos feito isso em oficinas para professores.

6.2 Elaboração e condução de oficinas

Nesta seção, demonstraremos como temos desenvolvido oficinas para professores de inglês do EFII e EM, visando a sensibilizá-los quanto à importância de se trabalhar com textos autênticos. Além disso, apresentamos ou revisamos os pressupostos da LC, chamando a atenção quanto à utilidade das ferramentas baseadas em LC para realizar tal tarefa.

Para as oficinas, tivemos dois grupos diferentes, A (cinco docentes) e B (sete docentes), todos do Estado de São Paulo e profissionais das escolas que utilizam o material do SME. Em diferentes datas, os dois grupos trabalharam com dois textos autênticos adaptados[13]: grupo A, com o texto *Myths and legends: the Loch Ness monster*; grupo B, com o texto *The impact of nutrition on your health*. Esses textos fazem parte do livro do 8º Ano do SME que estava em elaboração e, portanto, com o qual os docentes ainda não tinham tido contato.

[13] Por causa de restrições editoriais (número de páginas), os textos são adaptados no que se refere à supressão de trechos ou parágrafos. Contudo, características, tais como gênero e léxico, não são alteradas.

Feita essa divisão, as oficinas ocorreram da seguinte forma:

- Oficina I (doravante I): os grupos A e B leram os textos e propuseram dois ou três exercícios que explorassem a compreensão linguística (lexical) de seus alunos.
- Oficina II (doravante II), continuação de I e com os mesmos grupos: A e B observaram como os textos em que eles trabalharam em I foram efetivamente abordados no livro do 8º Ano – ou seja, quais exercícios foram propostos, levando-se em consideração a convencionalidade (mais especificamente, as colocações presentes nos textos).

Ainda em II, os docentes tiveram o primeiro contato com ferramentas da Linguística de *Corpus*, já que os exercícios propostos para a compreensão linguística dos textos autênticos do livro de inglês do 8º Ano são:

a) baseados na análise obtida na ferramenta WPI, no tocante à frequência das palavras e às colocações.
b) elaborados levando-se em consideração o seguinte:
   - o ensino DDL (*Data-Driven Learning*), ou Aprendizagem Direcionada por Dados, que propõe que o aluno infira regras e usos de fenômenos linguísticos a partir da análise de *corpus* (JOHNS, 1991);
   - o uso de linhas de concordância para o ensino (TRIBBLE; JONES 1997; BERBER SARDINHA, 2004; GAVIOLI, 2005;);
   - os "três I's" – *Illustration, Interaction and Induction* (CARTER; McCARTHY, cf. XIAO; McENERY, 2005);
   - a Modelagem, que postula modelar os dados dos padrões a serem ensinados, com base em dados autênticos – ou seja, *corpora* (CARTER, 1998).

Todavia, considerando-se a quantidade de informações e a duração da Oficina II (quatro horas), apenas introduzimos o conceito do que é LC e para que pode ser utilizada, discorrendo brevemente sobre os concordanciadores, enfim, a base teórica da abordagem do material – e que será aprofundada nas próximas oficinas, focando também na ferramenta WPI.

7 Análise dos resultados

Descreveremos, nesta seção, os resultados obtidos nas duas oficinas-piloto, I e II.

Na I, tanto os grupos A quanto B propuseram exercícios em que os alunos deveriam sublinhar ou copiar as palavras cognatas, por exemplo:

- Texto A: *famous, enormous* e *creature*.
- Texto B: *important, reduce* e *adults*.

Além do exercício de reconhecimento dos cognatos, o grupo A ainda solicitou que os alunos lessem o que haviam sublinhado e apontassem a ideia principal do texto. O grupo B, todavia, não deu sequência ao trabalho com os cognatos, preferindo trabalhar com outra lista de palavras do texto: *food, today, tomorrow, future, health(y), dietary, child / adulthood, carry* e *benefits*.

Embora o levantamento dos cognatos seja uma técnica importante para auxiliar o aluno a compreender um texto, pudemos notar que a preocupação maior dos docentes era trabalhar "listas" – ou seja, reconhecer palavras isoladas, não explorando suas (possíveis) relações, tampouco sua classe morfológica e/ou gramatical. Tal preocupação ficou evidente principalmente no caso do grupo B, que abandonou a primeira lista (cognatos) por uma segunda. Quando perguntados qual o critério utilizado para a escolha dessa segunda lista, os docentes informaram que julgavam tais palavras "importantes", sem elaborar o porquê. Insistimos que *food, today* e *tomorrow* são bem conhecidas, já que os alunos do 8º Ano não são iniciantes em inglês. Não obtivemos resposta à nossa indagação.

As outras questões propostas pelos dois grupos saíram do escopo que havíamos delineado (exercícios de compreensão linguística, com foco no léxico): o grupo A solicitou que os alunos sublinhassem todos os verbos no *Simple Past*, classificando-os em *regular* e *irregular*. Já o grupo B apresentou quatro perguntas relacionadas à compreensão das ideias do texto.

Em II, distribuímos os textos novamente, e sua versão já analisada na ferramenta WP, conforme as Figuras 13 e 14.

Figura 13 – Reprodução de Texto A (*Myths and legends: the Loch Ness monster*), analisado em WPI
Fonte: Elaborada pela autora

Figura 14 – Reprodução de Texto B (*Why is it important?*), analisado em WPI
Fonte: Elaborada pela autora

Os docentes de ambos os grupos ficaram "maravilhados" quando viram os textos com as palavras destacadas, de acordo com a frequência. Disseram ainda que tal organização facilitava a visualização de expressões, algo que eles não haviam notado quando do trabalho com os textos na Oficina I. Por exemplo:

- Em A: *spread rumours, take a photograph, conduct a survey* e *conclusive proof*.
- Em B: (*un*)*healthy lifestyle* / *eating habits* / *weight* e *chronic* / *heart disease*(*s*).

Apresentamos, então, como essas colocações foram abordadas no LD do SME, demonstrando o modo como usamos o COCA para nos auxiliar na busca por outras possíveis colocações e outros exemplos. Inicialmente, introduzimos a interface da ferramenta WPI e explicamos como analisar um texto nela (para exemplificar, mostramos neste trabalho o Texto B). Depois, projetamos o texto já analisado, conforme Figura 15; e ampliamos a análise do texto (Figura 16) e da lista de frequência das palavras do texto (Figura 17).

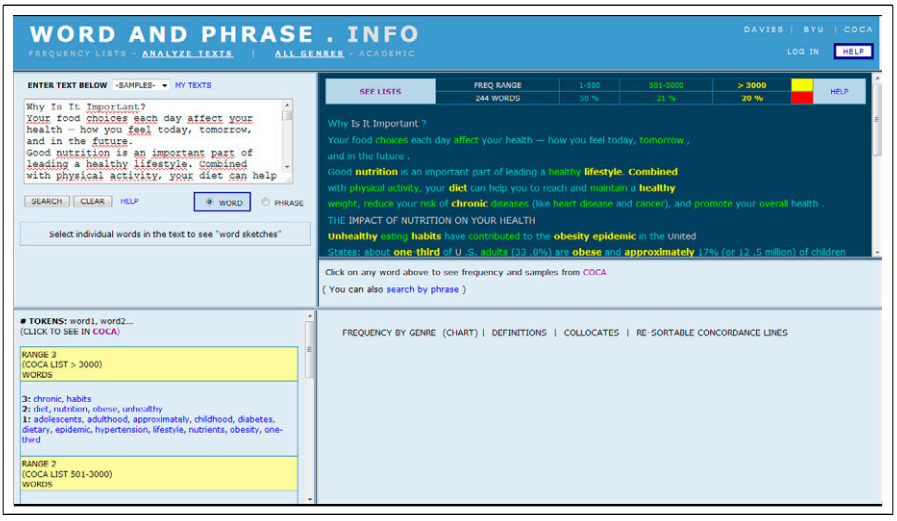


Figura 15 – WPI com a análise do Texto B
Fonte: Elaborada pela autora

Figura 16 – WPI com o Texto B ampliado
Fonte: Elaborada pela autora

Figura 17 – WPI com a lista de frequência de palavras do Texto B
Fonte: Elaborada pela autora

Com base na leitura da análise do texto (em que ficaram evidentes colocações como *chronic / heart disease(s)* e *healthy lifestyle / weight*) e na lista de frequência de palavras, explicamos como selecionamos as palavras que comporiam o glossário: *leading, one-third, increasingly, dietary, throughout, overall*. Para o trabalho lexical,

escolhemos as mais frequentes nas Escalas 3 (*chronic*, *lifestyle*) e 2 (*healthy*, *weight* e *disease*).

Focamos, então, na palavra lexical que mais ocorreu no texto, *healthy* (sete ocorrências). Clicamos nela e obtivemos duzentas linhas de concordância (a palavra possui 32.628 ocorrências no total, no COCA), como podemos observar na parte de baixo da interface, na Figura 18, e essas mesmas concordâncias expandidas na Figura 19. Ainda na Figura 17, destacamos, à esquerda das concordâncias, a distribuição da ocorrência de *healthy* por gênero (Spok – *spoken* – textos orais; Fic – *fiction* – ficção; Mag – *magazines* – revistas; News – *newspapers* – jornais; Acad – *academic* – produção acadêmica), com prevalência nas revistas (10.103 ocorrências).

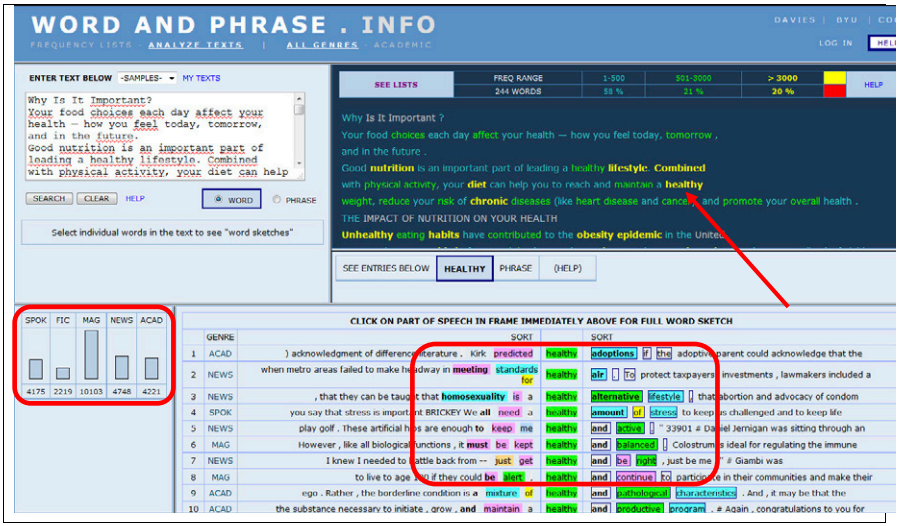


Figura 18 – Texto B e concordâncias com *healthy*
Fonte: Elaborada pela autora

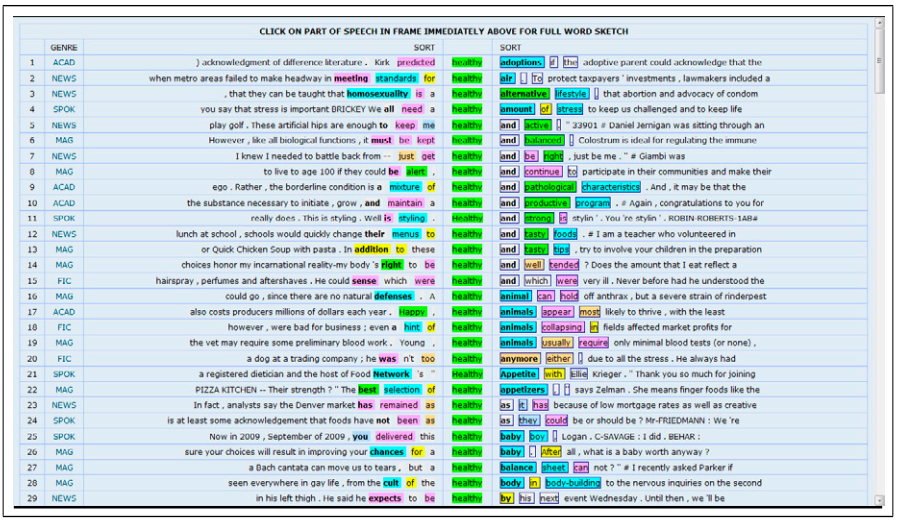


Figura 19 – Concordâncias com *healthy* (expandido, alinhado pela direita)
Fonte: Elaborada pela autora

Prosseguimos com as explicações, apontando para as ocorrências das seguintes colocações com *healthy*, que haviam aparecido no Texto B:

– ~ *lifestyle* (367 oc. no COCA) – Figura 20;
– ~ *weight* (345 oc. no COCA) – Figura 21;
– EAT[14] ~ (279 oc. no COCA) – Figura 22;
– STAY ~ (839 oc. no COCA) – Figura 23.

[14] A grafia em versalete indica que estamos tratando do lema – ou seja, no caso de *eat*, de todas as ocorrências de suas formas verbais.

Figura 20 – Concordâncias com *healthy*, com foco em *healthy lifestyle* (alinhamento à direita)
Fonte: Elaborada pela autora

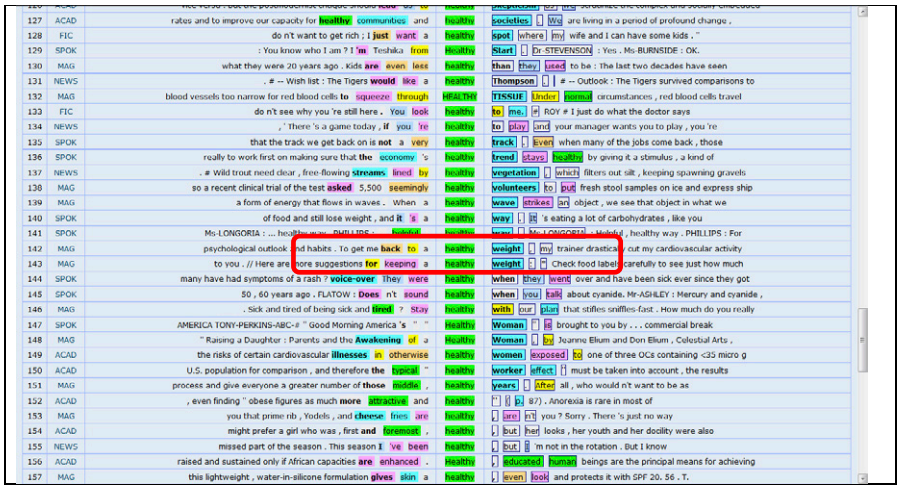


Figura 21 – Concordâncias com *healthy*, com foco em *healthy weight* (alinhamento à direita)
Fonte: Elaborada pela autora

Figura 22 – Concordâncias com *healthy*, com foco em EAT *healthy* (alinhamento à esquerda)
Fonte: Elaborada pela autora

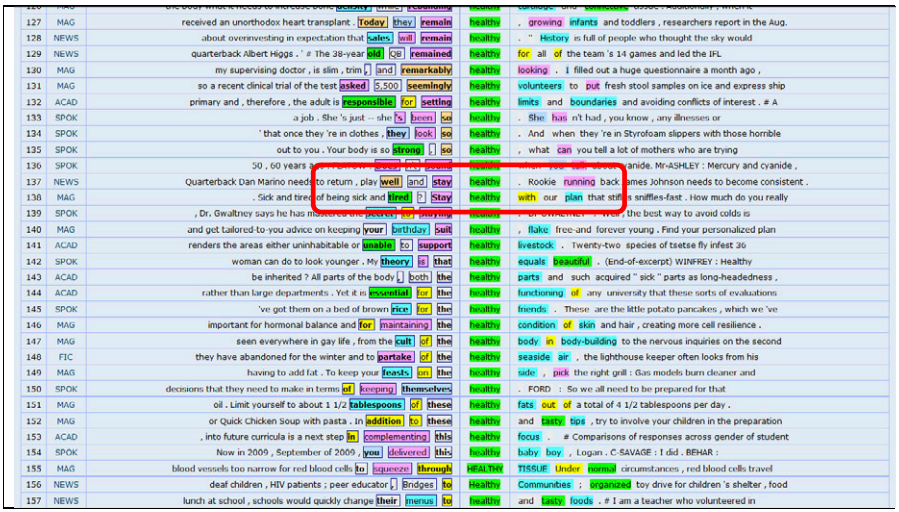


Figura 23 – Concordâncias com *healthy*, com foco em STAY *healthy* (alinhamento à esquerda)
Fonte: Elaborada pela autora

Fizemos o mesmo com *disease*: clicamos na palavra e obtivemos duzentas linhas de concordância (a palavra possui 56.970 ocorrências no COCA), como podemos observar na parte de baixo da interface, na Figura 24. Ainda nessa figura, destacamos, à esquerda das concordâncias, a distribuição da ocorrência de *disease* por gênero, com predominância nas revistas (15.258 ocorrências).

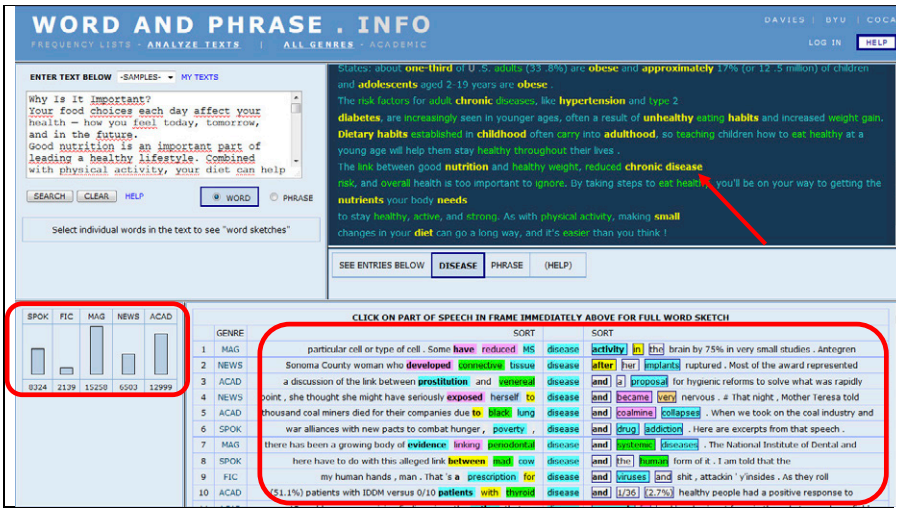


Figura 24 – Texto B e concordâncias com *disease*
Fonte: Elaborada pela autora

Apontamos, então, para as ocorrências das seguintes colocações com *disease*, que haviam aparecido no Texto B:

– *chronic* DISEASE (1.106 oc. no COCA) – Figura 25;
– *heart* DISEASE: (6.314 oc. no COCA) – Figura 25;
– *risk of* ~ DISEASE: (784 oc. no COCA) – Figura 26.

Figura 25 – Concordâncias com *disease*, com foco em *chronic* DISEASE e *heart* DISEASE (alinhamento à esquerda)
Fonte: Elaborada pela autora

1	2011	ACAD	SocialWork	A	B	C	, I. (2006) .	Burnout	and	risk	of cardiovascular disease	: Evidence , possible causal paths , and pr
2	2011	ACAD	AmJPubHealth	A	B	C	to influence physical	inactivity	and	risk	of cardiovascular diseases	(see the box on this page) . # Strategies
3	2007	MAG	Prevention	A	B	C	protective effect of hormones like estrogen	,	and	risk	of heart disease	climbs steeply . " DAY TWO Stress on the
4	2009	NEWS	USAToday	A	B	C	otherwise healthy people with	no	apparent	risk	of severe disease	. Peri calls some cases 1918 esque , refer
5	2012	ACAD	PracticeNurse	A	B	C	that she has no idea she	is	at	risk	of serious disease	. # When you next see Mr Martin he expl
6	2002	SPOK	NPR_ATC	A	B	C	they asked more than 12,000	middle-aged	men	at	risk	of heart disease	if they had had a vacation the previous y
7	2011	ACAD	PracticeNurse	A	B	C	Consequently they are	more	at	risk	of mosquito-transmitted diseases	such as dengue fever , malaria and yellov
8	2011	ACAD	PracticeNurse	A	B	C	: FIGURE 1 . IDENTIFYING	PATIENTS	AT	RISK	OF CARDIOVASCULAR DISEASE	(CVD) 2 # DIAGRAM : FIGURE 2 . DRUG T
9	2009	MAG	GoodHousekeeping	A	B	C	M.S. , R.D. , who counsels	people	at	risk	of heart disease	. And if you have n't already banished tra
10	2008	MAG	USAToday	A	B	C	change to the total future	global	population	at	risk	of that disease	. For that , one has to use the results fror
11	2006	NEWS	SanFranChron	A	B	C	source . # Adult men and	women	at	risk	of heart disease	: -- Can reduce risk of heart disease by ea
12	2012	MAG	Prevention	A	B	C	" Undiagnosed thyroid problems can put	you	at	risk	of heart disease	, infertility , and miscarriage " # Too little
13	2013	MAG	Prevention	A	B	C	and high cholesterol , that puts	you	at	risk	of heart disease	. # 3 . Research has linked the acidity in c
14	2008	ACAD	PhysicalEduc	A	B	C	may be related to decreased	BMI	,	decreased	risk	of chronic diseases	and associated mortality , and decreased
15	1994	MAG	Prevention	A	B	C	to 83 deaths . There was	no	elevated	risk	of heart disease	among women whose BMIs were under
16	2011	ACAD	AmJPubHealth	A	B	C	, Willett WC , Manson JE	,	Elevated	risk	of cardiovascular disease	prior to clinical diagnosis of type 2 diabet
17	2002	SPOK	NPR_Science	A	B	C	second year , that there was	an	excess	risk	of heart disease	and stroke , but they were very small nur
18	2012	ACAD	EnvironmentalHealth	A	B	C	with 0.76% absolute	reduction	in	excess	risk	of cardiovascular disease	for every 10% increase in air conditioner
19	2004	NEWS	Denver	A	B	C	women who do n't exercise have	a	greater	risk	of heart disease	than fewer women who do exercise . Th

Figura 26 – Concordâncias com disease, com foco em *risk of ~ DISEASE* (alinhamento à esquerda)
Fonte: Elaborada pela autora

Finalmente, os docentes tiveram acesso a alguns dos exercícios, já diagramados no livro, conforme Figuras 27 e 28.

3. Read the fragments from the text.

I. In the 1960s, the Loch Ness Investigation Bureau **conducted** a ten-year observational survey.

II. Over the years, local population **spread** rumours far and wide about "strange events" at Loch Ness.

III. To this day, there is no conclusive proof to suggest that the monster is a reality.

IV. The first recorded **sighting** of the monster was in the year 565.

V. Around this time, there were many independent **sightings** [...].

- Now,

a) Underline the noun that completes the idea of the verb **conduct** in sentence I.

b) Underline the noun that completes the idea of the verb **spread** in sentence II.

c) Underline the adjective that describes the noun **proof** in sentence III.

d) Underline the adjectives that describe the noun **sighting** in sentences IV and V.

4. Read other examples of collocations with conduct, spread, proof and sighting. Then, complete the diagrams.

Scientists affirm they are	**conducting**	a research like this in more than ten countries.
Dr. Hudd	**conducted**	a study focusing on stress among university students.
The detectives affirmed that they would	**conduct**	an investigation next week.

Adapted from: <http://corpus.byu.edu/coca>. Accessed in: September, 2016.

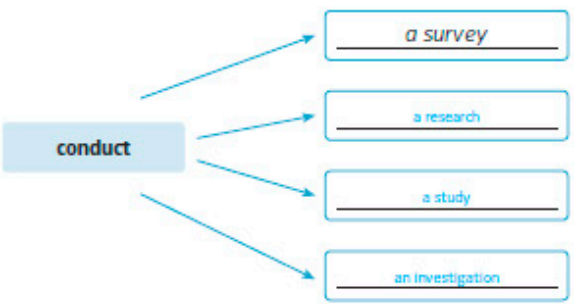


Figura 27 – Exercícios com o Texto A – reprodução do livro do 8º Ano do SME
Fonte: Elaborada pela autora

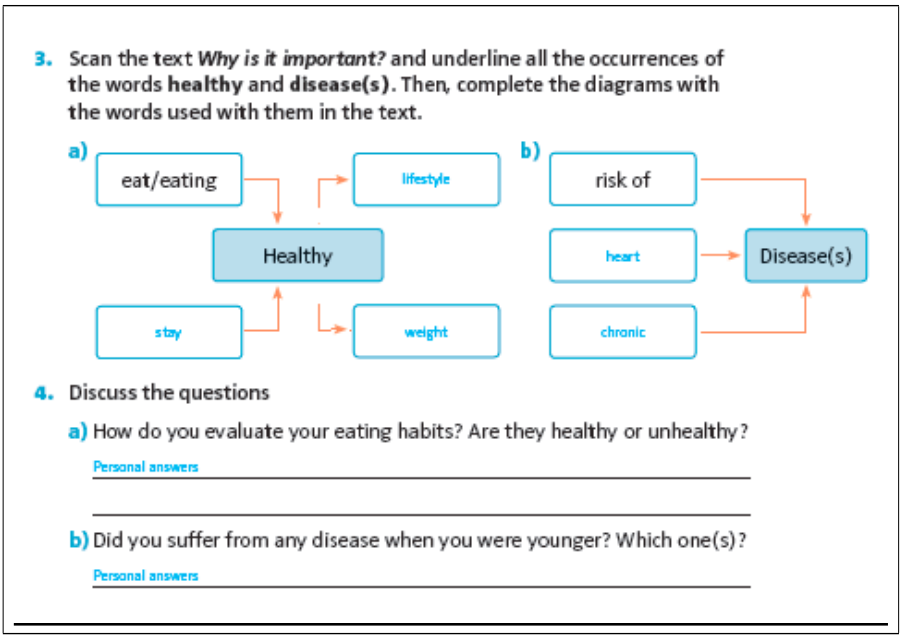


Figura 28 – Exercícios com o Texto B – reprodução do livro do 8º Ano do SME
Fonte: Elaborada pela autora

Os docentes afirmaram, então, que passaram a entender qual era a proposta do livro do SME: abordar poucas palavras, em vez de listas, mas aprofundar seu estudo, por meio da análise do modo como elas se relacionam com as outras palavras à sua volta em um contexto autêntico de uso.

Para a oficina III, ainda a ser realizada, solicitamos que os docentes escolham um texto autêntico curto, façam sua análise no WPI e elaborem duas ou três questões para analisar a compreensão linguística desse texto. Esperamos que tais questões possam refletir o que os docentes aprenderam na Oficina II.

8 Considerações finais

Neste trabalho, procuramos demonstrar como temos elaborado oficinas para sensibilizar os professores de inglês quanto à importância de se usar textos autênticos e ferramentas da LC para o ensino de línguas. Nessas oficinas, pretendemos que esses profissionais possam também entender como o LD para o ensino de inglês do SME é elaborado com base nesses pressupostos, diferindo da abordagem tradicional do mercado de livros de ELT, que pouco explora a convencionalidade.

Sabemos que esse é apenas um recorte do trabalho que teremos de fazer com os professores das mais de duzentas escolas que utilizam esse material constituído

de textos autênticos e informado por *corpus*. Temos ciência também de que nem todos se sentirão "maravilhados" com os resultados obtidos com a pesquisa em *corpus*, seja porque não "gostam de lidar com computador", porque "não têm tempo" para pesquisa, ou porque preferem "tudo pronto".

Contudo, as duas oficinas-piloto confirmam o que McCarthy (2008) afirmou há uma década: sem investir na capacitação docente para o uso de ferramentas da Linguística de *Corpus*, não adiantará nada publicar material baseado em *corpora*.

Referências

AIJMER, K. Corpora *and language teaching*. Amsterdam: John Benjamins Publishing, 2009. (Studies in *Corpus* Linguistics, Vol.33).

ASTON, G. *Corpora* in language pedagogy: matching theory and practice. In: COOK, G.; SEIDLHOFER, B. (Org.). *Principle and practice in applied linguistic*: studies in honour of H.G. Widdowson. Oxford, Oxford University Press, 1995, p. 257-270.

BERBER SARDINHA, A. P. *Linguística de* Corpus. Barueri: Manole, 2004.

BURTON, G. *Corpora* and coursebooks: destined to be strangers forever? *Corpora 2012*, v. 7 (1), p. 91-108, 2012.

CARTER, R. Orders of reality: Cancode, communication and culture. *ELT Journal*, Oxford, v. 52, n. 1, p. 43-56, jan. 1998.

CORACINI, M. J. R. F. *Interpretação, autoria e legitimação do livro didático*: língua materna e língua estrangeira. 1. ed. Campinas: Pontes, 1999.

FIRTH, J. R. Modes of Meaning. *Papers in Linguistics 1934-51*. Oxford: Oxford University Press, 1999, p. 190-215.

GAVIOLI, L. *Exploring* Corpora *for ESP Learning*. Amsterdam: John Benjamins Publishing, 2005. (Studies in *Corpus* Linguistics, Vol. 21).

GRIGOLETTO, M. Leitura e funcionamento discursivo do livro didático. In: CORACINI, M. J. *Interpretação, autoria e legitimação do livro didático*: língua materna e língua estrangeira. 1. ed. Campinas: Pontes, 1999, p. 67-77.

HANNA, V. L. H. *Línguas estrangeiras*: o ensino em um contexto cultural. São Paulo: Editora Mackenzie, 2012. (Col. Conexão Inicial, Vol. 2).

JOHNS, T. Should you be persuaded: two samples of data-driven learning materials. In: JOHNS, T. J.; KING, P. (Ed.). Classroom Concordancing. *ELR Journal 4*. Birmingham: University of Birmingham, 1991, p. 1-16.

KRAMSCH, C. J. The cultural discourse of foreign language textbooks. In: TÍLIO, R.; FERREIRA, A. de J. (Ed.). *Innovations and challenges in language teaching and materials development* – Inovações e desafios na produção de materiais didáticos para o ensino de línguas. Campinas: Pontes Editores, 2017, p. 13-58.

MCCARTHY, M. Accessing and interpreting *corpus* information in the teacher education context. *Language Teaching*, 41(4), p. 563-574, 2008.

McENERY, T.; XIAO, R.; TONO, Y. Corpus-*based Language Studies*: An advanced resource book. London: Routledge, 2006. (Routledge Applied Linguistics series).

MISHAN, F. *Designing Authenticity into Language Learning Materials*. Bristol: Intellect Books, 2005.

O'KEEFE, A.; MCCARTHY, M.; CARTER, R. *From* Corpus *to Classroom:* language use and language teaching. Cambridge: Cambridge University Press, 2007.

SANTOS, A. G. dos. *Working closely with* corpora. *Proposta de ensino de colocações adverbiais, sob a luz da Linguística de* Corpus. Dissertação (mestrado em Letras). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2011.

_____. Developing ELT coursebooks with *corpora*: the case of "Sistema Mackenzie de Ensino". In: FORMATO, F.; HARDIE, A. (Ed.). Corpus *Linguistics 2015*. Lancaster: UCREL, 2015, p. 293-294.

SCOTT, M. Aprendizagem direcionada por dados: uma homenagem a Tim Johns (1936-2009). In: VIANA, V.; TAGNIN, S. (Org.). Corpora *no ensino de línguas estrangeiras*. São Paulo: Hub Editorial, 2010, p. 7-11.

SEIDLHOFER, B. (Ed.) *Controversies in Applied Linguistics*. Oxford: Oxford Universtiy Press, 2003.

SILVA, B. F. da; LIMA, C. R. V. *Inglês 8º Ano, livro 2*. 2. ed. Vol. 2. Ed. Pedagógica: Andréa Geroldo dos Santos. São Paulo: Sistema Mackenzie de Ensino, 2017. 232p. (Col. Crescer em Sabedoria).

SINCLAIR, J. M.; RENOUF, A. A lexical syllabus for language learning. In: CARTER, R.; McCARTHY, M. (Org.). *Vocabulary and language teaching*. London: Longman, 1988, p. 140-160.

SINCLAIR, J. Lexical grammar. *Naujoji Metodologija*, 24, p. 191-203, 2000. Disponível em <http://donelaitis.vdu.lt/publikacijos/sinclair.pdf>. Acesso em: mar. 2010.

SOUZA, D. M. de. Do monumento ao documento. In: CORACINI, M. J. R. F. (Org.). *O jogo discursivo na aula de leitura*: língua materna e língua estrangeira. 2. ed. Campinas: Pontes, 2002, p. 113-117.

TAGNIN, S. E. O. *O jeito que a gente diz*: combinações consagradas em inglês e português. São Paulo: Disal, 2013.

TITONE, R. *Teaching Foreign Languages*: An historical sketch. Washington: Georgetown University Press, 1968.

TOGNINI-BONELLI, E. Corpus *Linguistics at work*. Amsterdam: John Benjamins, 2001.

TOMLINSON, B. *Developing Materials for Language Teaching*. London: Continuum, 2003.

_____. *English Language Learning Materials*. A Critical Review. Londres: Continuum, 2008.

TRIBBLE, C.; JONES, G. *Concordances in the classroom*. A resource guide for teachers. Houston: Athelstan Publications, 1997.

WIDDOWSON, H. G. *Aspects of Language Teaching*. Oxford: Oxford University Press. 1990.

_____. On the limitations of linguistics applied. *Applied Linguistics*, 21(1), p. 3-25, 2000.

XIAO, R.; McENERY, T. Corpora *and language education*. 2005. Manuscript. Available at: <http://www.corpus4u.org/archive/index.php/t-75.htm>. Acesso em: 2 out. 2010.

Brazilian students' use of English academic vocabulary: an exploratory study

O uso de vocabulário acadêmico em língua inglesa por estudantes brasileiros: um estudo exploratório

Larissa Goulart da Silva
Marine Laisa Matte
Simone Sarmento

Abstract: The aim of this chapter is to present an investigation on how Brazilian students use academic vocabulary. The following research questions are addressed: a) What is the vocabulary profile of assignments written by Brazilian students?; b)How does it compare to the vocabulary profile of other academic *corpora*?; c)What words in the AWL do Brazilian students use?, and d) How does the use of academic words differ between Brazilian students and students represented in the British Academic Written English (BAWE) *corpus*? To answer these questions, the Brazilian Academic Written English (BrAWE) *corpus* was analysed using Range and Sketch Engine. The results point that BrAWE presents a similar number of academic words as other academic *corpora*. However, the word forms selected by these students differ as they underuse affixation processes.

Keywords: English for Academic Purposes. Academic Vocabulary. Lexical Profile.

Larissa Goulart da Silva – Professora na University of Nebraska, mestrado em Ensino de Língua Inglesa pela Warwick University, bolsista da comissão Fulbright – goulart.larissa@huskers.unl.edu.
Marine Laisa Matte – Mestranda do Programa de Pós-Graduação em Estudos da Linguagem na UFRGS, graduada em Letras pela mesma universidade – marine.laisa@gmail.com.
Simone Sarmento – Professora no Programa de Pós-Graduação em Estudos da Linguagem na UFRGS – simone.sarmento@ufrgs.br.

Resumo: Este capítulo objetiva investigar como os alunos brasileiros utilizam vocabulário acadêmico. As seguintes perguntas de pesquisa são abordadas: a) Qual é o perfil lexical dos trabalhos escritos por alunos brasileiros?; b) Como ele se compara ao perfil lexical de outros *corpora* acadêmicos?; c) Quais palavras da AWL são utilizadas por eles?, e d) Como o uso de palavras acadêmicas difere entre os alunos brasileiros e os representados no *Corpus* British Academic Written English (BAWE)? Para responder essas perguntas, o *corpus* Brazilian Academic Written English (BrAWE) foi analisado utilizando Range e Sketch Engine. Os resultados sugerem que os alunos brasileiros utilizam a mesma quantidade de palavras acadêmicas na comparação com o que aparece em outros *corpora* acadêmicos. Entretanto, as formas de palavras selecionadas por eles diferem, pois utilizam menos processos de afixação.

Palavras-chave: Inglês para fins acadêmicos. Vocabulário acadêmico. Perfil lexical.

1 Introduction

Having an extended vocabulary is essential for reading and writing in a foreign language. Some studies (LAUFER, 1989; HIRSH; NATION, 1992; HSUEH-CHO; NATION, 2000; NATION, 2013) argue that it is necessary to master around 95% to 98% of the words in a given text to understand it properly. Formal writing requires students to have a vast knowledge of academic words and to actively use them on written assignments in order to achieve academic success (KAUR; HEGELHEIMER, 2005; COXHEAD, 2011). This study uses a *corpus* linguistics approach to investigate how Brazilian students use academic vocabulary when writing assignments in British universities. It is our understanding that academic words are lexical items that are recurrent in academic texts (CHUNG; NATION, 2003; CLARK; ISHIDA, 2005) and, therefore, are not usually acquired in everyday interaction. This investigation seeks to answer the following questions:

a) What is the vocabulary profile of assignments written by Brazilian students?
b) How does it compare to the vocabulary profile of other academic *corpora*?
c) What words in the Academic Word List (AWL) do Brazilian students use?
d) How does the use of academic words differ between Brazilian students and students represented in the British Academic Written English (BAWE) *corpus*?

2 Academic vocabulary in language learning

For the purposes of this study, it is worth distinguishing between general academic vocabulary and discipline specific (academic) vocabulary. General academic vocabulary, also referred simply as academic vocabulary[1], is a group of words that are frequent in academic texts across different disciplines. On the other hand, discipline specific vocabulary or technical words/terms (FISHER; FREY, 2008; HARMON; WOOD; HEDRICK, 2008) are recurrent words in one academic domain, for example, *abdominal, laboratory* and *melanoma* are more recurrent in the field of health sciences (LEI and LIU, 2016). Beck et al. (2002) call general academic vocabulary Tier 2, and discipline specific words Tier 3. This means that these groups of words are the second and third most frequent ones in a *corpus* of academic texts.

In this study the AWL[2] will be used as it represents general academic English (HYLAND; TSE, 2007) and is widely used as a reference for writers of teaching materials, EAP researchers, and as a source for The Vocabulary Level Test (SCHMITT; SCHMITT; CLAPHAM, 2001).

2.1 AWL coverage in other corpora

Coxhead (2000) and Hyland and Tse (2007) investigated the lexical profile of texts in two academic *corpora*. They described the use of high-frequency words (GSL) and academic words (AWL) using Range[3]. In Coxhead's (2000) academic *corpus*, both the AWL and the GSL represent 86.1% of the running words in the *corpus*. The AWL accounts for 10.0% of the tokens and the GSL for 76.1%. However, there are some differences in coverage in each *subcorpus*, as can be seen in Table 1 below. Commerce had the highest AWL coverage with 12.0%, followed by arts 9.3%, law 9.4% and sciences 9.1%.

[1] *Commence*, *obtain* and *indicate* are some examples of academic words.

[2] We are aware of the criticisms the AWL has been receiving. The AWL has been criticised for a number of reasons: (i) for being dated (GARDNER and DAVIES, 2014); (ii) because it was compiled on top of the General Service List (GSL) (WEST, 1953); and (iii) for using word families (NAGY and TOWNSEND, 2012). Nevertheless, for the reasons stated in the text, we will be using it in this research.

[3] Range is a software that analyses the frequency of a group of words in a *corpus*. According to its developers "Range provides a spectrum or distribution figure, a headword frequency figure, a family frequency figure, and a frequency figure for each of the texts the words occurs in" (NATION; HEATLEY, 2002).

Table 1 – Coverage of AWL and GSL words in Coxhead's (2000) Academic *corpus* (COXHEAD, 2000, p. 224)

Subcorpus	Academic Word List	General Service List		Total
		First 1,000 words	Second 1,000 words	
Arts	9.3	73.0	4.4	86.7
Commerce	12.0	71.6	5.2	88.8
Law	9.4	75.0	4.1	88.5
Science	9.1	65.7	5.0	79.8

Hyland and Tse's (2007) academic *corpus* "includes a range of professional and learner's texts representing key academic genres across a broad span of disciplines" (p.238). In this *corpus*, the two lists combined represent 85% of the running words (AWL 10,6% and GSL 74,4%). Regarding the AWL, in every seven words one is unknown. Table 2 below depicts the coverage in each *subcorpus*. As already noticed by Coxhead (2000) the combined coverage of the AWL and the GSL is slightly smaller in the science *subcorpus* than in other *subcorpora*.

Table 2 – Coverage of AWL and GSL words in Hyland and Tse's (2007) Academic *corpus* (HYLAND; TSE, 2007, p. 240)

	Frequency words	AWL items	Mean	Coverage %		
				AWL	GSL	Overall
Engineering	551,891	61,408	108	11.1	73.3	84.4
Social Sciences	1,822,660	200,393	352	11.0	77.0	88.0
Sciences	838,926	78,234	137	9.3	69.0	78.3
Overall	3,213,477	340,035	597	10.6	74.0	84.7

Thus, based on previous investigations, it is possible to say that the AWL typically covers from 8 to 10%, while the GSL covers from 65 to 75% of an academic *corpus*. These figures will be compared to the data from the Brazilian Academic Written English (BrAWE) *corpus*.

3 Corpora

Two *corpora* will be used in this study, the British Academic Written English (BAWE) *corpus* and the Brazilian Academic Written English (BrAWE) *corpus*. The BAWE *corpus* was compiled throughout the years of 2004 and 2007 in three

British universities – Warwick, Reading and Oxford Brooke and it contains a total of 2,897 texts, and 8,506,995 tokens. The texts represent students' academic writing in English. The *corpus* is composed of assignments which were part of students' graduate and Master's course evaluations, and, in order to be included in the *corpus*, the assignments must have been graded with merit, honor or distinction. The texts were written to display content knowledge rather than language skills, and are, this way, considered as authentic material. BAWE is divided into 13 genre families – Case Study, Critique, Design Specification, Explanation, Exercise, Essay, Empathy Writing, Literature Survey, Methodology Recount, Narrative Recount, Problem Question, Proposal and Research Report – and four main areas – Social Sciences (SS), Arts and Humanities (AH), Physical Sciences (PS), and Life Sciences (LS). Texts in the BAWE *corpus* contain a series of other contextual information, such as, students' level of education, their grades, previous study background, gender, among other information. Nevertheless, students who contributed to BAWE had a financial motivation to proceed with their submission.

The Brazilian Academic Written English *Corpus* (BrAWE)[4] (SILVA, 2017) comprises assignments written by Brazilian students who were in undergraduate programs in the UK. The *corpus*, compiled in 2015 and 2016, contains 380 texts representing 59 universities in the UK (as opposed to three in BAWE), comprising 670,314 tokens. Most of the students who collaborated with the *corpus* were Science without Borders (SwB)[5] students. All assignments included in this *corpus* received at least a pass grade. As well as in BAWE, BrAWE is divided into four main areas – Social Sciences (SS), Arts and Humanities (AH), Physical Sciences (PS), and Life Sciences (LS) and 13 genre families – Case Study, Critique, Design Specification, Explanation, Exercise, Essay, Empathy Writing, Literature Survey, Methodology Recount, Narrative Recount, Problem Question, Proposal and Research Report. Table 3 depicts the distribution of texts and tokens in each BrAWE *subcorpora*. The largest *subcorpora* is the PS, followed by LS. This is due to the large number of SwB students, which were mainly from STEM areas.

[4] We would also like to highlight one study that has already been conducted using BrAWE which is Matte (2017). This author analysed the uses of connectors in BrAWE and BAWE. The results of this quantitative study show that overall there is an overuse of connectors by Brazilians represented in the *corpus*.

[5] Science without Borders was a scientific mobility program created by the Brazilian government in 2011. The aim of the program was to promote the internationalization of Brazilian universities by sending students to an academic year in universities abroad and attracting researchers to Brazilian universities.

Table 3 – Numbers of texts and words by disciplinary group at BrAWE

	AH		SS		LS		PS		Total	
	Texts	**Tokens**	**Texts**	**Tokens**	**Texts**	**Tokens**	**Texts**	**Tokens**	**Texts**	**Tokens**
Case Study			5	15,326	9	20,908	18	41,866	32	78,100
Critique			7	15,341	16	25,782	19	36,053	42	77,176
Design							18	36,093	18	36,093
Empathy Writing										
Essay	4	7,887	13	20,906	46	82,975	31	48,401	94	160,169
Exercise			1	1,594	7	6,829	28	35,236	36	43,659
Explanation			7	11,371	11	14,976	29	54,266	47	80,613
Literature Survey					5	11,923	1	3,418	6	15,341
Methodology Recount					19	24,790	31	41,593	50	66,383
Narrative Recount					1	1,457	3	2,375	4	3,832
Problem Question			2	3,369	3	4,306	3	3,602	8	11,277
Proposal					2	4,078	12	17,554	14	21,632
Research Report					11	26,955	18	49,084	29	76,039
Total	4	7,887	35	67,907	130	224,979	211	369,541	380	670,314

3.1 Methodology

The lexical profile of Brazilian students' assignments was analysed using Range (NATION; HEATLEY, 2002). This software allows the researcher to compare the vocabulary of up to 32 text files. Nation and Heatley (2002) explain that "for each word in the texts, it (Range) provides a range or distribution figure, a headword frequency figure, a family frequency figure, and a frequency figure for each of the texts the word occurs in". However, in this study we compare the words in the *corpus* to pre-selected word lists. Thus, using AWL and GSL as baselists, the software has provided the percentage of academic words used in the texts written by Brazilian students.

Apart from that, Sketch Engine was used to check the frequency and the concordance lines of AWL words in BrAWE. In order to do that we used the AWL in Sketch Engine Whitelist tool. When a whitelist is added, the outcome of a word list shows only the frequency and the occurrences of the words in this whitelist. Furthermore, to compare how academic words are used in context in BAWE and BrAWE we selected the 40 highest frequency academic words in both *corpora* and analysed the concordance lines for the top three highest frequency words in each of them.

It is important to point out that our purpose is not to consider native speakers as baseline data, but, rather, excellent assignments, regardless of the writers' nationality. Thus, BAWE will be used as baseline data, since it is composed solely by highly graded assignments.

To verify if there is any significant difference in word frequencies in both *corpora*, the Log-likelihood (LL) test will be used. According to Rayson (2002), the

LL is appropriate to establish the statistical significance of the different frequencies of a word (or expression) between two *corpora* of different sizes. The author (2002, p. 58) lists a series of advantages for using the LL test in preference to other statistical tests, among them is the fact that "LL has been shown to be better 'in general' than the chi-squared test" and that "the Mann-Whitney test is suitable only for mid to high frequency words and for comparing *corpora* of the same size" (p. 58). If the LL outcome is 6.43 or higher, it means that there is a 99% chance that the difference between the two *corpora* is not random. When the result of this test is negative (-), this means that the first *corpus* includes proportionally a lower frequency of the word analysed than the second *corpus*.

4 Analysis

To visualize BrAWE lexical profile, Table 4 presents the coverage of the AWL, of the GSL and of off-list words divided per genre and area of study. In the first row the number between brackets represents the number of tokens in each *subcorpora*

Table 4 – AWL coverage in the *corpus* of Brazilian student

	AH (7,887)			SS (67,907)			LS (224,979)			PS (369,541)			Total (670,314)		
	GSL (%)	AWL (%)	OFF (%)	GSL (%)	AWL (%)	OFF (%)	GSL (%)	AWL (%)	OFF (%)	GSL (%)	AWL (%)	OFF (%)	GSL (%)	AWL (%)	OFF (%)
Case Study				78,60	9,70	11,70	74,10	8,90	17,00	79,00	10,40	10,50	77,60	9,90	12,50
Critique				79,70	11,00	9,40	77,00	11,00	12,00	77,20	11,10	11,80	77,60	11,00	11,40
Design										78,70	10,40	10,80	78,70	10,40	10,80
Empathy Writing															
Essay	81,70	5,20	13,10	80,60	9,40	10,00	69,50	9,60	20,90	77,30	10,50	11,30	73,90	9,60	16,50
Exercise				86,40	5,20	8,40	70,80	8,40	20,80	78,70	9,90	11,40	77,60	9,50	12,90
Explanation							73,00	8,00	19,10	78,60	8,90	12,50	77,90	8,90	13,20
Literature Survey							71,00	9,30	19,80	62,40	7,50	30,10	69,30	8,90	21,80
Methodology Recount							69,30	8,20	22,50	79,20	9,40	11,30	75,60	8,90	15,50
Narrative Recount							74,50	11,80	13,80	80,90	10,40	8,60	78,40	10,90	10,60
Problem Question							78,60	12,40	9,10	78,50	7,80	13,70	80,50	9,70	9,80
Proposal							76,30	11,30	12,40	77,80	12,40	9,80	77,40	12,20	10,30
Research Report							69,10	9,10	21,80	75,40	10,70	13,90	73,10	10,20	16,70
Total	81,70	5,20	13,10	80,30	9,80	9,90	71,30	9,40	19,30	77,90	10,20	12,00	76,00	9,80	14,20

The BrAWE *corpus* presents 76% of words covered by GSL, and 9.8% covered by AWL, whereas 14,2% are not covered by any list. On previous investigations, the AWL covered from 9 to 11% of academic written English *corpora*, with some small variations that could be attributed to discipline specificities (COBB; HORST, 2004; KONSTANTAKIS, 2007). Considering the different areas, the AWL covers around the same percentage of words in all BrAWE *subcorpora* except for Arts and Humanities, which is as low as 5,2%. However, this result should be taken as something to be further explored in the future, since AH contains only

7,887 words, thus not being representative enough to allow for generalizations. In addition, similar to previous studies, the Life Sciences (CHEN; GE, 2007; COBB; HORST, 2004; CHUNG; NATION, 2003) *subcorpus* has a slightly lower AWL coverage than the Physical Sciences (WARD, 2009; HYLAND; TSE, 2007), with the AWL covering 9.40% and 10.20% respectively. A salient difference between the lexical profile of LS and PS is the amount of off-list words, with 19,3% for LS and for 12% PS. PS would be closer to the 10% occurrence of off-list words predicted by Nation and Coxhead (2001) and LS a lot higher.

4.1 Life Sciences' lexical profile

Previous studies in Life Sciences *corpora* showed a somehow lower coverage of academic words: Chung and Nation (2003), 8.60%; Cobb and Horst (2004), 6.72% and Chen and Ge (2007), 10.07%. While in Chung and Nation (2003) this is due to the amount of off-list words, which is 71.1%, in Cobb and Horst (2007) the AWL and the GSL combined cover 81.57% which falls in the coverage proposed by Coxhead and Nation (2001). However, as in the *corpus* of Brazilian students almost 20% of the words in LS are neither in the GSL nor in the AWL, we decided to explore the off-list words in this *subcorpus*. The figure below represents the lexical profile of one text within this *subcorpus*. The words in red are off-list words, the words in yellow are from the AWL, and the blue and green words appear in the GSL.

Figure 1 – Excerpt from LS text

The high percentage of off-list words are discipline specific words (chloride, potassium, patients etc.). There are also some blended words, like, "kcqnumber", or "clcnumber". Still, this is a small excerpt from a *subcorpus* of 224,979 tokens. Looking at the list of the 50 most frequent off-list words in the LS *corpus* (below), it is possible to assume that they are mainly related to health issues. These findings suggest that technical vocabulary plays an important role in academic writing for disciplines in LS.

Table 5 – Top 50 off-list words in Life Sciences

Word	Raw frequency	Word	Raw frequency
AL	887	BRAZIL	81
ET	875	CLIMATE	81
PATIENTS	355	MALARIA	74
CELLS	291	CALCIUM	71
DNA	230	REFERENCE	69
CANCER	225	HCV	68
CELL	195	PCR	67
SPECIES	170	MEMBRANE	65
ML	166	MUTATIONS	65
GENES	152	TISSUES	65
DRUG	126	LIVER	64
PROTEIN	126	MUSCLE	64
DRUGS	124	X	64
BACTERIA	118	DIET	61
PROTEINS	112	CARBON	59
GENE	106	MBC	59
CLINICAL	102	VITAMIN	56
INFECTION	100	MOLECULES	55
TISSUE	99	MARINE	54
CHRONIC	98	ORGANISMS	54
SYMPTOMS	93	VIRUSES	54
DIAGNOSIS	87	DIABETES	53
BIOLOGICAL	84	PATHWAY	53
GENETIC	84	CORRELATION	52
ACTIVATION	81	THERAPY	52

4.2 Physical Sciences' lexical profile

Turning to the AWL coverage in the PS *subcorpus*, it is possible to compare it to the *subcorpus* of Engineering in Hyland and Tse (2007) (11.10%) and Ward (2009) (11.30%). The AWL coverage in BrAWE-PS of 10.20% is slightly lower than the figures found on previous *corpus* studies. Besides, the use of off-list words in this *corpus* is lower than in LS, only 12%. Figure 2 below presents the lexical profile of one text in the PS *subcorpus*.

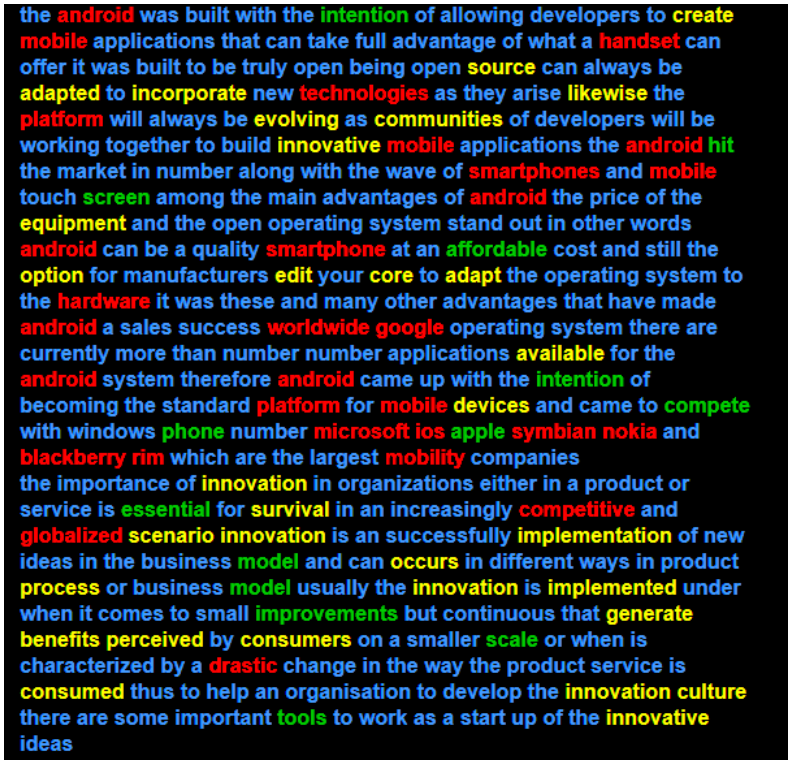


Figure 2 – Excerpt from PS *subcorpus*

The 50 most frequent off-list words in this *subcorpus* (Table 6, below) are mainly related to discipline specific terms. Nevertheless, when comparing the frequency of off-list words in PS and LS, it seems that knowing these words in LS leads to a larger impact in writing than in PS, even though the *corpus* of PS is larger, the mean frequency of the top off-list words in LS is two-thirds higher than the frequency in PS.

Table 6 – Top 50 off-list words in Physical Sciences

Word	Raw frequency	Word	Raw frequency
AL	388	ORGANIC	83
ET	386	RENEWABLE	83
TURBINE	216	TORQUE	82
URBAN	171	LONDON	81
MM	147	PEAK	80
CARBON	138	DIAGRAM	79
SOFTWARE	126	HUGE	77
SOLAR	126	GENERATOR	76
GRAPH	124	COMPETITIVE	75
LASER	111	DIAMETER	75
CIRCUIT	110	COLUMN	74
BRAZIL	107	DENSITY	74
VOLTAGE	107	AXIS	73
SPECIES	106	MATLAB	72
X	103	CLIMATE	71
CONCRETE	101	PITCH	71
FIG	96	TECHNOLOGIES	71
FLUID	95	CO2	70
STORAGE	95	ROTOR	69
ARCHITECTURE	92	TRAFFIC	69
UNILEVER	92	PH	68
UK	91	VELOCITY	68
PLOT	87	CELLS	67
BIOMASS	86	LINEAR	67
COEFFICIENT	84	HEIGHT	65

5 The use of academic vocabulary: BAWE compared to BrAWE

In order to verify the words in the Academic Word List (AWL) used by Brazilian students and the differences between BAWE and BrAWE, we generated a word list using Sketch Engine whitelist tool for the PS and LS sections of both *corpora*.

The AWL has 570 word families, which are translated into 3,120 word forms including different spellings of the same word, such as "labor" and "labour" or "analyse" and "analyze". Taking into account the complete list, texts in the BAWE *subcorpus* used 999 word forms, and BrAWE used 998 word forms. Table 5 presents the 40 most frequent AWL words in both *corpora*.

The table below depicts the frequency of 54 words, which are the 40 most frequent words in BAWE, from which, 26 are also from the top 40 in BrAWE. The bottom 14 words are part of the top 40 in BrAWE, but not in BAWE. The first column represents a word position in BAWE word list, the second is the word itself, and the third column depicts its frequency. The fourth column presents

the word position in BrAWE, the fifth is the word, and the sixth column shows its frequency. Finally, the last column presents the log-likelihood results showing how statistically significant the frequency difference is (not).

Out of the 54 words shown in the table, twenty words are used more frequently in BAWE, being 14 of them statistically significant (LL 6.43 or higher). These words are marked in italics. Conversely, 34 present a higher frequency of use in BrAWE, these are the rows with negative LL results, being 25 of these differences statistically significant (LL 6.43 or higher). This high amount of negative LL results might suggest a wider range of academic vocabulary in BAWE as compared to BrAWE, since Brazilian students presented lower lexical variation. Furthermore, although the AWL does not account for parts of speech, it is possible to see that the most frequent academic words are either nouns or verbs, followed by adjectives.

5.2 Summary of key findings

BrAWE is composed of 76% of high frequency words (GSL), 9.8% of Academic words (AWL) and 14.2% of off-list words. Thus, it is possible to assume that Brazilian students' written assignments do not differ in the use of the academic words from other academic *corpora* presented in the literature review. An important aspect to be taken into consideration is the high use of off-list words when compared to previous studies, especially in the LS *subcorpus*, since the AWL and the GSL combined usually cover around 85% to 90% of an academic *corpus*. As already mentioned, off-list words are usually discipline specific words. This extended use of discipline specific words in both *corpora* corroborates Hyland and Tse (2007) and Durrant's (2014) arguments that discipline specific terms are as important as academic words in academic English.

The aim of this research was to investigate the use of academic vocabulary by Brazilian students. More specifically, it has explored the AWL coverage in a *corpus* of Brazilian students and compared how Brazilian students use the words in the AWL to how students represented in the BAWE *corpus* use these words.

The analysis of the AWL coverage showed that Brazilian students use academic words to the same extent as observed in other academic *corpora* presented in the literature review. It also revealed a significant difference between the lexical profile of Life Sciences and Physical Sciences *subcorpora*, especially if we take off-list words into account. Life Sciences texts relied more on discipline specific words, corroborating Hyland and Tse's (2007) view that the division between academic and discipline specific vocabulary is not clear. Table 7 below shows the 40 most frequent academic words in both BAWE and BrAWE. The first column represents the word's position in BAWE or BrAWE, the second column shows the word form, and the third column presents the frequency. The shaded words at the end of the

table represent the words that are among the 40 most frequent academic words in BrAWE, but which are not among the 40 most frequent academic words in BAWE.

Table 7 – Most frequent academic words

BAWE			BrAWE			LL
Position	Word	Frequency	Position	Word	Frequency	
1	data	1829	3	data	707	-0,44
2	process	1529	1	process	792	-52,18
3	found	1235	5	found	538	-8,16
4	project	1054	4	project	628	-80,27
5	design	1047	8	Design	449	-5,49
6	method	967	9	Method	340	1,08
7	analysis	931	6	analysis	523	-52,18
8	required	880	29	required	208	39,39
9	research	857	14	research	266	7,56
10	function	853	12	function	286	2,77
11	structure	816	28	structure	213	23,84
12	energy	776	2	energy	709	-280,53
13	area	748	7	Area	493	-89,64
14	significant	734	30	significant	206	14,31
15	evidence	636	127	evidence	101	79,42
16	factors	631	10	Factors	328	-22,05
17	available	617	37	available	197	4,01
18	ratio	609	67	Ratio	151	22,52
19	period	607	80	Period	132	36,1
20	environment	602	16	environment	254	-2,41
21	range	600	20	Range	244	-1,11
22	potential	599	36	potential	198	2,45
23	methods	592	21	methods	238	-0,8
24	similar	591	41	similar	190	3,53
25	team	567	113	team	109	47,28
26	issues	563	56	issues	163	8,91
27	areas	550	17	areas	251	-6,43
28	individual	503	71	individual	144	8,61
29	normal	498	78	normal	133	12,94
30	section	492	68	section	150	5,14
31	output	484	25	output	229	-8,1
32	role	464	90	role	127	10,53
33	stress	463	15	stress	260	-25,9
34	specific	462	11	specific	293	-47
35	strategy	451	109	strategy	112	16,56
36	response	444	58	response	161	74,26
37	approach	439	33	approach	201	-5,31
38	processes	438	24	processes	230	-16,37
39	physical	435	53	physical	165	-0,01
40	reaction	426	42	reaction	190	-3.83
70	concentration	340	13	concentration	281	-91.3
73	environmental	331	18	environmental	249	-65.3
67	site	354	19	site	248	-53.99
54	obtained	396	22	obtained	235	-29.57
83	equation	310	23	equation	231	-59.17
74	region	328	26	Region	225	-46.14
60	maximum	377	27	maximum	220	-25.87
50	final	401	31	Final	206	-12.93
58	complex	382	32	complex	206	-16.89
66	negative	359	34	negative	200	-19.21
64	technology	365	35	technology	199	-17.24
62	positive	374	38	positive	192	-12
79	achieved	316	39	achieve	192	-26.44
228	construction	146	40	construction	191	-128.01

The investigation of the AWL words selected by Brazilian students and students in BAWE suggested that fixed expressions influence both the underuse and overuse of academic words by Brazilian students. In some cases, these students transfer expressions from Portuguese into English or do not make use of common expression in their field of study. Furthermore, the comparative analysis of the top frequency AWL words between the LS and PS *subcorpora* suggested that these two areas of study use these academic words in distinct ways.

5.3 Research evaluation and limitations

The main limitation of this study is the size and representativeness of the BrAWE *corpus*. Although statistical significance tests were possible in this investigation, the *corpus* was not balanced. Therefore, the disciplines of Physical Sciences and Life Sciences were overrepresented. Furthermore, the *corpus* compiled for this study could be expanded to contain more assignments. However, due to time constraints, this was not possible.

Investigation of the word forms in all sublists could have exhibited other language aspects that Brazilian students use differently than students in BAWE. In addition, the first part of the research could have been sounder if the AWL coverage of BAWE was verified as well, since the comparison of the AWL coverage would, then, be between two similar *corpora*.

Finally, we consider the use of the AWL as one of the limitations of this study. Although this academic word list is the most widely used in English for Academic Purposes' research, we believe that the results of the last section might have been more instructive if the word list used as a reference was based on lemmas, rather than on word forms. A word list based on lemmas could provide some insights related to the use of academic words in each part of speech, which is not possible with the AWL.

5.4 Pedagogical implications

The results point that BrAWE presents a similar number of academic words as other academic *corpora*. However, it was possible to notice some peculiarities in the use of academic vocabulary by Brazilian students. Considering these specific characteristics, some pedagogical recommendations for the teaching of English for Academic Purposes (EAP) for Brazilian students are presented below.

As a first point, the results indicate that there is a considerable difference in the lexical profile of texts in the fields of Life Science and Physical Sciences. This difference is reflected not only on the percentage of the use of academic words, but also on how these words are used in these fields of study. These findings suggest

that LS and PS students would benefit more from their EAP classes if they took lessons in separate groups as argued by Hyland and Tsé (2007).

In addition, in classes with students from different disciplines, teachers could group students from the same field of study to work on specific pedagogical tasks. This way, classroom practices could be more focused on the academic vocabulary of those specific disciplines. Furthermore, instructors could ask students to bring examples of texts of their field of study to the classroom and carry out some *corpus* analysis so that students could identify different lexical choices in these texts. The main point is that language instructors need to be aware that there is a considerable difference in the use of academic vocabulary between LS and PS and this should be taken into consideration when preparing their lessons.

Another further topic should be addressed in EAP classes for Brazilian students: the use of discipline specific fixed expressions. This could be done by having students explore the lexical bundles in BAWE for the specific disciplines, or by creating pedagogical materials based on Hyland (2008) or Simpson-Vlach and Ellis's (2010) lists of fixed expressions recurrent in academic texts.

The final pedagogical contribution is related to the *corpus* compiled for this study. This *corpus* can be used by instructors to provide students with examples of Brazilian academic writing which would probably look more familiar than texts produced by native speakers, for instance.

5.5 Suggestions for further research

In addition to the pedagogical implications presented above, we would like to suggest some ideas for future research to foster Brazilian EAP expertise.

The first one would be to compare the coverage of academic vocabulary in BAWE and the *corpus* of Brazilian students. This would allow the researcher to draw more informed conclusions related to the coverage of academic vocabulary in Brazilian students' writings.

A second aspect to be explored in the future is the differences in the use of Academic Vocabulary across genres. From table 4 presented in the results section, we could notice that some genres, such as Proposals and Critique, present higher AWL coverage than others, such as Literature Survey, Explanation and Methodology Recount. However, the study presented in this chapter did not explore genre differences in the use of the AWL.

In addition, another topic that could be addressed is the lexical bundles in Brazilian students' academic writing. Many researchers (SIMPSON-VLACH; ELLIS, 2010; HYLAND, 2008; BIBER; CONRAD; CORTES, 2004) claim that lexical bundles are the fabric of language and some of them have explored the use of lexical bundles in academic texts. Thus, it might be interesting to investigate

the similarities and differences in the use of these units of language by Brazilian students.

References

BECK, I. et al. *Bringing words to life*: Robust vocabulary instruction. New York: Guilford, 2002.

BIBER, D.; CONRAD, S.; CORTES, V. If you look at Lexical bundles in university teaching and textbooks. *Applied Linguistics*, v. 25, n. 3, p. 371-405, 2004.

CHEN, Q.; GE, C. A *corpus*-based lexical study on frequency and distribution of

Coxhead's AWL word families in medical research articles. *English for Specific*

Purposes, v. 26, p. 502-514, 2007.

CHUNG, T.; NATION, P. Technical vocabulary in specialised texts. *Reading in a foreign language*, v. 15, n. 2, p. 103-116, 2003.

CLARK, M.; ISHIDA, S. Vocabulary knowledge differences between placed and promoted EAP students. *Journal of English for academic purposes*, v. 4, p. 225-238, 2005.

COBB, T.; HORST, M. Is there room for an AWL in French? In: BOGAARDS, P.; LAUFER, B. (Ed.). *Vocabulary in a second language*: Selection, acquisition, and testing. Amsterdam: John Benjamins, 2004, p. 15-38.

COXHEAD, A. A new academic word list. *TESOL Quarterly*, v. 34, n. 2, p. 213-238, 2000.

COXHEAD, A. The Academic Word List 10 years on: research and teaching implications. *TESOL Quarterly,* v. 4, n. 2, p. 355-362, 2011.

COXHEAD, A.; NATION, P. The specialized vocabulary of English for academic purposes. In: FLOWERDEW, J.; PEACOCK, M. (Ed.). *Research perspectives on English for academic purposes*. Cambridge: Cambridge University Press, 2001, p. 252-267.

DURRANT, P. Discipline and Level Specificity in University Students' Written Vocabulary. *Applied Linguistics*, v. 35. n. 3, p. 328-356, 2014.

FISHER, D.; FREY, N. *Word wise and content rich: five essential steps to teaching academic vocabulary.* Portsmouth: Heinemann, 2008.

GARDNER, D.; DAVIES, M. A new academic vocabulary list. *Applied Linguistics*, v. 35, n. 3, p. 305-327, 2014.

HARMON, J. M.; WOOD, K. D.; HEDRICK, W. B. Vocabulary instruction in middle and secondary content classrooms: Understandings and direction from research. In: FARSTRUP, A. E.; SAMUELS, S. J. (Ed.). *What research has to say about vocabulary instruction*. Newark: International Reading Association, 2008, p. 150-181.

HIRSH, D.; NATION, P. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, v. 8, n. 2, p. 689-696, 1992.

HSUE-CHO, M.; NATION, P. Unknown vocabulary density and reading comprehension. *Reading in a foreign language*, v. 13, n. 1, p. 403-430, 2000.

HYLAND, K.; TSE, P. Is there "an academic vocabulary"? *TESOL Quarterly*, v. 41, n. 2, p. 235-253, 2007.

HYLAND, K. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, v. 27, p. 4-21, 2008.

KAUR, J.; HEGELHEIMER, V. ESL students' use of concordance in the transfer of academic word knowledge: An exploratory study. *Computer Assisted Language Learning*, v. 18, n. 4, p. 287-310, 2007.

KONSTANTAKIS, N. Creating a business word list for teaching business English. *Estudios de Linguistica Inglesa Aplicada*, v. 7, p. 79-102, 2007.

LAUFER, B. What percentage of text-lexis is essential for comprehension? In: LAUREN, C.; NORDMAN, M. (Ed.). *Special Language*: from human thinking to thinking machines. Clevedon: Multilingual Matters Ltd., 1989, p. 316-323.

LEI, L.; LIU, D. A new medical academic word list: a *corpus* based study with enhanced methodology. *Journal of English for academic purposes*, v. 22, p. 42-53, 2016.

MATTE, M. L. *A* corpus-*based study of connectors in student academic writing*. Final Paper. Universidade Federal do Rio Grande do Sul, 2017.

NATION, P. *Learning vocabulary in another language.* Cambridge: Cambridge University Press, 2013.

NATION, P.; HEATLEY, A. *Range*: A program for the analysis of vocabulary in texts. 2002. Retrieved from: <http://www.victoria.ac.nz/lals/resources/range>.

SCHMITT, N.; SCHMITT, D.; CLAPHAM, C. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, v. 18, n. 1, p. 55-88, 2001.

SILVA, L. G. Compilation of a Brazilian Written English *Corpus*. *E-scrita*, v. 8, n. 2, p. 32-47, 2017.

SIMPSON-VLACH, R.; ELLIS, N. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, v. 31, n. 4, p. 487-512, 2010.

WARD, J. A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, v. 28, p. 170-182, 2009.

WEST, M. *A general service list of English words.* London: Longman, 1953.

Atividades de compreensão oral com base em *corpora* de *TED Talks*: um estudo piloto

Listening comprehension activities based on *corpora* of TED Talks: a pilot study

Luciano Franco da Silva
Paula Tavares Pinto
Elen Dias

Resumo: Este artigo apresenta alguns exemplos sobre como o uso de *corpora* orais podem ser adotados para a criação de atividades de compreensão oral para alunos em nível A2 e B1 em língua inglesa. Para tanto, a Linguística de *Corpus* foi utilizada como embasamento teórico-metodológico nas análises e descrições lexicais. Os *corpora* de estudo foram compilados a partir das transcrições de 71 palestras retiradas do *site* TED (www.ted.com) e de 56 transcrições das animações do *site* TED-ED (https://ed.ted.com/). Com base nessas palestras, atividades didáticas foram criadas e, posteriormente, aplicadas em um minicurso ministrado em uma faculdade de tecnologia no noroeste do estado de São Paulo.

Palavras-chave: Ensino e aprendizagem de língua estrangeira. Inglês com fins acadêmicos. Linguística de *Corpus*. Atividades didáticas de compreensão oral em língua inglesa.

Luciano Franco da Silva – Mestre em Estudos Linguísticos pela Universidade Estadual "Júlio de Mesquita Filho" (Unesp/Ibilce), bolsista da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – luciano.francco@gmail.com.
Paula Tavares Pinto – Professora da Universidade Estadual "Júlio de Mesquita Filho", doutora pela Universidade Estadual "Júlio de Mesquita Filho" (Unesp/Ibilce) – paula@ibilce.unesp.br.
Elen Dias – Professora da Faculdade de Tecnologia de São Paulo – FATEC Jales, doutora pela Universidade Estadual "Júlio de Mesquita Filho" (Unesp/Ibilce) – elen.dias@fatec.sp.gov.br.

Abstract: This paper presents some examples on how oral *corpora* can be used to create listening comprehension activities for students at levels A2 and B1 of English. To that end, *Corpus* Linguistics was adopted as a theoretical-methodological basis for lexical analyzes and descriptions. The *corpora* of study were compiled from transcripts of 71 lectures taken from the TED website (www.ted.com) and 56 transcripts from the TED-ED website animations (https://ed.ted.com/). Based on those talks, didactic activities were created and applied in a mini-course taught at a Technology College in the Northwest of the state of São Paulo.

Keywords: Foreign Language Teaching and Learning. English for Academic Purposes. *Corpus* Linguistics. Didactic activities for listening comprehension in English.

1 Introdução

A preocupação com o desenvolvimento da compreensão oral (CO) no ensino de línguas estrangeiras (LEs) foi, até o final de século XX, deixada em segundo plano, uma vez que era entendida por muitos como uma habilidade passiva que o aluno desenvolvia conforme aprendia a ler e a escrever na segunda língua (PAIVA, 2014; ROST, 2011; SATO, 2011). Essa concepção ainda persiste em muitos contextos de ensino até hoje.

Brown (2001) e Paiva (2014) lembram que um dos possíveis motivos para esse "atraso" em se reconhecer a importância da CO no ensino de línguas provém da ideia de que a produção, em especial a oral, seja o maior indicador de proficiência que um falante de segunda língua possa demonstrar. Isso justifica o fato de que, no decorrer da história do ensino de línguas, as habilidades de produção (*Speaking* e *Writing*) sempre receberam maior prestígio do que as habilidades de compreensão (*Reading* e *Listening*).

Visando a expor o aluno a novas formas de contato com a língua, a Linguística de *Corpus* (LC) apresenta-se como uma excelente metodologia para o auxílio no ensino de LEs (O'KEEFFE; MCCARTHY; CARTER, 2007; CHENG, 2010; McENERY; XIAO, 2011; CHARLES, 2012), em especial no que se refere ao uso de *corpora* orais (LUZÓN et al., 2007; MAURANEN, 2007; ADOLPHS; KNIGHT, 2010), que vêm ganhando, cada vez mais, espaço nas pesquisas científicas graças aos avanços tecnológicos.

Nessa linha de pensamento, este artigo basear-se-á nas relações entre os dados obtidos da compilação de *corpora* provenientes dos *sites* TED e TED-ED e da preparação de materiais didáticos com o intuito de melhorar a CO de alunos que participaram de um minicurso de Inglês para Fins Acadêmicos (IFA), oferecido em uma instituição de ensino superior no interior do estado da São Paulo.

O presente artigo visa a apresentar os principais conceitos acerca do uso de *corpora* orais para o ensino de LE (LUZÓN et al., 2007; MAURANEN, 2007; MCENERY; XIAO, 2011; GOH, 2012). Além disso, pretende-se observar e

descrever *se* e *como* o uso de *corpora* poderá auxiliar no desenvolvimento de atividades didáticas de CO em língua inglesa, a fim de contribuir com a melhora do processo de ensino-aprendizagem.

2 Fundamentação Teórica

Nesta seção, são examinados os pressupostos teóricos que norteiam este artigo. Será apresentado o referencial teórico sobre a Linguística de *Corpus* e como essa metodologia de análise linguística influencia o ensino de línguas.

2.1 Linguística de Corpus do Ensino de Línguas

Para estabelecermos algumas considerações a respeito da LC, tanto para a análise linguística quanto para o ensino de línguas, é necessário, primeiramente, entender o sentido do termo *corpus* dentro da grande área da Linguística.

Em termos mais abrangentes, um *corpus*, segundo Tagnin (2011), é uma coletânea de textos[1] compilados em formato eletrônico a partir de critérios específicos para servirem como objeto de estudos linguísticos. Todavia, é preciso ter em mente que nem todo texto pode ser considerado um *corpus*, visto que existem diversos critérios que devem ser respeitados em sua compilação (BERBER SARDINHA, 2004; SINCLAIR, 2005). Dessa forma, neste trabalho, utilizaremos as noções de *corpus* com base nas leituras de Granger (2002) e Sinclair (2005), cujas definições, que serão apresentadas a seguir, complementam-se e delimitam bem o conceito de um *corpus*.

Primeiramente, Sinclair (2005) define um *corpus* como:

> Uma coleção de porções de linguagem em formato de textos eletrônicos, selecionados de acordo com critérios externos para representar, o máximo possível, uma língua ou variedades de uma língua como fonte de dados para a pesquisa linguística[2] (SINCLAIR, 2005).

O conceito de *corpus* adotado por Granger (2002) complementa a definição de Sinclair (2005):

[1] Para a definição de *textos*, entendemos tanto aqueles pertencentes ao uso oral da língua quanto ao escrito.

[2] "*A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as source of data for linguistic research*". (SINCLAIR, 2005)

> Linguística de *Corpus* pode ser melhor definida como uma metodologia linguística que é baseada no uso de coleções eletrônicas de textos ocorridos naturalmente. Ela não é um novo ramo da linguística, nem uma nova teoria da linguagem, mas a própria natureza das evidências que ela utiliza a tornam uma metodologia particularmente poderosa, que possui o potencial de mudar a perspectiva sobre a qual vemos a língua[3]. (GRANGER, 2002, p. 3)

Vale lembrar que essas porções de linguagem devem provir de textos autênticos, ou seja, aqueles que não foram produzidos com o propósito de integrarem o *corpus* (BERBER SARDINHA, 2004; O'KEEFFE; McCARTHY; CARTER, 2007). Em relação aos critérios para a escolha dos textos, Sinclair (2005) e Berber Sardinha (2004) elencam alguns dos mais comuns utilizados nas pesquisas. São eles: *i*) modo (oral ou escrito); *ii*) composição (jornais, revistas, palestras, artigos científicos); *iii*) origem e data; e *iv*) extensão e representatividade.

Acrescentando ao que foi mencionado acima, podemos afirmar que a LC é uma metodologia de cunho empirista (CHENG, 2010; McENERY; XIAO, 2011), que analisa a língua como um sistema probabilístico, a partir da exploração sistemática de um *corpus*. Pelo fato de essas pesquisas visarem a descrições da língua em uso, muitos pesquisadores também utilizam os dados obtidos dos *corpora* para uma análise crítica em currículos e materiais de ensino de línguas estrangeiras (O'KEEFFE; McCARTHY; CARTER, 2007; CHENG, 2010; McENERY; XIAO, 2011; BERBER SARDINHA, 2015).

Além disso, Hunston (2002) e Berber Sardinha (2009) destacam a importância dos *softwares* de análises linguísticas, afirmando que eles oferecem uma nova perspectiva ao pesquisador por meio de listas de frequências, fraseologias e colocações, revelando, dessa forma, aspectos da língua que seriam despercebidos em outros métodos de análises.

Graças aos avanços tecnológicos nas últimas décadas, *corpora* provenientes da modalidade oral têm ganhado espaço nas pesquisas linguísticas (LUZÓN et al., 2007; MAURANEN, 2007; ADOLPHS; KNIGHT, 2010). É interessante notar que, na grande maioria dos casos, *corpora* orais tendem a ser muito menores em tamanho e, consequentemente, apresentam um número menor de ocorrências quando comparados a suas contrapartes escritas (O'KEEFFE; McCARTHY; CARTER, 2007; ADOLPHS; KNIGHT, 2010).

No entanto, diversas pesquisas indicam que eles são excelentes recursos para variados tipos de pesquisas com interesse no uso natural da língua (ADOLPHS; KNIGHT, 2010; CHENG, 2010, 2007; LUZÓN et al., 2007; MAURANEN,

[3] "Corpus *linguistics can be best defined as a linguistic methodology which is founded on the use of electronic collections of naturally occurring texts. It is neither a new branch of linguistics nor a new theory of language, but the very nature of the evidence it uses makes it a particularly powerful methodology, one which has the potential to change perspectives on language".* (GRANGER, 2002, p. 3)

2007), tais como, entonação, linguagem corporal e estrutura do discurso, visto que esses são aspectos do uso da língua difíceis de serem analisados em *corpora* provenientes de textos escritos (ADOLPHS; KNIGHT, 2010).

Dentre as diversas utilizações de *corpora* no ensino, Charles (2012) destaca três deles, que são: *i) repetição de padrões linguísticos* dentro de uma quantidade massiva de dados, revelando, assim, evidências objetivas e não mais subjetivas na análise linguística; *ii) descrição de características específicas do discurso acadêmico*, assim como a fraseologia de diferentes disciplinas e gêneros acadêmicos; por último, *iii) compilação de listas de palavras acadêmicas*[4], tais como o trabalho de Coxhead (2000).

Complementando os usos de *corpora* mencionados acima, Flowerdew (2001) apresenta o uso de *corpora* pequenos na produção de materiais voltados ao IFA, visto que, para o contexto acadêmico, eles são mais especializados em termos de tópico e gênero. Além disso, pesquisas com base em *corpus* podem oferecer recursos mais realistas, ilustrativos e atuais para a criação de materiais de ensino (CHENG, 2010), ideia que também é defendida por McCarthy (2001 apud O'KEEFFE; McCARTHY; CARTER, 2007, p. 21) ao afirmar que:

> A Linguística de *Corpus* representa uma mudança pioneira nas técnicas e métodos científicos, e provavelmente, antecipa mudanças tecnológicas que mudarão nossa noção de educação, papel do professor e contexto cultural dos serviços de educação e a mediação de teorias e técnicas[5].

Conforme mencionado anteriormente, a LC tem uma natureza empírica com o foco na língua em contextos reais de comunicação; sendo assim, sua principal contribuição para o ensino de LE é o de demonstrar, por meio de textos autênticos, características e padrões que não são percebidos somente pela intuição do pesquisador.

Nessa mesma linha de pensamento, O'Keeffe, McCarthy e Carter (2007) e Walsh (2010) apresentam uma contradição em relação à preparação de materiais didáticos de LE, afirmando que, em sua maioria, eles são baseados na intuição de como a língua é utilizada, ao invés de apresentar evidências reais de seu uso. Como solução para esse problema, os autores supramencionados defendem a ideia de que materiais com base em *corpora* devem colocar o aluno em contato direto com a língua autêntica.

---

[4] Mais informações estão disponíveis no *site*: http://www.victoria.ac.nz/lals/resources/academicwordlist/

[5] No original: "Corpus *Linguistics represents cutting-edge change in terms of scientific techniques and methods and probably foreshadows even more profound technological shifts that will "impinge upon our long-held notions of education, roles of teachers, the cultural context of the delivery of educational services and the mediation of theory and technique".* (MCCARTHY, 2001 *apud* O'KEEFFE; McCARTHY; CARTER, 2007, p. 21)

2.2 As TED Talks e sua aplicação em contextos de ensino

As *TED Talks*, acrônimo de *Technology, Entertainment and Design*, são séries de conferências sem fins lucrativos iniciadas em 1984 e, hoje, realizadas em diversos locais do mundo com o intuito de disseminar a produção científica e intelectual. Desde 2007, as conferências são disponibilizadas *on-line* e legendadas para mais de 80 línguas, e, atualmente, muitas palestras tornam-se virais na internet, alcançando milhões de visualizações.

Além de sua relevância sociocultural, as palestras do *TED Talks* têm ganhado espaço no contexto acadêmico, em especial no ensino de inglês como língua estrangeira (TAKAESU, 2013). Segundo o autor, as *TED Talks* são uma excelente fonte de insumo para o desenvolvimento da CO.

Nesse sentido, o uso deste recurso é de alta relevância para o ensino de línguas, pois oferece ao aluno oportunidade de acesso ao uso autêntico da língua. Takaesu (2013 *apud* FIELD, 2002, p. 244) reforça o uso das *TED Talks* em atividades de CO, afirmando que as palestras colocam os alunos em contato com exemplos de hesitações, falsos começos, reduções de palavras e diferentes sotaques, características autênticas da linguagem oral e muito pouco abordadas nos materiais de ensino de LE.

2.3 Considerações sobre a compreensão de palestras em LE

Considerando a habilidade de CO uma competência comunicativa essencial para pessoas inseridas em contextos acadêmicos (CHAUDRON; LOSCHKY; COOK, 1994; FLOWERDEW, 1994), é importante se considerar as características pertinentes ao discurso acadêmico (RICHARDS, 1983; FLOWERDEW, 1994), em especial sobre o processo da compreensão de palestras, visto que é de grande valor para o professor de LE. A esse respeito, Flowerdew (2004, p. 8) afirma que sua importância se justifica por

> poder sugerir maneiras apropriadas de encorajar alunos de LE a ouvir palestras. Por um lado, melhora a metodologia de ensino de línguas e por outro aprimora as estratégias de ensino. Além disso, informações sobre este processo podem guiar palestrantes em como apresentarem seus conteúdos, garantindo assim, uma melhora na compreensão do conteúdo[6].

[6] No original: "[...] *can suggest appropriate ways to encourage second language learners to listen to lectures. It can thus feed into ESL teaching methodology, on the one hand, and leaner strategy training, on the other. In addition, information about the lecture comprehension process can guide content lecturers in how to present their lectures to ensure optimal comprehension".* (FLOWERDEW, 2004, p. 8)

Richards (1983) elenca as seguintes habilidades de CO necessárias para contextos acadêmicos, em especial para a compreensão de palestras.

1. Habilidade em identificar o propósito e o escopo da palestra.
2. Habilidade em identificar o tópico da palestra, e acompanhar o seu desenvolvimento.
3. Habilidade para identificar a relação entre as unidades do discurso (e.g. ideias principais, generalizações, hipóteses, argumentos e exemplos).
4. Habilidade em identificar o papel dos marcadores de discurso dentro da estrutura da palestra (e.g. conjunções e advérbios).
5. Habilidade em inferir relações (e.g. causa, efeito, conclusão).
6. Habilidade em reconhecer itens lexicais chaves relacionados ao assunto/tópico.
7. Habilidade em deduzir palavras de acordo com o contexto.
8. Habilidade em reconhecer marcadores de coesão.
9. Habilidade em reconhecer diferentes entonações para obter informações (e.g. tom de voz, volume e ritmo).
10. Habilidade em detectar a postura do falante em relação ao assunto da palestra.
11. Habilidade em seguir diferentes tipos de palestras: falada, gravada, áudio visual).
12. Habilidade de acompanhar a palestra independente do sotaque e da velocidade de fala do palestrante.
13. Familiaridade com os diferentes estilos de palestras: formal, conversa, leitura, informal.
14. Familiaridade com os diferentes tipos de registro: escrito *versus* coloquial.
15. Habilidade em reconhecer informações irrelevantes: piadas, digressões.
16. Habilidade em reconhecer informações não verbais como marcadores de ênfase e postura.
17. Conhecimento dos acordos em sala de aula (e.g. turnos de perguntas, pedidos).
18. Habilidade em reconhecer incumbências deixadas aos ouvintes (e.g. avisos, sugestões, recomendações, instruções e conselhos). (RICHARDS, 1983, p. 229-230)

Flowerdew (1994), ao comparar as habilidades de CO gerais e acadêmicas, afirma que elas se diferenciam em graus e tipos. Habilidades de conhecimento prévio e separação de informações mais relevantes são diferenciadas por graus, pois também são necessárias para a CO geral, porém em menor grau de importância. Já as habilidades de concentração em longos períodos de diálogos, de fazer anotações e de interação com as diferentes fontes de informações são tipos de habilidades mais associadas ao contexto acadêmico (FLOWERDEW, 1994).

3 Metodologia

A presente seção apresentará os materiais e os passos metodológicos utilizados nesta pesquisa. Serão feitas as considerações necessárias acerca das ferramentas

de análise lexical utilizadas, como o *software* AntConc® e alguns de seus recursos (ferramentas Word List, Keyword List e N-Gram). Assim como sobre a compilação do *corpus* de estudo, englobando os critérios de seleção, os passos para a sua coleta. Após, discutiremos os critérios obrigatórios e opcionais que envolvem a criação de atividades com base em *corpora*.

Ao final desta seção relataremos os passos metodológicos para a criação das atividades, elencando a taxonomia dos exercícios e a separação das atividades em *pre-listening*, *listening* e *post-listening*.

3.1 Detalhamento do corpus de estudo e ferramentas utilizadas

O *corpus* de estudo para a pesquisa foi compilado a partir das transcrições de 71 palestras retiradas do *site* TED (www.ted.com) e de 56 transcrições das animações do *site* TED-ED (https://ed.ted.com/). Esse *corpus* totalizou 152.261 itens (*tokens*) e 18.012 formas (*types*). Já as apresentações do TED-ED compilaram um *subcorpus* de 40.156 itens e 9.304 formas.

Em relação aos critérios para a seleção das palestras, buscou-se compilar um conjunto de dados que abordasse diversos assuntos e representasse as diferentes variedades da língua inglesa. Portanto, os principais fundamentos para a seleção dos textos eram ser pertencentes à modalidade oral e ao gênero textual *palestra acadêmica*.

O segundo passo foi salvar os 127 *scripts*, disponibilizados publicamente na internet, em uma pasta eletrônica; após, eles foram salvos em extensão *.txt, o que tornou os arquivos passíveis de análise pelo *software* AntConc® e, desta forma, foi possível acessar e avaliar todos os dados contidos no *corpus.*

O *software* para análise linguística AntConc® foi criado por Laurence Anthony (2002) da Universidade de Waseda, no Japão. Trata-se de um concordanciador usado para criar listas de ocorrências de uma palavra ou linhas de concordância[7] em uma quantidade definida de textos. Esse concordanciador pode ser adquirido sem custos via *download* [8].

O AntConc® é formado por um conjunto de ferramentas que permitem a realização de buscas e cálculos estatísticos de ocorrências de palavras em um *corpus*, como podemos perceber na Figura 1 abaixo:

[7] Utilizaremos nesta pesquisa a definição de Tagnin (2011) que define linhas de concordância como a relação de todas as ocorrências de uma palavra de busca em um *corpus* junto com seu cotexto. Em geral, é apresentada em posição central.
[8] Disponível para download em http://www.laurenceanthony.net/software.html.

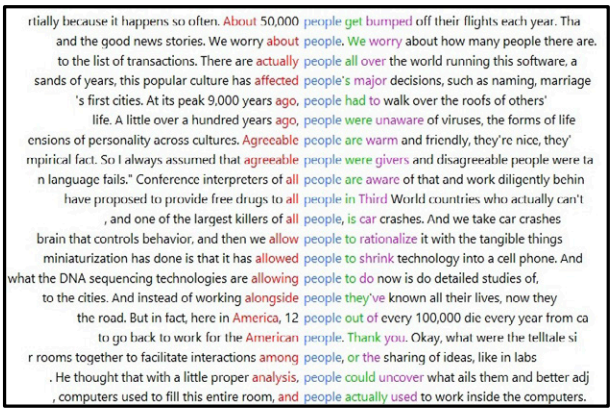


Figura 1 – Demonstração da ferramenta *concordance* no AntConc®
Fonte: Dados da pesquisa

O AntConc® analisa textos automaticamente, facilitando a coleta e a análise de dados. Como esse programa permite buscas ilimitadas para a formação de Word Lists, linhas de concordância e extração de palavras-chave, notamos que a demanda de pesquisadores e aprendizes que utilizam essa ferramenta para otimizarem seus estudos é crescente e, por esse motivo, apresentaremos, abaixo, algumas considerações sobre as ferramentas Word List, Keyword List e Cluster/N-Grams.

A ferramenta Wordlist produz listas de palavras que são organizadas de acordo com a sua frequência dentro do *corpus*, o que permite ao pesquisador rapidamente descobrir quais são os itens mais frequentes dentro de seu *corpus* de estudo, assim como mostrado na Figura 2:

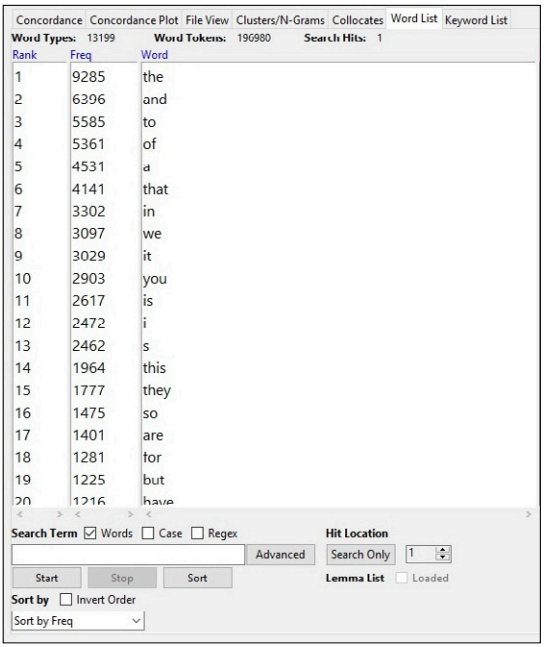


Figura 2 – Ferramenta Wordlist
Fonte: Dados da pesquisa

Segundo Viana (2010), essa ferramenta é de extrema importância porque,

> Com a geração de listas de palavras, o foco na proposição de uma ideia fica em segundo plano para dar espaço às escolhas lexicais que forma feitas na realização desta tarefa. O fato de um autor utilizar um número maior ou menor de determinado recurso linguístico consiste num provável traço de sua expressão estilística. De forma semelhante, se uma construção aparece mais frequentemente num determinado registro, ela é passível de ser compreendida como um traço inerente à constituição dele (VIANA, 2010, p. 44).

Já a ferramenta Key Word List apresenta, por meio de análises estatísticas, quais palavras são mais significativas quando comparadas com outro *corpus* de referência. Recomenda-se que o *corpus* de referência seja, no mínimo, cinco vezes maior do que o de estudo (BERBER SARDINHA, 2009). Justamente por isso, Berber Sardinha (2009) afirma que as palavras-chave não são o mesmo que palavras "importantes", visto que elas sempre serão relativas ao *corpus* de referência utilizado.

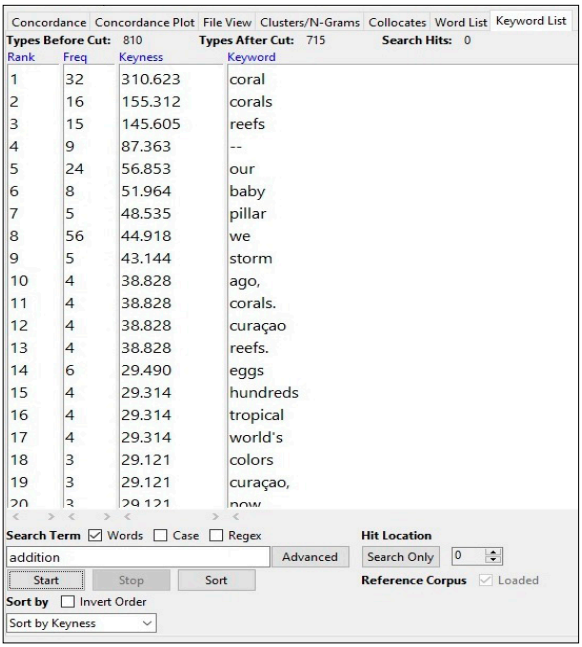


Figura 3 – Ferramenta Keyword List
Fonte: Dados da pesquisa

Na figura acima, ao compararmos um pequeno *corpus* de 810 palavras, compilado a partir de somente uma das palestras do *TED Talks* (MARHAVER, 2015), com a seção de ciências biológicas do *corpus* BASE, percebemos que o item lexical *coral* aparece em primeiro lugar com uma frequência de 32 ocorrências no *corpus* de estudo e com uma chavicidade[9] de 310.263. Todavia, percebe-se que o item *we*, apesar de aparecer com frequência numérica superior, com 56 ocorrências, possui chavicidade de 44.918, ou seja, esse item é estatisticamente menos significante no *corpus* de estudo do que o item *coral*.

Dessa forma, Berber Sardinha (2009, p. 194) elenca quatro fins para o uso desse recurso: *i)* identificar a temática de um *corpus* ou de um texto; *ii)* descrever a organização interna dos textos; *iii)* localizar marcas indicativas de posicionamento ideológico; e *iv)* traçar um perfil lexical de um autor ou de outros indivíduos.

Visto que uma lista de palavras não precisa ser necessariamente composta por formas únicas (VIANA, 2010), a ferramenta Cluster permite que o *software*

[9] O termo *chavicidade* é o resultado de um procedimento estatístico que levanta o quão importante cada palavra-chave é para o *corpus* de pesquisa em relação ao de referência. Quanto maior o valor apresentado, maior a relevância da palavra em questão (VIANA, 2010).

faça uma busca pelo *corpus* e liste os agrupamentos de palavras de acordo com a quantidade previamente estabelecida pelo pesquisador.

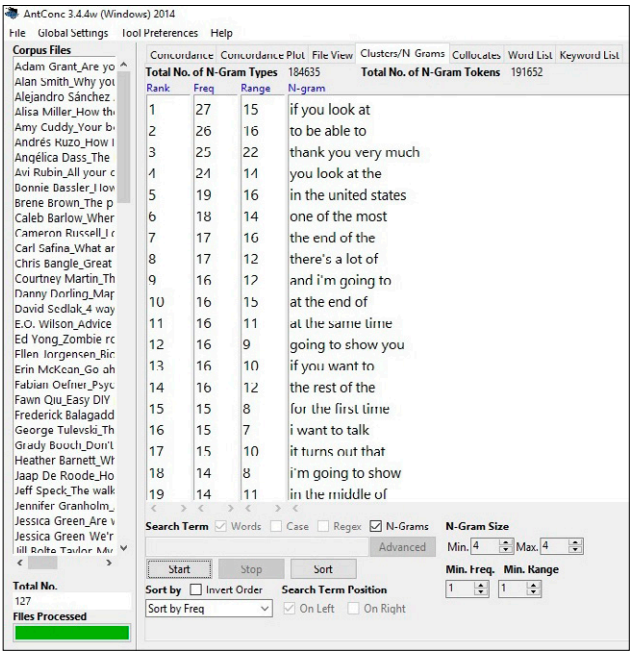


Figura 4 – Ferramenta Cluster
Fonte: Dados da pesquisa

Viana (2010) defende que as mesmas informações encontradas na lista com palavras individuais podem ser, também, aplicadas aos agrupamentos. Dessa forma, a gama de análises linguísticas, aplica-se tanto para quem interessa observar palavras isoladas quanto palavras em grupos. Além do mais, a existência de grandes quantidades desses padrões reafirma o caráter probabilístico e associativo da linguagem (BERBER SARDINHA, 2009).

Após a apresentação e descrição das principais ferramentas de análise lexical oferecidas pelo AntConc®, são relatados, na próxima subseção, quais os critérios estabelecidos para a criação das atividades que foram desenvolvidas.

3.2 Critérios para a preparação das atividades

As atividades apresentadas neste artigo respeitaram os critérios para a preparação de atividades com base em *corpora*, baseados no trabalho de Delfino (2016). A pesquisadora afirma que tais critérios foram desenvolvidos em um curso

intitulado "*Corpus*: ensino e análise", oferecido pelo Programa de Pós-Graduação em Estudos da Linguagem na PUC-SP em 2004 e ministrado pelo prof. dr. Tony Berber Sardinha. Ao total, são elencados 28 critérios, dentre os quais 11 são obrigatórios (O) e 17 são não obrigatórios (NO). Abaixo, segue a descrição de tais critérios.

3.2.1 Critérios obrigatórios

Antes de descrevermos detalhadamente cada item, o Quadro 1, abaixo, apresenta as características gerais dos critérios obrigatórios para atividades com base em *corpora*.

Quadro 1 – Lista dos critérios obrigatórios para a preparação de atividades

O1 –	O exercício faz uso de *corpus*;
O2 –	O exercício precisa ter enunciados claros;
O3 –	O exercício tem como foco principal o padrão lexicogramatical;
O4 –	O exercício é ético;
O5 –	O exercício é replicável;
O6 –	O exercício é motivador;
O7 –	O exercício não simplifica a língua usada nos textos/concordâncias/lista de palavras etc.;
O8 –	O exercício deve apresentar nível de dificuldade adequado;
O9 –	O exercício contém conteúdo relevante para o aluno e para a construção do conhecimento em inglês;
O10 –	O professor é facilitador e não distribuidor de conhecimento;
O11 –	O aluno é descobridor, pesquisador e não recipiente de conhecimento.

Fonte: DELFINO, 2016, p. 56

Critério O1 – O exercício usa o corpus. Esse critério foi respeitado ao utilizarmos *corpus* baseado em apresentações do *TED Talks* e *corpora* COCA e BASE.

Critério O2 – O exercício precisa ter enunciados claros. Tal critério é de extrema relevância em qualquer atividade desenvolvida, visto que o aluno precisa de um direcionamento compreensível sobre o que fazer durante o exercício apresentado.

Critério O3 – O exercício tem como foco principal o padrão lexico-gramatical. As atividades devem apresentar padrões da língua ao aluno para que ele saiba como os reconhecer. Esse critério ficou evidente durante as atividades de agrupamentos lexicais apresentadas.

Critério O4 – O exercício é ético. As atividades não devem expor o aluno a situações que sejam constrangedoras ou que agridam a sua integridade.

Critério O5 – O exercício é reaplicável. Todos os exercícios apresentados neste artigo podem ser impressos e reutilizados, inclusive, todos os vídeos estão disponíveis para *download*.

Critério O6 – O exercício é motivador. Novamente, tal critério deve ser obrigatório em todos os exercícios preparados por professores, visto que a motivação é um elemento fundamental na aprendizagem de uma segunda língua; é necessário que o aluno sinta prazer em realizar a atividade e em participar das aulas.

Critério O7 – O exercício não simplifica a língua utilizada nos textos e concordâncias/listas de palavras. Houve muitas discussões sobre a autenticidade do material apresentado ao aluno (MCENERY; XIAO, 2011), no entanto, ao utilizarmos a LC para ensino, é fundamental que a língua em uso não seja de nenhuma forma adaptada e/ou facilitada para que ocorra a compreensão dos alunos. O que deve acontecer, na verdade, é uma melhor adaptação da proposta do exercício e, não, do texto apresentado.

Critério O8 – O exercício deve apresentar nível de dificuldade adequado. Esse critério se relaciona intimamente com o sétimo, visto que parte da motivação do aluno vem do quanto ele consegue produzir em sala de aula.

Critério O9 – O exercício contém conteúdo relevante para o aluno e para a construção do conhecimento em inglês. Novamente, é possível afirmar que as atividades preencheram esse requisito ao analisarmos os *feedbacks* recebidos dos participantes. Há uma grande quantidade de comentários positivos em relação à escolha dos temas apresentados aos alunos.

Critério O10 – O professor é facilitador e não distribuidor do conhecimento. Para cumprir esse critério, é necessário que o professor compartilhe a responsabilidade de aprendizagem com seus alunos, apresentando a eles caminhos para que façam suas próprias conclusões sobre o uso da língua. Delfino (2016) lembra que a atenção do professor não deve estar somente no ensino do conteúdo, mas, sim, em estimular o aluno a buscar novas fontes de conhecimento.

Critério O11 – O aluno é descobridor, pesquisador e não recipiente de conhecimento. Complementando o critério O10, as atividades com base em *corpora* devem instigar o aluno a descobrir evidências linguísticas com base em padrões e em frequência dos dados.

Vistos os critérios obrigatórios para a criação de atividades com base em *corpora* descritos acima, na próxima seção, serão descritos os critérios opcionais que também podem estar presentes em tais atividades.

3.2.2 Critérios não obrigatórios

Aqui serão descritos os critérios não obrigatórios (NO), isto é, características que não precisam ser necessariamente seguidas pelo professor no momento

da confecção de atividades com *corpora* (DELFINO, 2016). O quadro abaixo apresenta as características gerais para tais critérios. Cada um dos 17 itens será descrito com maiores detalhes a seguir.

Quadro 2 – Lista dos critérios não obrigatórios para a preparação de atividades

NO1 – O exercício é participativo;
NO2 – O exercício é colaborativo;
NO3 – O exercício exige pouco tempo de preparação por parte do professor;
NO4 – O exercício trabalha com o conceito de padronização;
NO5 – O exercício lida com o conceito de frequência;
NO6 – O exercício trabalha com o conceito de variação;
NO7 – O exercício trabalha com o conceito de variedade textual;
NO8 – O exercício incorpora mídias diferentes;
NO9 – O exercício incorpora linhas de concordância;
NO10 – O exercício incorpora textos;
NO11 – O exercício incorpora lista de frequência de palavras;
NO12 – O exercício incorpora lista de palavras-chave;
NO13 – O exercício faz com que os alunos trabalhem diretamente com *corpora*;
NO14 – O exercício incorpora diagramas e diferentes formas de visualização;
NO15 – O exercício é de fácil adaptação;
NO16 – O exercício desenvolve a autonomia;
NO17 – O exercício ensina e não testa.

Fonte: DELFINO, 2016, p. 56

Critério NO1 – O exercício é participativo. Isso significa que a atividade em questão deve estimular, sempre que possível, a participação do aluno.

Critério NO2 – O exercício é colaborativo. Concomitante ao critério descrito acima, as atividades devem, sempre que possível, incentivar os alunos a chegar a conclusões dos exercícios de forma colaborativa, por meio de *pair works* ou discussões em grupos.

Critério NO3 – O exercício exige pouco tempo de preparação por parte do professor. Assim como na área de LinFE, a pesquisa com *corpora* pode exigir certa demanda de tempo do professor na preparação de materiais. Todavia, o uso de *softwares* para análises linguísticas (AntConc® e Word Smith Tools®) e de *corpora* disponíveis *on-line* (COCA, BASE, MICUSP, MICASE, entre outros) é grande facilitador desse processo, otimizando o tempo gasto conforme o professor domine melhor o uso desses recursos.

Critério NO4 – O exercício trabalha com o conceito de padronização. Sempre que for conveniente, a atividade deve apresentar ao aluno linhas de concordância para que ele perceba as tendências de combinações lexico-gramaticais da língua.

Critério NO5 – O exercício lida com o conceito de frequência. Visto que a LC analisa a língua por uma abordagem probabilística (BERBER SARDINHA, 2004). Esse critério visa a apresentar ao aluno, sempre que possível, quais as formas linguísticas mais frequentes encontradas em um determinado *corpus* de estudo.

Critério NO6 – O exercício trabalha com o conceito de variação. O objetivo é fazer com que o aluno tenha contato com os diversos tipos de variação lexical em sala de aula.

Critério NO7 – O exercício trabalha com o conceito de variedade textual. Assim como no critério acima, o objetivo é fazer com que os alunos tenham contato com diferentes registros de uso da língua.

Critério NO8 – O exercício incorpora mídias diferentes. Com os avanços tecnológicos, está cada vez mais fácil o professor de línguas utilizar diversos recursos em sala de aula para estimular seus alunos, tais como Facebook®, Youtube®, Twitter® e, até mesmo, as *TED Talks*, conforme é apresentado neste artigo.

Critério NO9 – O exercício incorpora linhas de concordância. Tal critério destaca o uso das linhas como forma de apresentar ao aluno o item lexical estudado (nódulo). Essa técnica permite que o aluno melhor visualize os padrões linguísticos encontrados no *corpus* de estudo, possibilitando, assim, que sejam feitas inferências e generalizações sobre a língua.

Critério NO10 – O exercício incorpora textos. Tal critério se faz presente nesta pesquisa por apresentar aos alunos exercícios de CE com base em textos retirados da internet que enfocaram o assunto abordado na aula.

Critério NO11 – O exercício incorpora a lista de frequência de palavras. Essa técnica facilita a percepção do aluno sobre quais são as palavras mais frequentes em um *corpus* de estudo.

Critério NO12 – O exercício incorpora a lista de palavras-chave. Ajuda o aluno a perceber qual o tema abordado no *corpus*, por meio de atividades de previsão, inferência textual.

Critério NO13 – O exercício faz com que os alunos trabalhem diretamente com corpora.

Critério NO14 – O exercício incorpora diagramas e diferentes formas de visualização. Esse critério colabora com os diferentes estilos de aprendizagem dos alunos, facilitando a compreensão por meio de atividades mais visuais, auditivas, cinestésicas, entre outras.

Critério NO15 – O exercício é de fácil adaptação. A atividade pode ser reproduzida para outras turmas e adaptada para alunos com diferentes níveis de proficiência linguística, facilitando, assim, o trabalho do professor.

Critério NO16 – O exercício desenvolve a autonomia do aluno contribuindo para a sua independência e responsabilidade pelo aprendizado.

Critério NO17 – O exercício ensina e não testa. As atividades preparadas devem ser utilizadas em sala como ferramentas para o ensino, e não para avaliações sobre a língua.

Depois de descritos os dois tipos de critérios nas seções acima, passa-se agora à descrição dos processos envolvidos na criação de algumas das atividades que foram propostas durante a pesquisa de mestrado.

3.3 Exemplo de atividades preparadas com base em corpora

Serão detalhados, agora, os processos metodológicos aplicados durante a criação de algumas atividades, visto que tais processos são de extrema importância para a criação de um elo entre as teorias sobre o ensino de LE e a prática em SA. As atividades propostas estão divididas em três etapas: *pre-listening*, *listening* e *post-listening*, conforme descrito abaixo.

3.3.1 Análise das atividades de *pre-listening*

As atividades de *pre-listening* tiveram por objetivo ajudar os alunos a obterem conhecimento de mundo, indicar o contexto dos exercícios e estimular os alunos a uma participação mais ativa em SA. Visto isso, passa-se à descrição de tais exercícios e de como eles se relacionam com a metodologia desta pesquisa.

Quadro 3 – Exemplo de atividade de *pre-listening* #1

LISTENING – WARM-UP

You will watch a video from the National History Museum about the importance of Coral Reefs. Before listening, write down three main ideas you think will be mentioned in the video.

From: https://www.youtube.com/watch?v=eNqbSi_6KdA&t=1s

01._____.
02._____.
03._____.

Now watch the video and check with a partner if your predictions were correct.

Fonte: Elaborado pelos autores

O quadro acima apresenta uma atividade com o intuito de incentivar o aluno a, primeiro, tentar prever o assunto que será abordado durante a aula, utilizando seu próprio conhecimento de mundo, para, então, observar e analisar imagens, gráficos e diagramas apresentados durante o vídeo como forma de ativação do conhecimento prévio. Caso o aluno tenha pouco ou nenhum tipo de conhecimento prévio sobre o assunto, a atividade é responsável por suprir essa fragilidade.

Visto que a atividade acima não foi produzida com base no *corpus* de estudo, ela não precisou responder a todos os critérios obrigatórios de preparação de atividades descritos anteriormente. Todavia, com exceção dos critérios *O1– o exercício faz uso de corpus* e *O3 – o exercício tem como foco o padrão lexico-gramatical*, todos os critérios obrigatórios foram preenchidos; entre os não obrigatórios, podemos observar a presença do *NO1 – o exercício é participativo*, *NO2 – o exercício é colaborativo*, *NO8 – o exercício incorpora mídias diferentes* e *NO15 – o exercício é de fácil adaptação*.

Outras atividades de *pre-listening* também foram preparadas, conforme exposto abaixo. O Quadro 4 apresenta um exercício que abordou o ensino de agrupamentos lexicais. A LC mostrou-se presente em todos os passos metodológicos, cumprindo todos os critérios obrigatórios e diversos não obrigatórios, além de desenvolver, no aluno, a estratégia de compreensão proposta pelo modelo ascendente, ou seja, a combinação das palavras, dos sons e da gramática presente no texto.

Quadro 4 – Exemplo de atividade de *pre-listening* #2

Listen to different excerpts from original TED Talks, fill in the blanks using the correct bundle.

01. "[…] *so,* _____ *the green data points, which is air that's outside, you'll see that there's a large amount of microbial diversity* […]"

02. *"[…] but before I do that,* _____ *a little bit about the past […]"*

03. *"[…] and yes, I agree* _____ *little ego satisfaction in being able to say, "Hey, I was the first one to discover that […]"*

04. *"[…]* _____ *most of the world's major religions, you will find seekers – Moses, Jesus, Buddha, Muhammad […]"*

05. *"[…] and without understanding what attackers can do and the security risks from the beginning,* _____ *danger in this […]"*

Fonte: Elaborado pelos autores

Nesse recorte da atividade, é possível perceber uma maior influência do modelo ascendente e os tipos de estratégias que o aluno deve utilizar para a realização do exercício. Em particular, duas estratégias contribuem diretamente para a CO do aluno; elas são: a compreensão de *detalhes presentes no texto* e o *reconhecimento de padrões de ordem de palavras*.

Nesse exercício, todos os critérios obrigatórios foram respeitados; já entre os não obrigatórios, estão presentes os critérios: *NO3 – o exercício exige pouco tempo de preparação por parte do professor*, *NO5 – o exercício lida com o conceito de frequência*, visto que os agrupamentos apresentados foram os mais recorrentes dentro do *corpus* de estudo, e *NO9 – o exercício incorpora linhas de concordância*.

Na próxima subseção, abordam-se mais algumas atividades preparadas durante a pesquisa.

3.3.2 Atividades de *listening*

O Quadro 5 apresenta uma atividade de CO do tipo de *cross out*, ou seja, o aluno deve ouvir a um recorte de uma *TED Talk* e escolher quais dos marcadores discursivos em destaque foram utilizados pelo palestrante. Nessa atividade, o vídeo da palestra foi recortado em partes menores, entre 2 e 3 minutos, utilizando o *software* Windows Movie Maker®; após isso, o professor-pesquisador utilizou o *script* (BROWN, 2010) do trecho recortado para a criação do exercício, conforme é apresentado abaixo.

Quadro 5 – Exemplo de atividade de *listening* #1

Listen to the beginning of a talk and cross out the extra words.

All right / So, I'll start with this: a couple years ago, an event planner called me because I was going to do a speaking event. ***And / Then*** she called, and she said, "I'm really struggling with how to write about you on the little flyer." And I thought, "***Now / Well***, what's the struggle?" and she said, "***Well / Anyway***, I saw you speak, and I'm going to call you a researcher, ***I think / In fact***, but ***I think / I'm afraid*** if I call you a researcher, no one will come, because they'll think you're boring and irrelevant." And ***maybe / I was like***, "Okay." And she said, "But the thing I liked about your talk is you're a storyteller. So, ***I think / of course*** what I'll do is ***just / perhaps*** call you a storyteller." and ***probably / of course***, the academic, insecure part of me was ***sort of / like***, "You're going to call me a what?" And she said, "I'm going to call you a storyteller." And I was like, "Why not 'magic pixie'?"

I was like, "Let me think about this for a second." I tried to call deep on my courage. And I thought, ***you know / you see***, I am a storyteller. I'm a qualitative researcher. I collect stories; that's what I do. And ***I think / maybe*** stories are just data with a soul. And ***apparently / maybe*** I'm ***maybe / just*** a storyteller. And so I said, "You know what? Why don't you ***just / probably*** say I'm a researcher-storyteller." And she went, "Ha ha. There's no such thing."

So / Anyway I'm a researcher-storyteller, and I'm going to talk to you today — we're talking about expanding perception — and ***now / so*** I want to talk to you and tell some stories about a piece of my research that fundamentally expanded my perception and really ***basically / actually*** changed the way that I live and love and work and parent.

Fonte: Elaborado pelos autores

De modo similar, abaixo também é apresentado um exercício com o foco no uso de Marcadores Discursivos. Novamente, foi recortada uma parte de uma *TED Talk* (SHELLENBERGER, 2016), e o *script* dela foi utilizado para a realização do exercício. Entretanto, dessa vez, o tipo de atividade foi de preenchimento de lacunas.

Nesse exercício, os alunos deveriam completar as lacunas presentes no texto de acordo com o áudio da palestra. Vale lembrar que os alunos já haviam sido introduzidos ao conceito e ao uso de Marcadores Discursivos, dessa forma, o exercício requereu que fossem completadas as lacunas com as palavras/expressões disponíveis previamente no exercício.

Quadro 6 – Exemplo de atividade de *listening* #2

Complete the blanks spaces using the words from the box, then listen again and check your answers.

I mean	Actually	You know	That means	So
In fact	Maybe (2x)	Well	In other words	Then

What we found is that the world is _____ at risk of losing four times more clean energy than we lost over the last 10 years. _____: we're not in a clean energy revolution; we're in a clean energy crisis. _____ it's understandable that engineers would look for a technical fix to the fears that people have of nuclear. But when you consider that these are big challenges to do, that they're going to take a long time to solve, there's this other issue, which is: Are those technical fixes really going to solve people's fears? Let's take safety. _____, despite what people think, it's hard to figure out how to make nuclear power much safer. _____, every medical journal that looks at it — this is the most recent study from the British journal, "Lancet," one of the most respected journals in the world — nuclear is the safest way to make reliable power *[…]* and the truth is that even if we get good at using that waste as fuel, there's always going to be some fuel left over. _____ there's always going to be people that think it's a big problem for reasons that _____ don't have as much to do with the actual waste as we think. _____, what about the weapons? _____ the most surprising thing is that we can't find any examples of countries that have nuclear power and _____, "Oh!" decide to go get a weapon. _____, it works the opposite. What we find is the only way we know how to get rid large numbers of nuclear weapons is by using the plutonium in the warheads as fuel in our nuclear power plants.

Fonte: Elaborado pelos autores

Ambas as atividades apresentadas respeitam todos os critérios obrigatórios referentes a exercícios preparados com base em *corpora*; em relação aos critérios não obrigatórios, é possível percebemos a presença dos critérios *NO3 – o exercício exige pouco tempo de preparação por parte do professor*, *NO8 – o exercício incorpora mídias diferentes* e *NO15 – o exercício é de fácil adaptação*.

A seguir, serão descritos os exercícios caracterizados como *post-listening*, ou seja, com a função de reforçar e finalizar o conteúdo discutido em sala. São apresentados exemplos desses exercícios e é exposto o modo como eles se relacionam com a metodologia do uso de *corpora* para o ensino de línguas.

3.3.3 Atividades de *post-listening*

Nesta subseção, será discutida a última classificação das atividades preparadas: as atividades utilizadas como *post-listening*. Tais exercícios tiveram o intuito de fechar o assunto discutido em sala e de ajudar os alunos a reforçarem o conteúdo que lhes foi ensinado, conforme se vê nos exemplos abaixo.

Quadro 7 – Exemplo de atividade de *post-listening* #1

Below there are some collocations related to the word "Democracy", discuss with a partner and complete the following exercises.

Liberal	Social	Representative	Political
Participatory	Toward	Constitutional	Restore
Direct	Movement	True	Promote

A. *Mr. Chavez's neighborhood associations that sought to bring* _____ ***democracy*** to people who had long been shut out of power by the elites.

B. *They also can* _____ ***democracy*** on the ground by their support for nonviolent groups and national reconciliation.

C. *The American Left believes that socially conservative values are incompatible with* _____ ***democracy****.*

D. *We need to re-establish this party as the party of* _____ ***democracy****. That is at the heart of what I believe.*

E. *Twenty years after the military withdrew from power, Brazilians realize that* _____ ***democracy*** is not economic democracy.

Fonte: Elaborado pelos autores

O quadro acima apresenta um recorte do exercício de compreensão escrita (CE) voltado ao ensino de colocações[10], tendo em vista que o assunto abordado na *TED Talk* foi sobre o funcionamento da democracia em Atenas. Após as discussões feitas em aula e a realização do exercício de CO, foi entregue aos alunos uma lista com 12 dos colocados mais frequentes com o item *Democracy*, retirado do Corpus *of Contemporary American English* (COCA), conforme Figura 5 abaixo.

[10] Tagnin (2011) explica colocações como a coocorrência de duas (ou mais) palavras numa frequência maior do que seria de se esperar caso a coocorrência fosse aleatória.

No exercício com colocações, os alunos deveriam ler algumas linhas de concordâncias retiradas do COCA e preencher as lacunas com o colocado que melhor completasse as linhas de concordância. Pelo fato de essa atividade poder estar um pouco acima do nível de aprendizado dos participantes, foi pedido que os alunos a realizassem em dupla, aumentando, assim, a aprendizagem colaborativa em sala de aula.

Nesse exercício, todos os critérios obrigatórios para a preparação de atividades com *corpora* foram respeitados; em relação aos critérios não obrigatórios, estão presentes o *NO2 – o exercício é colaborativo*, o *NO3 – o exercício exige pouco tempo de preparação do professor*, o *NO9 – o exercício incorpora linhas de concordância*, o *NO13 – o exercício faz com que os alunos trabalhem diretamente com* corpora, o *NO16 – o exercício desenvolve a autonomia* e o *NO17 – o exercício ensina e não testa*.

		CONTEXT	FREQ
1		LIBERAL	670
2		AMERICAN	477
3		POLITICAL	389
4		TOWARD	336
5		MOVEMENT	212
6		SOCIAL	180
7		REPRESENTATIVE	153
8		PROMOTING	140
9		PROMOTE	137
10		PARTICIPATORY	130
11		CONSTITUTIONAL	130
12		PARLIAMENTARY	114
13		RESTORE	114
14		WESTERN	114
15		MODERN	110
16		TRUE	104
17		DIRECT	95

Figura 5 – Colocados com o item *Democracy* retirado do COCA
Fonte: *Corpus* COCA

Outra atividade de *post-listening* proposta visou a melhorar a CE dos alunos por meio do uso de marcadores discursivos. Um dos objetivos foi criar oportunidades para que os alunos desenvolvessem o pensamento crítico-reflexivo sobre o

uso da língua e aplicassem o conteúdo aprendido durante a aula para a resolução da tarefa, conforme Quadro 8.

Quadro 8 – Exemplo de atividade de *post-listening* #2

Read the text below and answer the questions.

"*we need to break these down into doable steps. you want to write a family history, you can read some other family histories, get a sense for the style. Think about the questions you want to ask your relatives, set up appointments to interview them. Or you want to run a 5K*"

A) In your opinion, is it possible to understand the idea of the text?
B) Do you think the sentences in the text are well connected?
C) Read the text again, and try to complete it using the Discourse Markers in the box.

Then	So	Now
Maybe (3x)	First	And

"*1._____ 2._____ we need to break these down into doable steps. 3._____ 4._____ you want to write a family history. 5._____, you can read some other family histories, get a sense for the style. 6._____ 7._____think about the questions you want to ask your relatives,* set up ap*pointments to interview them. Or 8._____ you want to run a 5K.*"

D. What are the functions of the words "And", "*First*", "*Maybe*" and "*Then*" in this text?_____

Fonte: Elaborado pelos autores

Conforme mencionado, o exercício acima buscou ajudar os alunos a utilizarem de maneira mais eficaz os marcadores discursivos por meio da habilidade de CE e de discussões sobre o uso da língua, ativando, dessa maneira, a consciência linguística do aluno e reforçando o conteúdo aprendido em sala.

Em relação aos critérios para a preparação de atividades com *corpora*, todos os obrigatórios foram obedecidos. Entre os não obrigatórios, percebemos a presença do *NO1 – o exercício é participativo, NO2 – o exercício é colaborativo, NO10 – o exercício incorpora textos*, *NO16 – o exercício incorpora autonomia* e *NO17 – o exercício ensina e não testa*.

Ainda em relação às atividades com foco nos marcadores discursivos, o quadro abaixo apresenta mais um exercício utilizado como *post-listening*. Essa

atividade é uma sequência direta do exercício apresentado no Quadro 5. Aqui, os alunos devem, após a correção da atividade anterior, elencar quais as funções dos marcadores utilizados pela palestrante durante seu discurso.

Quadro 9 – Exemplo de atividade de *post-listening* #3

Now match the discourse markers according to its function.

OPEN/CLOSE A TOPIC	SEQUENCING	SHARED KNOWLEDGE	STANCE	HEDGES

Fonte: Elaborado pelos autores

Esse exercício mostra-se importante por transcender a simples identificação sonora dos itens lexicais trabalhados e requerer que o aluno reflita sobre o uso da língua em contextos autênticos de comunicação. Visto que tal atividade está diretamente ligada a um exercício anterior, ela também respeita todos os critérios obrigatórios para as atividades envolvendo *corpora*.

Dentre os não obrigatórios, podemos observar a presença dos *NO3 – o exercício exige pouco tempo de preparação do professor*, *NO10 – o exercício incorpora textos*, *NO14 – o exercício incorpora diferentes formas de visualização*, *NO16 – o exercício desenvolve a autonomia* e *NO17 – o exercício ensina e não testa*.

4 Conclusões

Este estudo, desenvolvido à luz de teorias sobre o uso da LC para o ensino de LE, teve como objetivo desenvolver, descrever e aplicar atividades didáticas baseadas em *corpora* em uma instituição de ensino superior tecnológica no interior do estado de São Paulo. Especificamente, procurou-se desenvolver atividades baseadas em um *corpus* oral de pequeno-médio porte a partir de apresentações dos *sites* TED e TED-ED. Após o desenvolvimento das atividades, elas foram aplicadas em um minicurso de 12 horas.

Graças aos dados coletados no decorrer da pesquisa de mestrado, é possível afirmar que o uso das *TED Talks* provou-se útil para a compilação de um *corpus* oral, assim como uma excelente ferramenta para o ensino de LE em contextos acadêmicos. Apesar de o *corpus* em estudo ser considerado pequeno em relação a

outros *corpora* existentes (BASE, COCA, BNC), ele mostrou-se funcional. Mais pesquisas envolvendo ampliação e novas análises do *corpus* de estudo estão sendo planejadas para estudos futuros.

Além disso, foi possível constatarmos que os exercícios foram bem recebidos pelos participantes do estudo e estavam em um nível de dificuldade adequado, segundo o *feedback* recebido dos alunos. Todavia, é importante lembrar que fatores como a carga horária do minicurso e a quantidade do conteúdo apresentado receberam alguns *feedbacks* negativos; esses dados, receberão maior atenção em estudos futuros.

Visto isso, os conhecimentos produzidos a partir desta pesquisa representam apenas um estudo de caso dentro de um dos campos que mais tem crescido na área da LC – *corpora* como recursos na criação de atividades de CO, bem como, a implementação e a avaliação das atividades desenvolvidas. Entretanto, faz-se necessário aprofundar o tema, principalmente no que diz respeito à compilação de *corpora* orais com base em fontes diversas, quebrando, assim, o paradigma tradicional do "falante nativo" como balizador da produção linguística, que ainda é dominante nas pesquisas em LC. Desse modo, esse tipo de pesquisa poderia propiciar, tanto a professores quanto a alunos, diferentes modelos pedagógicos na criação de atividades didáticas, além da simples utilização de *corpora* no processo de ensino-aprendizagem de línguas.

Referências

ADOLPHS, S.; KNIGHT, D. Building a spoken *corpus*: what are the basics? In: O'KEEFFE, A.; MCCARTHY, M. (Ed.). *The Routledge handbook of* corpus *linguistics*. New York: Routledge, 2010, p. 38-52.

ANTHONY, L. *AntConc.* (Version 3.4.4) [Computer Software]. Tokyo: Waseda University, 2011.

BERBER SARDINHA, T. Como usar a linguística de *corpus* no ensino de língua estrangeira: Por uma linguística de *corpus* educacional brasileira. In: VIANA, V.; TAGNIN, S. E. O. (Org.). Corpora *na Tradução*. 1. ed. São Paulo: Hub Editorial, 2015, p. 301-356.

_____. *Pesquisa em Linguística de* Corpus *com o Word Smith Tools*. Campinas: Mercado de Letras, 2009.

_____. *Linguística de* Corpus. Barueri: Manole, 2004.

BROWN, B. *The power of vulnerability.* TEDx Houston, jun. 2010. Disponível em <https://www.ted.com/talks/brene_brown_on_vulnerability>. Acesso em: 19 maio 2016.

BROWN, H. D. Teaching Listening. In: _____. *Teaching by principles*: An interactive approach to language pedagogy. 2. ed. New York: Longman, 2001, p. 247-266.

CARTER, R.; MCCARTHY, M. *Cambridge grammar of English*: a comprehensive guide; spoken and written English grammar and usage. Cambridge: Cambridge University Press, 2006.

CHARLES, M. English for Academic Purpose. In: PALTRIDGE, B.; STARFIELD, S. (Org.). *The Handbook of English for Specific Purposes*. Chichester: John Wiley & Sons, Ltd. 2012, p. 137-154.

CHAUDRON, C.; LOSCHKY, L.; COOK, J. Second language listening comprehension and lecture note-taking. In: FLOWERDEW, J. (Org.). *Academic listening*: Research perspectives. Cambridge: Cambridge University Press, 1994, p. 75-92.

CHENG, W. What can a *corpus* tell us about language teaching? In: O'KEEFFE, A.; MCCARTHY, M. (Ed.). *The Routledge handbook of* corpus *linguistics*. New York: Routledge, 2010, p. 319-332.

COXHEAD, A. A new academic word list. *TESOL quarterly*, v. 34, n. 2, 2000, p. 213-238.

DELFINO, M. C. N. *Uso de música para o ensino de inglês como língua estrangeira em um ambiente baseado em* Corpus. 2016. 159f. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem). LAEL, PUC, São Paulo, 2016.

FIELD, J. The changing face of listening. In: RICHARDS, J. C.; RENANDYA, W. A. (Ed.). *Methodology in language teaching*: An anthology of current practice. Cambridge: Cambridge University Press, 2002, p. 242-247.

FLOWERDEW, J. Research of relevance to second language lecture comprehension – an overview. In: _____. (Org.). *Academic listening*: Research perspectives. Cambridge: Cambridge University Press, 1994, p. 7-30.

FLOWERDEW, J.; MILLER, L. *Second language listening: Theory and practice*. Cambridge: Cambridge University Press, 2005.

FLOWERDEW, J.; PEACOCK, M. The EAP curriculum: Issues, methods, and challenges. In: _____. (Ed.). *Research perspectives on English for academic Purposes*. Cambridge: Cambridge University Press, 2001, p. 177-194.

FLOWERDEW, L. Needs analysis and curriculum development in ESP. In: PALTRIDGE, B.; STARFIELD, S. (Org.). *The Handbook of English for Specific Purposes*. Chichester: John Wiley & Sons, Ltd. 2012, p. 325-346.

_____. The exploitation of small learner *corpora* in EAP materials design. In: GHADESSY, M.; HENRY, A.; ROSEBERRY, R. L. (Ed.). *Small* corpus *studies and ELT*: theory and practice. Amsterdam: John Benjamins Publishing, 2001, p. 363-379.

GOH, C. C. M. ESP and Listening. In: PALTRIDGE, B.; STARFIELD, S. (Org.). *The Handbook of English for Specific Purposes*. Chichester: John Wiley & Sons, Ltd., 2012, p. 55-76.

GRANGER, S. A bird's-eye view of learner *corpus* research. In: GRANGER, S.; HUNG, J.; PETCH-TYSON, S. (Ed.). *Computer learner* corpora*, second language acquisition and foreign language teaching*. Amsterdam: Benjamins, 2002, p. 3-33.

LUZÓN, M. J.; CAMPOY, M. C.; SANCHEZ, M. M.; SALAZAR, P. Spoken *Corpora*: New Perspectives in Oral Language Use and Teaching. In: CAMPOY, M. C.; LUZÓN, M. J. (Ed.). *Spoken* corpora *in applied linguistics*. Bern: Peter Lang Pub, 2007, p. 3-30.

MARHAVER, K. *How we're growing baby corals to rebuild Reefs.* TED Mission Blue II, out. 2015. Disponível em: <https://www.ted.com/talks/kristen_marhaver_how_we_re_growing_baby_corals_to_rebuild_reefs>. Acesso em: 09 maio 2016.

MAURANEN, A. Investigating English as a lingua franca with a spoken *corpus*. In: CAMPOY, M. C.; LUZÓN, M. J. (Ed.). *Spoken* corpora *in applied linguistics*. Bern: Peter Lang Pub, 2007, p. 33-56.

MCENERY, T.; XIAO, R. What *corpora* can offer in language teaching and learning. In: HINKEL, E. (Ed.). *Handbook of research in second language teaching and learning.* New York: Routledge, 2011, p. 364-380.

O'KEEFFE, A.; MCCARTHY, M.; CARTER, R. *From* corpus *to classroom*: Language use and language teaching. Cambridge: Cambridge University Press, 2007.

PAIVA, V. L. M. P. O. Os desafios na produção de materiais didáticos para o ensino de línguas no ensino básico. *Revista (Com)Texto em Estudos Linguísticos*. Vitória, v. 8, n. 10.1, p. 344-357, 2014.

RICHARDS, J. C. Listening comprehension: Approach, design, procedure. *TESOL quarterly*, v. 17, n. 2, p. 219-240, 1983.

ROST, M. *Teaching and researching listening*. 2. ed. Harlow: Pearson Education Limited, 2011.

SATO, E. T. N. *A compreensão oral nos Cadernos de Língua Estrangeira Moderna - Inglês do Estado de S. Paulo.* Campinas: [s.n.], 2011. 137f. Dissertação (Mestrado em Linguística Aplicada). Instituto de Estudos da Linguagem, UNICAMP, Campinas, 2011.

SHELLENBERGER, M. *How fear of nuclear power is hurting the environment*. TEDSummit, 2016. Disponível em <https://www.ted.com/talks/michael_shellenberger_how_fear_of_nuclear_power_is_hurting_the_environment>. Acesso em: 03 jan. 2017.

SINCLAIR, J. M. *Corpus* and Text - Basic Principles. In: WYNNE, M. (Ed.). *Developing Linguistic* Corpora: a Guide to Good Practice. Oxford: Oxbow Books, 2005. Disponível em: <https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm>. Acesso em: 09 abr. 2017.

TAGNIN, S. E. O. Glossário de linguística de *corpus*. In: TAGNIN, S.; VIANA, V. Corpora *no ensino de línguas estrangeiras.* São Paulo: HUB Editorial, 2011, p. 357-361.

TAKAESU, A. TED talks as an extensive listening resource for EAP students. *Language Education in Asia*, 4, p. 150-162, 2013.

TED. Ideas worth spreading. Disponível em <https://www.ted.com/>. Acesso em 12 de outubro de 2017.

TED-ED. Lessons worth sharing. Disponível em <https://ed.ted.com/>. Acesso em 02 de setembro de 2017.

WALSH, S. What features of spoken and written *corpora* can be exploited in creating language teaching materials and syllabuses. In: O'KEEFFE, A.; MCCARTHY, M. (Ed.). *The Routledge handbook of* corpus *linguistics*. New York: Routledge, 2010, p. 319-332.

Índice remissivo

Developing a rule-based Brazilian Portuguese-to-Libras machine translation system

Proposta de um vocabulário bilíngue de festas populares brasileiras baseada em um estudo de corpus

Frames *de compreensão e* corpora*: estudo de caso com uso do Sketch Engine*

É possível falar em estilo da tradução em legendagem?

Elaboração de um protótipo de glossário bilíngue (português-inglês) de treinamento de força: subsídios para o tradutor

Dicionário Olímpico: *a Semântica de* Frames *encontra a lexicografia eletrônica*

Colocações especializadas na área do direito comercial internacional e proposta de glossário trilíngue

O uso de corpus paralelo e comparável para descrever padrões de uso na tradução de abreviaturas e acrônimos de termos médicos

Identificação de termos no discurso literário de fantasia da série Harry Potter em uma abordagem direcionada por corpus

Pommersche Korpora: *um conjunto de* corpora *dialetais da variedade brasileira do pomerano*

Construções de tópico do português brasileiro falado em áreas indígenas

Para a segmentação automática de fronteira na fala espontânea a partir de parâmetros prosódicos

Fluência e interação no inglês aeronáutico: uma análise baseada em pragmática e Linguística de Corpus

Aos professores, as colocações

Brazilian students' use of english academic vocabulary: an exploratory study

Atividades de compreensão oral com base em corpora *das* TED talks*: um estudo-piloto*

Os autores

ComentCorpus: *o uso de mecanismos linguísticos na detecção de ironia e sarcasmo para o português do Brasil em um* corpus *opinativo*

Gabriela Wick Pedro
Mestranda do Programa de Pós-Graduação em Linguística pela Universidade Federal de São Carlos (UFSCar), na linha Descrição, Análise e Processamento de Linguagem Natural, e graduada em Bacharelado em Linguística pela mesma universidade. Tem experiência e interesse em construção, análise e anotação de *corpus* aplicados em Processamento de Língua Natural (PLN) e Análise de Sentimentos.

Oto Araújo Vale
Graduou-se em Sciences du Langage pela Université de Paris VIII (1989), tem mestrado (DEA) em Sciences du Langage (DEA), pela mesma instituição (1990), e doutorado em Linguística e Língua Portuguesa pela Universidade Estadual Paulista Júlio de Mesquita Filho (2002). Foi professor visitante na Universidade do Algarve no programa Erasmus Mundus, da Comunidade Europeia (2008), e pesquisador convidado na Université Catholique de Louvain (2014/2015). É professor associado do Departamento de Letras da Universidade Federal de São Carlos, onde atua na graduação e na pós-graduação. Tem experiência na área de Linguística, com ênfase em Linguística Computacional, atuando principalmente nos seguintes temas: expressões cristalizadas, léxico-gramática, dicionários eletrônicos e Linguística de *Corpus*.

O discurso dos deputados na votação do impeachment*: a LC combinada à ACD*

Rozane Rodrigues Rebechi
Professora adjunta da Universidade Federal do Rio Grande do Sul (UFRGS), atuando na graduação e na pós-graduação do Instituto de Letras. Possui graduação em Tradução pela Faculdade Ibero-Americana e especialização em Tradução e mestrado e doutorado em Estudos Linguísticos e Literários em Inglês pela Universidade de São Paulo (USP). Suas principais áreas de atuação são Ensino de Inglês como Língua Estrangeira, Tradução e Terminologia, com enfoque na utilização da Linguística de *Corpus* como metodologia. É coautora do *Vocabulário para química* (São Paulo: SBS, 2007) e autora de artigos e capítulos de livros publicados no Brasil e no exterior. Atualmente, coordena o projeto Culinária para Fins Acadêmicos junto à UFRGS.

Clustering and hierarchical organization of opinion aspects: a corpus *study*

Thiago Alexandre Salgueiro Pardo
Possui graduação em Bacharelado em Ciência da Computação pela Universidade Federal de São Carlos (1999), mestrado em Ciência da Computação pela Universidade Federal de São Carlos (2002) e doutorado em Ciências da Computação e Matemática Computacional pela Universidade de São Paulo (2005), onde também realizou estágio de pós-doutorado (2005). Atualmente é professor associado da Universidade de São Paulo. Tem experiência na área de Inteligência Artificial, atuando principalmente nos temas de processamento de linguagem natural, ou linguística computacional, mais especificamente nas áreas de sumarização automática de textos, análise discursiva e mineração de opiniões.

Francielle Alves Vargas
Possui mestrado em ciência da computação e matemática computacional pela Universidade de São Paulo (ICMC-USP). Cursou Bacharelado em Linguística pela Universidade Federal de Minas Gerais (UFMG) e Bacharelado em Sistemas de Informação pela Pontifícia Universidade Católica de Minas Gerais (PUCMG). Tem experiência na área de Processamento de Línguas Naturais, sobretudo com Mineração de Opinião/Análise de Sentimentos e classificação de textos. Atualmente, integra o Núcleo Interinstitucional de Linguística Computacional da Universidade de São Paulo (NILC-USP).

Tony Berber Sardinha
Professor associado do Departamento de Linguística e do Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem da Pontifícia Universidade Católica de São Paulo (PUCSP). É Ph.D. pelo Departamento de Inglês da Universidade de Liverpool (Inglaterra) sob a orientação de Michael Hoey. Atua nas áreas de Linguística de *Corpus*, metáfora e Linguística Aplicada há mais de 20 anos e coordena o Grupo de Pesquisa em Linguística de *Corpus* (GELC).

Maria Claudia Nunes Delfino
Doutoranda no programa de Linguística Aplicada e Estudos da Linguagem da Pontifícia Universidade Católica de São Paulo (PUCSP), onde também obteve o título de mestre enfocando a Linguística de *Corpus* e Ensino de Línguas, ambos sob orientação de Tony Berber Sardinha. É especialista em Linguística e Ensino de Línguas pela Uniseb e trabalha atualmente na Faculdade de Tecnologia de São Paulo de Praia Grande (FATEC-PG). Membro do Grupo de Estudos em Linguística de *Corpus* (GELC), seu principal interesse é a aplicação de Linguística de *Corpus* no ensino e aprendizagem de língua estrangeira e o estudo das dimensões lexicais em letras de música.

Rafael Fonseca de Araújo
Mestre em Linguística Aplicada e Estudos da Linguagem da Universidade Católica de São Paulo (PUCSP) sob orientação de Tony Berber Sardinha. Professor de Inglês Instrumental em escolas técnicas estaduais do Centro Paula Souza e da Universidade Metropolitana de Santos (UNIMES). Membro do Grupo de Estudos em Linguística de *Corpus* (GELC), tem experiência na área de Linguística de *Corpus* e Análise Multidimensional, atuando principalmente na construção de *corpora* e linguagem de televisão, em especial *reality TV shows*.

Developing a rule-based Brazilian Portuguese-to-Libras machine translation system

Francisco Aulisio dos Santos Paiva
Licenciado em Matemática pelo Instituto Federal de Educação, Ciência e Tecnologia do Ceará (2013), possui mestrado em Matemática Aplicada pela Universidade Estadual de Campinas (2015). Doutorando em Engenharia Elétrica na Faculdade de Engenharia Elétrica e de Computação, FEEC/UNICAMP. Atualmente desenvolve pesquisas na área de Inteligência Artificial, mais especificamente em Processamento de Linguagem Natural e tradução automática de português brasileiro para língua de sinais brasileira.

Plínio Almeida Barbosa
Linguista com graduação em Engenharia Eletrônica pelo ITA (1988), mestrado em Engenharia Eletrônica e Computação pelo mesmo Instituto (1990) e doutorado em Signal-Image-Parole/Option Parole pelo INP de Grenoble, França (1994). Tem livre docência em Fonética e Fonologia pela UNICAMP (2006), onde é professor associado 3 que atua na área de Fonética experimental nos temas: análise e modelamento dinâmicos da prosódia da fala, teoria de osciladores acoplados, Libras e ensino de fonética do francês. Tem mais de 100 publicações em periódicos e anais de eventos especializados. É o autor dos livros "Incursões em torno do ritmo da fala" (Campinas: Pontes) e "Manual de Fonética Acústica Experimental" (São Paulo: Cortez), esse último com Sandra Madureira (PUC-SP).

Pablo Picasso Feliciano de Faria
Doutor em Linguística pela Universidade Estadual de Campinas, atualmente é docente na mesma instituição, com ênfase em Aquisição de Linguagem e Linguística Computacional. Tem interesse nas áreas de formalismos gramaticais, Aquisição de Linguagem, teorias de aprendibilidade, processamento automático (*parsing*, tradução etc.) e Psicolinguística. Sua formação é interdisciplinar, sendo bacharel em Ciências da Computação, com aperfeiçoamento em Sistemas de Informação. Atuou como desenvolvedor e analista de sistemas comerciais entre 1998 e 2007. Entre 2000 e 2006, atuou também como artista (cantor e violonista) profissional, tendo gravado um álbum solo, em 2006, interpretando canções autorais.

José Mario De Martino
Engenheiro Eletricista pela Faculdade de Engenharia Elétrica e de Computação, FEEC/UNICAMP (1981). Possui mestrado (1986), doutorado (2005) e livre docência (2016) pela FEEC. É professor associado da FEEC, coordenador do "Laboratório de Alto Desempenho e Ambiente 3D, Imersivo Interativo para Visualização Científica" da UNICAMP, membro do "Comitê de Ética em Pesquisa" da UNICAMP, do "Conselho Curador do Núcleo Softex/Campinas" e do "Conselho Técnico e Científico do SIDI Samsung". É coautor de mais de 90 publicações incluindo capítulos de livros e artigos em periódicos e em anais de eventos científicos. Possui seis pedidos de patente depositados junto ao INPI. Suas áreas de pesquisa incluem: computação gráfica, processamento de imagem, visão computacional e Processamento de Linguagem Natural.

Proposta de um vocabulário bilíngue de festas populares brasileiras baseada em um estudo de corpus

Giovana M. C. Marqueze
Licenciada em Letras pela Universidade Estadual de Londrina e mestranda do Programa de Estudos da Tradução do Departamento de Letras Modernas da Universidade de São Paulo, sob a orientação da profª. drª. Stella E. O. Tagnin. Dos seus oito anos de experiência como comissária de bordo de uma companhia área brasileira, o que possibilitou o contato com as mais variadas manifestações culturais do Brasil, aliados a seu trabalho como legendista de documentários sobre turismo, nasceu o interesse pela pesquisa acadêmica sobre termos ligados à cultura do país.

Frames *de compreensão e* corpora*: estudo de caso com uso do Sketch Engine*

Aline Nardes
Doutoranda e mestra em Linguística Aplicada pela Universidade do Vale do Rio dos Sinos (Unisinos). Licenciada em Letras Português/Inglês (Unisinos-Universidade de Coimbra). Atua como bolsista CAPES/PROSUP no grupo de pesquisa SemanTec (linha de pesquisa Texto, Léxico e Tecnologia). Coeditora da Revista Entrelinhas (Unisinos). Na dissertação de mestrado, estudou conceptualizações de feto anencéfalo no contexto da ADPF 54. Atualmente, investiga conceptualizações ligadas ao aborto em audiências públicas que discutem a Sugestão Legislativa 15/2014. Áreas de interesse: Semântica Cognitiva, Semântica de *Frames*, interfaces entre Linguística Cognitiva e teorias do texto/discurso e Linguística de *Corpus*.

Rove Chishman
Professora titular da Universidade do Vale do Rio dos Sinos (UNISINOS) e Bolsista Produtividade do CNPq. É mestre e doutora em Letras pela PUCRS e licenciada em Letras pela UFRGS. Em 2009, realizou estágio de Pós-Doutorado na University of Texas at Austin (EUA), com projeto na área da Linguística Cognitiva, com ênfase na teoria da Semântica de *Frames*. Tem pesquisa em desenvolvimento na área de semântica lexical computacional, com foco na construção de recursos lexicográficos eletrônicos com base no arcabouço teórico da Semântica de *Frames*. Coordena o grupo de pesquisa SemanTec. Além da Semântica de *Frames*, a pesquisadora também tem interesse em estudos vinculados à Semântica Cognitiva em um sentido amplo.

O estudo do estilo na legendagem: uma pesquisa baseada em corpus

Janailton Mick Vitor da Silva
Bolsista CAPES no Programa de Pós-Graduação em Estudos de Tradução (POSTRAD) na Universidade de Brasília (UnB). Licenciado em Letras - Língua Inglesa pela Universidade Federal de Campina Grande (UFCG). Atualmente, realiza pesquisas dentro dos Estudos da Tradução, mais especificamente em Tradução Audiovisual e nos Estudos da Tradução Baseados em *Corpus* (ETBC).

Alessandra Ramos de Oliveira Harden
Professora do quadro permanente do Departamento de Línguas Estrangeiras e Tradução da Universidade de Brasília desde 1996. Tem experiência na área de Letras, com ênfase em ensino de: tradução (teoria e prática), língua inglesa, redação e leitura (língua inglesa e portuguesa). Atualmente, realiza pesquisa em história da tradução, tradução de textos feministas e tradução audiovisual, com interesse especial em possibilidades de diálogo com o Direito, a História, a Educação e a Filosofia.

Elaboração de um protótipo de glossário bilíngue (português-inglês) de treinamento de força: subsídios para o tradutor

Márcia dos Santos Dornelles
Bacharel em Letras – Tradução (inglês e espanhol) pela Universidade Federal do Rio Grande do Sul, onde também concluiu uma especialização em Estudos Linguísticos do Texto e mestrado em Letras, na linha de pesquisa Lexicografia e Terminologia: relações textuais. É servidora técnico-administrativa da UFRGS, em exercício na Escola de Educação Física, Fisioterapia e Dança (ESEFID). Como autônoma, dedica-se à tradução (inglês e espanhol), especialmente de textos acadêmicos, incluindo diversos livros sobre temas relacionados à área de Educação Física, tais como Condicionamento Físico para a Saúde e Treinamento de Força. Também atua como editora de texto dos *Cadernos do IL*, periódico do Instituto de Letras da UFRGS.

Maria José Bocorny Finatto
Bolsista Produtividade-Pesquisa (PQ) do CNPq desde 2007. Integrante do grupo de pesquisa TERMISUL (UFRGS) desde 1993. Fundadora do Grupo de Pesquisa em Linguística de *Corpus* para região Sul - GELCORP-SUL, em 2010. Docente do Programa de Pós-Graduação em Letras da UFRGS desde 2002. Pós-Doutorada em Ciência da Computação junto o Núcleo Interinstitucional de Linguística Computacional (NILC) do ICMC-USP em 2011. Bolsista Estágio Sênior CAPES junto à Universidade de Évora, Portugal, em 2017. Temas de pesquisa: Acessibilidade Textual e Terminológica em temas de Saúde para Leigos (PQ 2017-2020), Linguística de *Corpus*, Terminologia, Linguística das Linguagens Especializadas baseada em *Corpus*, Processamento da Linguagem Natural (PLN), Lexicologia e Estatística Lexical, Lexicografia, Estudos do Texto, Tradução e Enunciação Científica, padrões do português popular escrito (Projeto PorPopular – www.ufrgs.br/textecc) e Educação a Distância. Desenvolve produtos *on-line* para aprendizes de tradução (http://www.ufrgs.br/textecc/traducao/).

Dicionário Olímpico: a Semântica de Frames *encontra a lexicografia eletrônica*

Rove Chishman
Professora titular da Universidade do Vale do Rio dos Sinos (UNISINOS) e Bolsista Produtividade do CNPq. É mestre e doutora em Letras pela PUCRS e licenciada em Letras pela UFRGS. Em 2009, realizou estágio de Pós-Doutorado na University of Texas at Austin (EUA), com projeto na área da Linguística Cognitiva, com ênfase na teoria da Semântica de *Frames*. Tem pesquisa em desenvolvimento na área de semântica lexical computacional, com foco na construção de recursos lexicográficos eletrônicos com base no arcabouço teórico da Semântica de *Frames*. Coordena o grupo de pesquisa SemanTec. Além da Semântica de *Frames*, a pesquisadora também tem interesse em estudos vinculados à Semântica Cognitiva em um sentido amplo.

Larissa Moreira Brangel
Possui graduação em Letras pela Universidade Federal do Rio Grande do Sul e mestrado e doutorado em Linguística pela mesma instituição. Atualmente, é pesquisadora de pós-doutorado (CAPES/PNPD) na Universidade do Vale do Rio dos Sinos e integra o grupo de pesquisa SemanTec, da mesma universidade. Tem interesse na interface entre Semântica Cognitiva e (meta)Lexicografia, com especial atenção à lexicografia pedagógica.

Sandra de Oliveira
Mestranda em Linguística Aplicada pela Universidade do Vale do Rio dos Sinos (UNISINOS) desde 2017 como bolsista Capes - Prosuc. Estuda Licenciatura - Letras Português e Inglês na Universidade do Vale do Rio dos Sinos desde 2012. Possui graduação em Tecnólogo em Processamento de Dados pela Universidade do Vale do Rio dos Sinos concluída em 1985. É membro do grupo de pesquisa SemanTec - Semântica e Tecnologia, atuando no projeto "Convergências entre Semântica de Frames e lexicografia computacional" sob orientação da profª. drª. Rove Chishman desde 2015. Seu principal interesse de pesquisa é a interface entre a Semântica Cognitiva, com destaque para a Semântica de *Frames*, e a Lexicografia Eletrônica.

Aline Nardes
Doutoranda e mestra em Linguística Aplicada pela Universidade do Vale do Rio dos Sinos (Unisinos). Licenciada em Letras Português/Inglês (Unisinos-Universidade de Coimbra). Atua como bolsista CAPES/PROSUP no grupo de pesquisa SemanTec (linha de pesquisa Texto, Léxico e Tecnologia). Coeditora da Revista Entrelinhas (Unisinos). Na dissertação de mestrado, estudou conceptualizações de feto anencéfalo no contexto da ADPF 54. Atualmente, investiga conceptualizações ligadas ao aborto em audiências públicas que discutem a Sugestão Legislativa 15/2014. Áreas de interesse: Semântica Cognitiva, Semântica de *Frames*, interfaces entre Linguística Cognitiva e teorias do texto/discurso e Linguística de *Corpus*.

Diego Spader
Mestre em Linguística Aplicada pela Universidade do Vale do Rio dos Sinos – UNISINOS e doutorando (CAPES/PROSUC) em Linguística Aplicada pela mesma instituição. Licenciado em Letras Português/Inglês (UNISINOS). Membro do grupo de pesquisa SemanTec - Semântica e Tecnologia (UNISINOS/CNPq), tendo atuado no desenvolvimento dos dicionários *Field: Dicionário de Expressões do Futebol* (CHISHMAN, 2014) e *Dicionário Olímpico* (CHISHMAN, 2016). Tem interesse nos seguintes temas e áreas de pesquisa: Semântica Lexical, Semântica Cognitiva (em especial a Semântica de *Frames*) e Lexicografia Computacional.

Bruna da Silva
Mestra pelo Programa de Pós-graduação em Linguística Aplicada da Universidade do Vale do Rio dos Sinos (UNISINOS). É licenciada em Letras Português pela mesma universidade. Membro do grupo de pesquisa SemanTec – Semântica e Tecnologia, coordenado pela profa. dra. Rove Chishman, atuou no desenvolvimento do projeto Dicionário Olímpico (CHISHMAN, 2016) e, atualmente, desenvolve pesquisa no âmbito do projeto Dicionário Paraolímpico. Tem como interesses de pesquisa a interface entre Semântica Cognitiva e Lexicografia Eletrônica e as reflexões teóricas e práticas sobre Lexicografia Eletrônica. Doutoranda em Linguística Aplicada da UNISINOS (bolsista CAPES – PROSUC), com foco na proposta de Lexicografia Eletrônica Cognitiva.

Colocações especializadas na área do direito comercial internacional e proposta de glossário trilíngue

Jean Michel Pimentel Rocha
Doutorando e mestre em Estudos Linguísticos pela Universidade Estadual Paulista (UNESP), onde também graduou-se em Licenciatura em Letras (Português/Inglês). É membro do grupo de pesquisa "Pedagogia do Léxico, da Tradução e Linguística de *Corpus*", atuando em trabalhos que envolvem Linguística de *Corpus*, Fraseologia (especialmente colocações e colocações especializadas) e Ensino de Inglês como LE.

Adriane Orenha-Ottaiano
Possui Bacharelado em Letras com Habilitação para Tradutor, pela Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP), mestrado em Estudos Linguísticos e Literários em Inglês, pela Universidade de São Paulo (USP), e doutorado em Estudos Linguísticos, pela UNESP. É líder do Grupo de Pesquisa "Pedagogia do Léxico, da Tradução e Linguística de *Corpus*" e Professora Assistente Doutora do Departamento de Letras Modernas, da UNESP, campus de São José do Rio Preto. Atua na Pós-Graduação em "Estudos Linguísticos" da UNESP, nas linhas de pesquisa "Estudos da Tradução" e "Pedagogia do léxico e da tradução a partir de *corpora*". Atua em pesquisas sobre Estudos da Tradução Baseados em *Corpus*, Ensino de Inglês como LE, Linguística de *Corpus* e Fraseologia/Fraseografia (foco em colocações).

O uso de corpus *paralelo e comparável para descrever padrões de uso na tradução de abreviaturas e acrônimos de termos médicos*

Márcia Moura da Silva
Com mestrado e doutorado em Estudos da Tradução pela Universidade Federal de Santa Catarina (UFSC), é professora adjunta no Instituto de Letras da Universidade Federal do Rio Grande do Sul (UFRGS). Entre seus interesses de pesquisa estão formação de tradutor, tradução do texto médico, tradução literária e tradução infantojuvenil. No momento coordena projeto de pesquisa que investiga a tradução de abreviaturas e acrônimos de termos médicos na área da reumatologia. É também pesquisadora em projeto que tem como tema a linguagem do patrimônio cultural brasileiro, com foco na conservação de bens culturais móveis em suporte papel, do grupo Termisul da UFRGS. É atual coordenadora do Núcleo de Estudos da Tradução Olga Fedossejeva (NET) na mesma instituição.

Gabriele Vasconcelos Paparelli
Graduanda do curso de Bacharelado em Letras Português/Inglês pela Universidade Federal do Rio Grande do Sul. Atualmente é bolsista de Iniciação Científica (PROBIC FAPERGS-UFRGS) em projeto de pesquisa que investiga a tradução de abreviaturas e acrônimos de termos médicos na área da reumatologia. Interessa-se pela área da Linguística de *Corpus*, especialmente no estudo da tradução de textos médicos e técnicos em geral.

Identificação de termos no discurso literário de fantasia da série Harry Potter em uma abordagem direcionada por corpus

Raphael Marco Oliveira Carneiro
Professor substituto assistente na área de Língua Inglesa do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU). Doutorando em Linguística e Linguística Aplicada pelo Programa de Pós-Graduação em Estudos Linguísticos (PPGEL) do ILEEL/UFU. Mestre em Linguística e Linguística aplicada pelo PPGEL da mesma instituição. Graduado do curso de Licenciatura Plena em Letras: Habilitação em Inglês e Literaturas de Língua Inglesa pela mesma universidade. Seus interesses de pesquisa incluem estilística; estudos da tradução; fraseologia e paremiologia; lexicologia, lexicografia e terminologia; e Linguística de *Corpus*.

Pommersche Korpora*: um conjunto de* corpora *dialetais da variedade brasileira do pomerano*

Neubiana Silva Veloso Beilke
Doutoranda em Estudos Linguísticos pela UFU/MG. Desenvolve atualmente o projeto: "A descrição do léxico pomerano a partir do *Pommersche Korpora* e uma proposta de ensino-aprendizagem orientado por dados". Mestre em Estudos Linguísticos pelo PPGEL/ILEEL/UFU. Criou o *Pommersche Korpora*, um banco de dados linguísticos em pomerano com um total de 129.666 palavras. Concluiu o projeto de pesquisa intitulado "*Pommersche Korpora:* uma proposta metodológica para compilação de *corpora* dialetais". Pesquisa, de modo geral, o pomerano e, de modo mais específico, a variedade brasileira do pomerano. Integrante dos grupos de pesquisa: Grupo de Pesquisa e Estudos em Linguística de *Corpus* (GPELC); Pesquisas em Léxico (Plex) e Grupo de Pesquisa em Sociogeolinguística (GPS). Consultora e parceira do Projeto Pomerando. Apoia o projeto de tradução do Evangelho de Lucas para o pomerano, conforme o trabalho que está sendo desenvolvido em Canguçu (RS).

Construções de tópico do português brasileiro falado em áreas indígenas

Edivalda Alves Araújo
Possui doutorado em Letras e Linguística pela Universidade Federal da Bahia e mestrado em Letras pela Universidade Federal de Minas Gerais. Atualmente é professora adjunto da Universidade Federal da Bahia, na área de Língua Portuguesa. Realiza pesquisas nas seguintes áreas: sintaxe, diacronia e variação linguística.

Wlianna Silva de Araújo
Graduanda em Letras Vernáculas pela Universidade Federal da Bahia. Atualmente desenvolve pesquisa na área de sintaxe sob perspectiva gerativista, no grupo de Sintaxe Histórica do Programa para a História da Língua Portuguesa (PROHPOR).

Para a segmentação automática de fronteira na fala espontânea a partir de parâmetros prosódicos

Bárbara Helohá Falcão Teixeira
Mestranda em Linguística Teórica e Descritiva, linha de pesquisa Estudos Linguísticos baseados em *corpora* no Programa de Pós-Graduação em Estudos Linguísticos da Universidade Federal de Minas Gerais. Atualmente, desenvolve pesquisas na área de Linguística de *Corpus*, com ênfase em Prosódia. É membro do projeto C-ORAL-BRASIL e integrante do Laboratório de Estudos Empíricos e Experimentais da Linguagem da Faculdade de Letras da Universidade Federal de Minas Gerais.

Plínio Almeida Barbosa
Professor da Universidade Estadual de Campinas. Tem formação em Engenharia Eletrônica e Linguística, com ênfase na área de Fonética experimental, atuando principalmente nos seguintes temas: análise e modelamento dinâmicos da prosódia da fala, prosódia experimental, teoria de sistemas dinâmicos e de osciladores acoplados, ciências da fala e da linguagem. É o autor do "Manual de Fonética Acústica Experimental" (São Paulo: Cortez), juntamente com Sandra Madureira e de "Incursões em torno do ritmo da fala" (Campinas: Pontes).

Tommaso Raso
Professor da Universidade Federal de Minas Gerais. Tem experiência na área de Linguística, tendo atuado principalmente nos seguintes temas: Filologia e Linguística Histórica, italiano, Pragmática, Linguística Textual, estudos da fala e linguística de *corpora*. Atualmente, trabalha principalmente em projetos voltados para a constituição de *corpora* de fala espontânea do PB e a análise da estrutura informacional e das ilocuções com base em *corpora* de fala espontânea. É coordenador do projeto C-ORAL-BRASIL.

Fluência e interação no inglês aeronáutico: uma análise baseada em pragmática e Linguística de Corpus

Malila Carvalho de Almeida Prado
Possui mestrado em Estudos Linguísticos e Literários em Inglês pela Universidade de São Paulo (USP). É doutoranda também em Estudos Linguísticos e Literários na USP, sob supervisão da profa. dra. Stella Tagnin. Estuda a linguagem utilizada por pilotos e controladores de tráfego aéreo em situações não rotineiras. Participa de dois grupos de estudos voltados ao inglês aeronáutico: o Grupo de Estudos de Inglês Aeronáutico, coordenado pela profa. dra. Patrícia Tosqui-Lucks, e o Research Group, da associação internacional ICAEA (International Civil Aviation English Association). Ministra aulas a pilotos da aviação civil há dez anos.

Aos professores, as colocações

Andréa Geroldo dos Santos
Doutoranda em Estudos Linguísticos e Literários pela FFLCH-USP, mestre pela mesma universidade, desenvolve estudos sobre Linguística de *Corpus*, com foco na produção de material didático para o ensino de inglês como segunda língua e educação bilíngue. Experiência de 20 anos em sala de aula (Inglês Geral e para Negócios), atuou como Coordenadora Pedagógica da Mastery Idiomas, professora de inglês na Cultura Inglesa e Editora Pedagógica de Inglês do Sistema Mackenzie de Ensino, onde também coordenou a produção editorial dos livros de Inglês e Português para o Ensino Médio. Atualmente, é Editora de Conteúdo Bilíngue na International School. Autora de capítulo do livro *Corpora* no ensino de línguas estrangeiras (Hub Editorial, 2010) e artigos aprovados em Anais de Congresso Internacional.

Brazilian students' use of English academic vocabulary: an exploratory study

Larissa Goulart da Silva
Professora assistente de português na Universidade do Nebraska, bolsista da Comissão Fulbright. Mestra em English Language Teaching pela Universidade de Warwick com bolsa do Hornby Trust/British Council, teve sua dissertação recomendada para o ELT Masters Dissertation Award do British Council em 2016, pelo qual recebeu Special Commendation. Graduada em Letras – Licenciatura Inglês com láurea acadêmica pela Universidade Federal do Rio Grande do Sul. Professora do programa Idiomas Sem Fronteiras na UFRGS. Bolsista do projeto de pesquisa o impacto do "Programa Nacional do Livro Didático no Cotidiano Escolar da Educação Linguística: Uma proposta de estudo Etnográfico", sob orientação da profa. dra. Simone Sarmento. Revisora do periódico *Bem Legal* e editora do periódico *Brazilian English Language Teaching Journal*.

Marine Laisa Matte
Graduada em Letras – Licenciatura com ênfase em Línguas Portuguesa e Língua Inglesa pela Universidade Federal do Rio Grande do Sul (UFRGS) e mestranda em Estudos da Linguagem, na linha de pesquisa Linguística Aplicada, do Programa de Pós-Graduação em Letras da mesma instituição. Atuou como bolsista de Iniciação Científica (voluntária) no projeto "A formação dos professores para o ensino de inglês para fins acadêmicos no programa Idiomas sem Fronteiras (UFRGS)", coordenado pela profa. dra. Simone Sarmento. Também atuou como professora de língua inglesa do programa Idiomas sem Fronteiras (IsF), ministrando aulas de inglês geral e de inglês para fins acadêmicos.

Simone Sarmento
Professora adjunta da Universidade Federal do Rio Grande do Sul (UFRGS). Possui doutorado em Terminologia e Lexicografia pela UFRGS (2008), mestrado em Language Studies pela University of Lancaster (2005) e mestrado em Linguística Aplicada pela UFRGS (2001). Realizou estágio pós-doutoral na Faculdade de Educação da University of British Columbia. É coordenadora pedagógica do Idiomas Sem Fronteiras-Inglês da UFRGS. Atuou no núcleo estruturante do programa Inglês sem Fronteiras junto à SESU/MEC. Foi vice-presidente de Língua Inglesa do Programa Idiomas sem Fronteiras na SESU/MEC. Seus principais interesses de pesquisa são na área de políticas educacionais linguísticas, políticas de internacionalização das universidades, material didático e formação de professores. Atualmente é coordenadora do projeto de pesquisa "O papel das políticas educacionais linguísticas na internacionalização das universidades: os efeitos do Idiomas sem Fronteiras".

Atividades de compreensão oral com base em corpora *das* TED Talks: *um estudo-piloto*

Luciano Franco da Silva
Mestre em Estudos Linguísticos pela Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP / Ibilce) – Bolsa CAPES, na linha de pesquisa Pedagogia do Léxico e da Tradução Baseada em *Corpora*. É membro do Grupo de Pesquisa Tradução, Terminologia e *Corpora* (UNESP/Ibilce). Tem experiência na área de Letras, com ênfase em Inglês, Ensino de Línguas para Fins Específicos, Inglês para Fins Acadêmicos e Linguística de *Corpus*.

Paula Tavares Pinto
Doutorado em Estudos Linguísticos pela Universidade Estadual Paulista Júlio de Mesquita Filho, tendo realizado estágio na Universidade de Manchester, Inglaterra (PDEE-CAPES). É pesquisadora de grupos cadastrados no CNPq e parecerista de revistas especializadas em Linguística. Atualmente é docente vinculada ao Departamento de Letras Modernas da UNESP, no *campus* de São José do Rio Preto. É coordenadora geral do Programa Idiomas sem Fronteiras na UNESP, do projeto de English Teaching Assistants (CAPES/Fulbright). Atua no Programa de Pós-Graduação em Estudos Linguísticos, nas linhas de Estudos da Tradução e Pedagogia do Léxico e da Tradução baseada em *Corpora*.

Elen Dias
Doutora em Linguística Aplicada pela UNESP. Atualmente, é professora titular na FATEC (Jales) e professora titular e membro do conselho do Curso de Letras da FEF. Tem experiência na área de Linguística Aplicada, atuando principalmente nos seguintes temas: formação de professores de línguas, avaliação, estratégias de ensino, ensino de línguas. Foi vice-presidente da Apliesp - Associação dos Professores de Língua Inglesa do Estado de São Paulo (2007) e membro de sua comissão científica (2007-2011). Atua também coordenadora do Núcleo de Estudos da Linguagem Fatec - Jales (NELF - Jales), além de ser membro integrante do Núcleo Docente Estruturante (NDE) do curso de Sistemas de Informação e Letras da Fundação Educacional de Fernandópolis.

orque é deixar a natureza de desampar

que hão de obrar; porque para não fazer remédio algum, sobre arguir édio algum, sobre arguir

aos parentes, e familiares da casa, porque tinham ouvido dizer, que câmaras sobreparto quisera

parecer que não é bastante,

com febre, câmaras, e faltas d

xidade no

ção

loquial; porque

embaraço,

PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS

INSTITUTO DE LETRAS UFRGS

ISBN 978-85-64522-36-7