BIANCA FRANCO PASQUALINI

LEITURA, TRADUÇÃO E MEDIDAS DE COMPLEXIDADE TEXTUAL EM CONTOS DA LITERATURA PARA LEITORES COM LETRAMENTO BÁSICO

Porto Alegre
2012

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS
ÁREA: ESTUDOS DA LINGUAGEM
LINHA DE PESQUISA: TEORIAS LINGUÍSTICAS DO LÉXICO: RELAÇÕES TEXTUAIS

LEITURA, TRADUÇÃO E MEDIDAS DE COMPLEXIDADE TEXTUAL EM CONTOS DA LITERATURA PARA LEITORES COM LETRAMENTO BÁSICO

BIANCA FRANCO PASQUALINI

ORIENTADORA**: PROFª. DRª. MARIA JOSÉ BOCORNY FINATTO**

Dissertação de Mestrado em Teorias Linguísticas do Léxico, apresentada como requisito para obtenção do título de Mestre pelo Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.

PORTO ALEGRE
2012

AGRADECIMENTOS

À minha mãe, que me apoiou incondicionalmente no percurso de produção desta pesquisa.

Ao meu marido e ao meu filho, pela paciência e amor constantes.

À minha irmã e ao meu sobrinho, pelas discussões inteligentes.

Aos professores do PPG-Letras/UFRGS, pela disposição, pelos conselhos e pelas aulas.

Aos colegas Aline Evers, Juliana Capitani, Janina Antonioli e Fabiano Gonçalves, pela parceria e amizade em todos os momentos.

À CAPES, pelo apoio institucional.

Agradeço especialmente:

À professora Maria José Bocorny Finatto, por apostar nas ideias desenvolvidas nesta dissertação e por acreditar na importância do papel social do linguista.

À Carolina Scarton e a todos os pesquisadores do Núcleo Interinstitucional de Linguística Computacional (NILC/USP), pela contribuição decisiva ao andamento da pesquisa.

nexo

is the

sexo

in the

léxico

- Alexandre Brito

RESUMO

Este trabalho trata dos temas da complexidade textual e de padrões de legibilidade a partir de um enfoque computacional, situando o tema em meio à descrição de textos originais e traduzidos, aproveitando postulados teóricos da Tradutologia, da Linguística de Corpus e do Processamento de Línguas Naturais. Investigou-se a suposição de que há traduções de literatura em língua inglesa produzidas no Brasil que tendem a gerar textos mais complexos do que seus originais, tendo como parâmetro o leitor brasileiro médio, cuja proficiência de leitura situa-se em nível básico. Para testar essa hipótese, processamos, usando as ferramentas Coh-Metrix e Coh-Metrix-Port, um conjunto de contos literários de vários autores em língua inglesa e suas traduções para o português brasileiro, e, como contraste, um conjunto de contos de autores brasileiros publicados na mesma época e suas traduções para o inglês. As ferramentas Coh-Metrix e Coh-Metrix-Port calculam parâmetros de coesão, coerência e inteligibilidade textual em diferentes níveis linguísticos, e as métricas estudadas foram as linguística e gramaticalmente equivalentes entre as duas línguas. Foi realizado também um teste estatístico (*t-Student*), para cada métrica e entre as traduções, para avaliar a diferença entre as médias significativas dentre resultados obtidos. Por fim, são introduzidas tecnologias tipicamente usadas em Linguística Computacional, como a Aprendizagem de Máquina (AM), para o aprofundamento da análise. Os resultados indicam que as traduções para o português produziram textos mais complexos do que seus textos-fonte em algumas das medidas analisadas, e que tais traduções não são adequadas para leitores com nível de letramento básico. Além disso, o índice Flesch de legibilidade mostrou-se como a medida mais discriminante entre textos traduzidos do inglês para o português brasileiro e textos escritos originalmente em português. Conclui-se que é importante: a) revisar equivalências de medidas de complexidade entre o sistema Coh-Metrix para o inglês e para o português; b) propor medidas específicas das línguas estudadas; e c) ampliar os critérios de adequação para além do nível lexical.

Palavras-chave: complexidade textual; leitura; processamento de língua natural

ABSTRACT

This work analyzes textual complexity and readability patterns from a computational perspective, situating the problem through the description of original and translated texts, based on Translation Studies, *Corpus* Linguistics and Natural Language Processing theoretical postulates. We investigated the hypothesis that there are English literature translations made in Brazil that tend to generate more complex texts than their originals, considering – as parameter – the typical Brazilian reader, whose reading skills are at a basic level according to official data. To test this hypothesis, we processed –using the Coh-Metrix and Coh-Metrix-Port tools – a set of literary short stories by various authors in English and their translations into Brazilian Portuguese, and – as contrast – a set of short stories by Brazilian literature authors from the same period and their translations into English. The Coh-Metrix and Coh-Metrix-Port tools calculate cohesion, coherence and textual intelligibility parameters at different linguistic levels, and the metrics studied were the linguistic and grammatical equivalents between the two languages. We also carried out a statistical test (t-test) for each metric, and between translations, to assess whether the difference between the mean results are significant. Finally, we introduced Computational Linguistics methods such as Machine Learning, to improve the results obtained with the mentioned tools. The results indicate that translations into Portuguese are more complex than their source texts in some of the measures analyzed and they are not suitable for readers with basic reading skills. We conclude that it is important to: a) review complexity metrics of equivalence between Coh-Metrix system for English and Portuguese; b) propose specific metrics for the languages studied, and c) expand the criteria of adaptation beyond the lexical level.

Keywords: text readability; reading; natural language processing

SUMÁRIO

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

INTRODUÇÃO

Esta dissertação surgiu a partir da percepção de um fenômeno recorrente durante a minha prática profissional como revisora de traduções literárias. Tal fenômeno não podia ser caracterizado como erro de tradução, tampouco como inconsistência de estilo, mas sim como uma dissonância entre o nível de complexidade do texto original e o nível de complexidade do texto traduzido, tendo em mente o nível de proficiência de leitura dos leitores. Alguns exemplos que ilustram essa impressão são os seguintes:

- Exemplo 1:

TEXTO-FONTE[1]: "I had found the spell of the picture in an absolute life-likeliness of expression, which, at first startling, finally confounded, subdued, and appalled me. With deep and reverent awe I replaced the candelabrum in its former position."

TRADUÇÃO[2]: "Descobrira que o encanto do retrato estava na expressão de uma absoluta aparência de vida que a princípio me espantou para afinal confundir-me, dominar-me e aterrar-me. Com profundo e reverente temor, tornei a pôr o candelabro em sua primitiva posição."

- Exemplo 2:

TEXTO-FONTE[3]**:** Misery is manifold. The wretchedness of earth is multiform. Overreaching the wide horizon as the rainbow, its hues are as various as the hues of that arch – as distinct too, yet as intimately blended. Overreaching the wide horizon as the rainbow! How is it that from beauty I have derived a type of unloveliness?

TRADUÇÃO[4]: A desgraça é variada. O infortúnio da terra é multiforme. Arqueando-se sobre o vasto horizonte como o arco-íris, suas cores são como as deste, variadas, distintas e, contudo, nitidamente misturadas. Arqueando-se sobre o vasto horizonte como o arco-íris! Como de um exemplo de beleza, derivei eu uma imagem de desencanto?

---

[1]POE, Edgar Allan. "The Oval Portrait". (Online) Disponível em < http://poestories.com >. Acesso em 25 de junho de 2010.

[2] POE, Edgar Allan. "O retrato oval". In: *Contos de terror, mistério e morte*. Tradução de Oscar Mendes. 7ª. ed. Rio de Janeiro: Nova Fronteira, 1981.

[3] POE, Edgar Allan. "Berenice". (Online) Disponível em < http://poestories.com >. Acesso em 25 de junho de 2010.

[4] POE, Edgar Allan. "Berenice". In: *Contos de terror, mistério e morte*. Tradução de Oscar Mendes. 7ª. ed. Rio de Janeiro: Nova Fronteira, 1981.

Com vistas a verificar se essa percepção seria verdadeira, passei a realizar uma série de estudos e de pequenos experimentos (a seguir revisados), até que ingressei no mestrado do PPG Letras da UFRGS.

Constituída a investigação, passei a considerar como leitor o leitor brasileiro médio com proficiência de leitura básica, correspondente ao Ensino Fundamental, conforme apontaram, em 2009, pesquisas do Indicador de Alfabetismo Funcional (INAF).

Para a investigação proposta, a hipótese inicial é a de que:

— Textos literários em inglês traduzidos para o português brasileiro tendem a ser mais complexos do que os seus textos-fonte.

Essa ideia de maior dificuldade constitutiva do texto traduzido em relação ao texto-fonte, obviamente associada a uma maior dificuldade de compreensão de leitura para o leitor brasileiro médio, conforme me parece, deveria poder ser percebida de modo especial pelas características e pela variedade do vocabulário empregado no texto. A feição do vocabulário, naturalmente relacionada a outros elementos, conformaria um texto mais complexo e poderia ser investigada a partir de uma abordagem lexical quantitativa do texto. Dessa maneira, a segunda hipótese é a seguinte:

— O índice Flesch (medida lexical de avaliação do nível de complexidade textual oriunda de trabalhos na área de Processamento de Língua Natural) é um recurso importante para um trabalho linguístico de avaliação de complexidade textual.

Considerando a natureza multifacetada dessas hipóteses, é evidente que, para investigá-las, foi preciso recorrer a pesquisas empreendidas em diversas áreas do conhecimento, desde os Estudos de Tradução até estudos de Processamento de Língua Natural, passando pelos Estudos Literários e pela Linguística de Corpus. E essas hipóteses estão acompanhadas das seguintes questões:

— Se textos traduzidos do inglês para o português tenderem a ser mais complexos que os textos-fonte, as traduções do português para o inglês também seriam mais complexas?

— Qual a contribuição de uma comparação entre a complexidade textual de originais e traduções, nos moldes da pesquisa de PLN, para os estudos linguísticos em geral?

Assim, a motivação principal deste trabalho é tentar explicar a dissonância aqui nomeada como "complexidade" ou "inteligibilidade" textual. O objetivo é o de criar subsídios para ajudar o profissional do texto e o tradutor a perceber algumas das sutilezas coesivas envolvidas na construção do texto de chegada e a tomar decisões tradutórias compatíveis com o nível de proficiência de leitura dos leitores a quem as traduções se destinam. Em outras palavras, o objetivo é contribuir para que ele possa adequar o texto ao leitor.

Desde já, é importante frisar que não considero uma tradução voltada para um público leitor específico uma tradução "adaptada". Uma adaptação, segundo a lógica comunicativa e funcionalista da tipologia proposta por Reiss (HURTADO ALBIR, 2008), implicaria mudanças semióticas na tradução da obra (ECO, 2007), o que não é o caso aqui. Além disso, saliento que o objetivo principal deste trabalho não é o de propor um modelo de tradução de textos literários para leitores com baixa proficiência de leitura. O objetivo central desta investigação é, sim, avaliar se o nível de complexidade das traduções de um conjunto de textos selecionados para o português brasileiro seria compatível com o nível de complexidade dos respectivos textos-fonte.

Garantir aos leitores o acesso a textos cujo nível de complexidade e inteligibilidade seja compatível à sua proficiência de leitura é assegurar-lhes o direito a participar da cultura humana de todas as épocas. Compartilhar vivências através da narrativa é um componente fundamental da construção da experiência humana e da vida em sociedade.

A sofisticação da linguagem acompanha a evolução da organização social ao longo do tempo e reflete em si mesma o desenvolvimento do aparato cognitivo humano necessário para o processamento da linguagem. Trazendo unidade e organicidade aos primeiros grupos nômades que se aventuravam pela Terra perseguindo migrações de grandes animais, como búfalos, por exemplo, a linguagem também cumpriu o papel de, através da narrativa de grandes feitos, perpetuar o impermanente. Sentados ao redor do fogo após uma caçada extenuante, caçadores pré-históricos narravam ao grupo suas

estratégias para vencer obstáculos e superar adversidades. A narrativa assume suprema importância no imaginário desses primeiros falantes, pois, por meio dela, a experiência é revivida entre todos. Conforme esses primeiros grupos humanos foram tornando-se menos nômades, as narrativas tornaram-se ligadas também ao local geográfico, não somente aos membros dos grupos sociais em formação (ROGER FISHER, p. 74, 2009). Assim, falar uma determinada língua passa a ser também partilhar da cultura de quem a fala em uma região específica. E o advento da escrita possibilitou o registro perene das narrativas antes transmitidas pela oralidade. Escrita que depende de um leitor com proficiência de leitura compatível à sua complexidade.

Das fogueiras pré-históricas até hoje, a narrativa nunca deixou de preencher a necessidade ancestral de compartilhamento de experiências. Conhecer uma história é fazer parte dela, é entender as motivações por trás dos comportamentos humanos, é exercitar a empatia, é ampliar a experiência humana para além dos limites do indivíduo, é aprender com a experiência do outro. Hoje, mais do que nunca, com a solidificação dos Estados nacionais a partir do século XVIII, a narrativa é essencial para a compreensão mútua entre os povos. E, evidentemente, a tradução é um dos elementos que torna isso possível, sobretudo no que diz respeito às narrativas literárias.

No entanto, traduzir não basta. É preciso levar em conta a contraparte do processo tradutório, que é o leitor a quem as traduções se destinam. E o letramento é um pré-requisito indispensável para que se tenha acesso à narrativa literária – seja ela fruto de tradução ou não. Ler é um direito universal e, para o exercício pleno desse direito, é preciso garantir condições não só de acesso, mas também de *inteligibilidade* dos materiais de leitura. Para tanto, é indispensável levar em conta a proficiência de leitura dos leitores com quem se pretende dialogar.

A narrativa humana extrapola as barreiras do literário, e a ciência ocupa lugar de destaque no imaginário do homem moderno. Nas últimas décadas, presenciamos um desenvolvimento tecnológico sem precedentes na história da humanidade – desenvolvimento cujo impacto se manifesta em absolutamente todas as atividades e ocupações humanas. Desde o surgimento do computador para uso pessoal e da popularização da internet a partir da década de 90, passamos a ter acesso a uma fonte inesgotável de conhecimento e de informação. Testemunhamos os avanços na Genética,

com a conclusão do Projeto Genoma; a descoberta de matéria e energia escuras, que compõem quase 96% do universo; a visita da sonda Opportunity ao planeta Marte; a construção do Grande Colisor de Hádrons, que recentemente divulgou a descoberta de uma partícula mais veloz que a luz; a descoberta de fósseis hominídeos que vêm ajudando a montar o quebra-cabeça da evolução da espécie humana na Terra. Hoje, a curiosidade humana não encontra limites – e a ciência nunca foi tão promissora. A narrativa acompanha toda essa explosão do imaginário humano em todos os gêneros literários. Realidade e imaginário se misturam em cenários futuristas e visionários que instigam a criatividade humana e impõem desafios a serem superados, como, por exemplo, reproduzir em um computador a capacidade humana de comunicar-se através da linguagem.

No epicentro de todos esses avanços e descobertas, um ponto em comum: os avanços na Ciência da Computação. E a Linguística, como não poderia deixar de ser, também vem se transformando em decorrência desses avanços. A Linguística de Corpus, no âmbito dos Estudos da Linguagem, e a Linguística Computacional, no âmbito da Inteligência Artificial, são algumas das áreas mais beneficiadas pelas novas tecnologias, uma vez que o processamento de *corpora* de dimensões gigantescas e o processamento de grandes volumes de informação são, hoje, uma trivialidade.

Evidentemente, falar em narrativa é falar em leitura, e falar em leitura é falar em textos, o que é também falar em leitor. Os pontos de vista a partir dos quais a leitura, o texto e o leitor têm sido abordados, no Brasil e no mundo, envolvem grandes áreas do conhecimento, como a Educação, a Psicologia, a Linguística e a Sociologia, entre outras.

Desenhado esse quadro, vale dizer que esta dissertação vincula-se à Linguística, fazendo incursões breves em áreas correlatas e afins, como o Processamento de Língua Natural, sempre sob a perspectiva dos Estudos da Linguagem, a fim de fazer um recorte bastante específico: o encontro do leitor com o texto literário do gênero conto. Mas quem é esse leitor? O leitor que será o foco desta pesquisa, como mencionado anteriormente, tem nível básico de alfabetização (INAF, 2009). Além disso, a maioria dos leitores brasileiros neste início de século tem mais de 15 anos e pertence às classes C ou D. [5]

[5] Mais adiante, será explicitada essa categoria.

Considerando-se a proficiência de leitura da maioria dos leitores brasileiros, na faixa de letramento considerada básica, surge um impasse: em que medida os projetos editoriais de textos literários preocupam-se com o nível de letramento dos leitores em potencial da obra publicada? Como mensurar o nível de inteligibilidade textual de obras literárias, pensando-se em um público leitor com níveis básicos de alfabetização? E, indo mais além: como definir "inteligibilidade textual"?

Por fim, avaliar os textos em sua complexidade faz-se ainda mais importante à medida que iniciativas recentes do Ministério da Educação do Brasil (MEC) visam popularizar o acesso a clássicos da literatura nacional e internacional para *neoleitores*[6] por meio de versões mais facilitadas dos textos para um primeiro contato de leitores iniciantes, em projetos como o "Leitura para Todos" e "É Só o Começo". Essas políticas públicas brasileiras baseiam-se em dados de pesquisas sobre alfabetismo e leitura, como as realizadas pelo Indicador de Alfabetismo Funcional (INAF), uma parceria entre a organização Ação Educativa e o Instituto Paulo Montenegro, e a pesquisa "Retratos da Leitura no Brasil" (RLB), do Instituto Pró-Livro, em parceria com a Imprensa Oficial do Estado de São Paulo.

Levando tudo isso em conta, neste trabalho os materiais utilizados foram contos literários em inglês traduzidos para o português e contos literários da literatura brasileira traduzidos para o inglês. Evidentemente, é preciso definir o que se considera "conto literário" – o que será aprofundado mais adiante (ver o item 2.2 do Capítulo 2); por ora, basta dizer que consideramos aqui especialmente os cânones literários, ou autores consagrados da literatura mundial e da literatura em língua portuguesa. Dentro desse universo, selecionamos o gênero conto, pois, além de serem textos curtos, contos são narrativas completas (com início, meio e fim), o que facilitou a coleta de um número razoável de textos integrais de vários autores e tradutores para, assim, comparar os textos-fonte às suas traduções para o português brasileiro (ver Capítulo 4, para uma descrição detalhada do *corpus*).

[6] Conforme o MEC: "Jovens com mais de 15 anos e adultos que participam do programa Brasil Alfabetizado em todo o país e nas escolas públicas com turmas de educação de jovens e adultos (EJA)." http://portal.mec.gov.br.

Os textos foram separados em dois grupos, um de contos em inglês e suas traduções para o português, e o outro de contos em português e suas traduções para o inglês. Para fazer um levantamento quantitativo de elementos lexicais dos textos selecionados, algumas ferramentas que calculam índices que avaliam a coesão, a coerência e a dificuldade de compreensão de um texto em diferentes níveis de análise linguística foram usadas (para mais detalhes, ver Capítulo 4). Tais recursos fazem uso de contagens e médias de incidências de determinados itens lexicais, como, por exemplo, negações, conectivos, pronomes pessoais, etc. São elas:

- Coh-Metrix: ferramenta desenvolvida por Graesser e colaboradores (2004), da Universidade de Memphis, a partir de estudos em Linguística Cognitiva. Os contos em inglês – tanto originais quanto traduções – foram processados por este recurso.
- Coh-Metrix-Port: adaptação da ferramenta Coh-Metrix para o português brasileiro, realizada por Carol Scarton (2009), pesquisadora do Núcleo Interinstitucional de Linguística Computacional (Nilc) da Universidade de São Paulo (USP). Os textos em português – tanto originais quanto traduções – foram processados por este recurso.

Nem todas as métricas foram adaptadas para o português, por diversos impedimentos de ordem técnica. Foi preciso, então, listar as métricas cuja equivalência é total entre ambos os recursos. Em seguida, calculou-se a média dos resultados das métricas equivalentes em ambas as ferramentas para os textos com mais de 15 mil caracteres, que foram processados em duas etapas cada um, uma vez que a ferramenta Coh-Metrix limita o processamento a 15 mil caracteres por texto. O segundo passo foi realizar o teste estatístico *t-Student*[7], para cada métrica e entre os grupos de textos, para avaliar se a diferença entre as médias dos resultados obtidos era significativa.

[7] O teste *t-Student* consiste em usar os dados de uma amostra para calcular uma estatística e contrastá-la com a distribuição *t-Student* a fim de determinar a probabilidade de se ter obtido o resultado observado, caso a hipótese nula seja verdadeira. Uma hipótese nula geralmente afirma que não existe relação entre dois fenômenos medidos. Em outras palavras, esse teste é indicado para situações em que os dados provêm de fontes diferentes, para testar se a variação entre eles tem relevância estatística importante, como no caso das métricas calculadas pelas ferramentas Coh-Metrix e Coh-Metrix-Port para textos em português e em inglês. Ver FRIES, Stephan. Useful statistics for corpus linguistics. In SÁNCHEZ, Aquilino; ALMELA, Moisés (eds.). *A mosaic of corpus linguistics: selected approaches*. pp. 269-291. Frankfurt: Peter Lang, 2010.

Após essa etapa e da análise qualitativa dos resultados (ver Capítulo 6), julgamos necessário entender melhor de que forma as métricas e medidas textuais aplicadas pelas ferramentas Coh-Metrix e Coh-Metrix-Port se relacionam entre si de acordo com a natureza dos textos analisados. Ou seja: perguntamos-nos quais as métricas mais características a cada um dos grupos de textos trabalhados. Para responder a essa pergunta, recorremos a técnicas de classificação estatística de Aprendizagem de Máquina (AM). A AM, área vinculada à Inteligência Artificial, desenvolve sistemas automáticos de aquisição e integração de conhecimento a partir de uma base de dados. Neste trabalho, usamos o sistema de classificação em árvore de decisão através do programa Weka:

- Weka: programa desenvolvido por pesquisadores da Universidade de Waikato, na Nova Zelândia. É um programa usado para analisar e fazer predições estatísticas a partir de informações contidas em um banco de dados. Os resultados obtidos das ferramentas Coh-Metrix e Coh-Metrix-Port foram processados com este recurso a fim de aferir as métricas mais características, sob o ponto de vista estatístico, na comparação entre os textos. O agrupamento das métricas foi feito em quatro conjuntos: métricas dos textos em inglês (todos); métricas dos textos em português (todos); métricas dos textos traduzidos (somente as traduções, em ambas as línguas trabalhadas); métricas dos textos originais (somente os textos originais, em ambas as línguas trabalhadas).

Todos os passos da pesquisa já aqui brevemente referidos serão aprofundados ao longo da dissertação. Antes, no entanto, será feito um breve resumo de nossos ensaios de estudo sobre o tema da inteligibilidade de traduções.

TRABALHOS ANTERIORES

Durante os meus anos de formação universitária, ainda como graduanda do curso de Letras da UFRGS, dei início à pesquisa sobre o tema da complexidade textual, aprofundado nesta dissertação. O primeiro trabalho se chamou “Análise de Traduções do Conto ‘O Retrato Oval’, de Edgar Allan Poe, e sua adequação para Leitores do Ensino Médio: Considerações Iniciais”, e foi apresentado no X Salão de Iniciação Científica da PUCRS (PASQUALINI, 2009). A pesquisa concentrou-se especificamente em duas

traduções do conto de Poe, investigando a possibilidade de se definir um perfil linguístico para uma tradução da obra desse autor a partir da determinação da comunidade interpretativa a que o texto traduzido se destina e das estratégias de tradução operadas pelos tradutores. Para isso, foi proposta uma abordagem baseada em *corpus* como instrumental do qual o tradutor poderia se servir a fim de manter o contato entre o texto de partida e a linguagem em uso pela comunidade interpretativa à qual a tradução se dirige, composta, como parâmetro referencial do estudo, por leitores com experiência de leitura compatível à de alunos do Ensino Médio. A observação das duas traduções do conto, em comparação com um *corpus* de língua portuguesa comum, ou seja, de natureza não literária e não especializada, possibilitou sugerir alguns elementos iniciais para a formulação de um projeto de tradução da obra de Poe, buscando, como mencionado anteriormente, traçar um padrão linguístico para futuras traduções da obra do autor para essa comunidade leitora.

A partir das perspectivas levantadas por esse trabalho inicial, um estudo voltado para a questão da complexidade textual foi apresentado no IX Encontro de Linguística de Corpus (ELC) e IV Escola Brasileira de Linguística Computacional (EBRALC): "Medidas de Complexidade Textual entre Traduções Brasileiras e Originais de Literatura Inglesa: Um Estudo-Piloto Baseado em *Corpus*" (PASQUALINI, FINATTO e EVERS, 2010). Esse trabalho colheu indicativos para investigar a seguinte hipótese: há determinadas traduções da literatura inglesa produzidas no Brasil que tendem a gerar textos mais complexos do que os textos-fonte. Assim, empreendemos um estudo quantitativo e qualitativo sobre padrões de vocabulário e de complexidade textual, tomando como *corpus* de estudo um conjunto de cinco contos de Edgar Allan Poe traduzidos para o português brasileiro por diferentes tradutores. Examinamos os textos-fonte e suas respectivas traduções por meio das ferramentas Coh-Metrix e Coh-Metrix-Port, que calculam índices que avaliam a coesão, a coerência e a dificuldade de compreensão de um texto em diferentes níveis de análise linguística. Os resultados do trabalho mostraram que as traduções brasileiras de Poe tenderam a produzir textos mais complexos do que seus textos-fonte no que tange a algumas das medidas analisadas. Além disso, vimos que tanto as traduções da literatura brasileira para o inglês quanto as traduções de textos científicos brasileiros para um padrão de inglês internacional fizeram o percurso inverso: o texto tornou-se menos complexo em função de diferentes fatores.

Esse foi o experimento inicial dos sistemas Coh-Metrix para comparar textos-fonte em inglês e suas traduções para o português e vice-versa. Os *corpora*, entretanto, eram pequenos.

De posse desses resultados, e confiantes de que valia a pena investir no uso das ferramentas Coh-Metrix e Coh-Metrix-Port, partimos para um terceiro estudo, apresentado no *8th Brazilian Symposium in Information and Human Language Technology* (STIL), em 2010, intitulado "Comparando Avaliações de Inteligibilidade Textual entre Originais e Traduções de Textos Literários" (PASQUALINI, SCARTON e FINATTO, 2011). O objetivo do trabalho, expandindo o anteriormente citado, foi examinar, contrastivamente, a inteligibilidade textual entre 28 contos de autores variados da literatura de língua inglesa e portuguesa do Brasil produzidos entre 1830 e 1940 e suas traduções, na direção inglês-português e português-inglês. Para a análise, foram utilizados os sistemas Coh-Metrix em inglês e português, com preferência para textos curtos, sem ultrapassar 30 mil caracteres cada (dada a limitação de processamento das ferramentas). As métricas estudadas dos dois sistemas foram as que, a princípio, fossem linguística e gramaticalmente equivalentes entre as duas línguas. Foi realizado o teste estatístico *t-Student*, para cada métrica e entre as traduções português-inglês e inglês-português, para avaliar se a diferença entre as médias dos resultados obtidos mostravam-se significativamente diferentes. Concluímos, nesse trabalho, que associar métricas de inteligibilidade textual entre originais e traduções pode render importantes *insights* tanto para estudos de PLN quanto para estudos linguísticos em geral, incluindo estudos de Tradutologia e estudos de Linguística de Corpus.

Foi a partir do trabalho de 2010, recém-descrito, que a pesquisa apresentada nesta dissertação foi construída. Uma vez que uma suposição de que textos literários em inglês traduzidos para o português brasileiro tendiam a ser mais complexos do que os seus textos-fonte foi parcialmente confirmada nos trabalhos anteriormente mencionados, nesta pesquisa o *corpus* foi mantido o mesmo, e as ferramentas Coh-Metrix e Coh-Metrix-Port também foram utilizadas, com aprofundamento da análise qualitativa dos resultados.

No entanto, a fim de (1) expandir a abordagem quantitativa aos resultados obtidos, e (2) definir quais as métricas mais características de cada agrupamento textual conforme a língua (inglês e português) e a natureza do texto (originais ou traduções),

acrescentamos a ferramenta Weka. A decisão por uma abordagem estatística como a proposta aqui teve também o objetivo de investigar o papel do índice Flesch (a seguir detalhado, no Capítulo 2) nos diferentes grupos de texto.

ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação está organizada em oito capítulos, com o acréscimo de quatro anexos. Os Capítulos 1 e 2 apresentam uma breve revisão da literatura, a qual foi dividida em duas partes. Na primeira, são abordados os pressupostos teóricos e os conceitos de língua e linguagem em que a pesquisa se baseia, com foco nos Estudos da Linguagem, Estudos do Léxico, Linguística de Corpus e Linguística Computacional, com destaque para a área de Aprendizagem de Máquina (AM), à qual nos acercamos não sem grandes esforços para um linguista sem qualquer formação específica em Estatística Avançada.

A segunda parte, no Capítulo 2, apresenta pressupostos teóricos que abrangem, de forma não ortodoxa, conceitos relativos à Leitura, de acordo com a visão da Crítica Literária, dos Estudos de Tradução e da Linguística Textual. Contudo, a revisão bibliográfica centrou-se apenas naquilo que esses campos do saber expressam sobre *Leitura*, sem fazer incursões em matérias não relacionadas a isso. A discussão sobre um assunto heterogêneo e multifacetado, como é o tema da Leitura, poderia ser enriquecida por conceitos de áreas como, por exemplo, a Educação e a Linguística Cognitiva. Entretanto, é preciso delimitar o alcance e reconhecer as limitações inerentes a uma dissertação de Mestrado como esta.

O Capítulo 3 discorre sobre o posicionamento do trabalho, apresentando uma síntese dos conceitos expostos nos capítulos 1 e 2, em que se retomam, complementam e reformulam algumas das ideias comentadas nos capítulos de revisão da literatura.

No Capítulo 4, apresentam-se as ferramentas usadas – Coh-Metrix, Coh-Metrix-Port e Weka – e o *corpus* de trabalho. Discorro sobre o processo de compilação dos textos, os problemas encontrados e as decisões tomadas, nem sempre fáceis.

No Capítulo 5 é feita a descrição dos procedimentos a serem empregados para a posterior análise dos resultados.

O Capítulo 6 apresenta os resultados e a descrição dos dados obtidos. A primeira parte é dedicada aos dados das ferramentas Coh-Metrix e Coh-Metrix-Port, e a segunda aos resultados obtidos com o Weka.

A partir dos resultados descritos no capítulo anterior, tem-se, no Capítulo 7 uma análise dos resultados, discutindo e problematizando algumas questões.

No Capítulo 8 são tecidas as considerações finais, em que se retomam as hipóteses do trabalho e em que são apontadas as perspectivas futuras ensejadas pela pesquisa.

Há também quatro seções anexas, em que estão incluídos: a lista de conectivos em inglês e em português usada pelas ferramentas Coh-Metrix e Coh-Metrix-Port (Anexo A); as tabelas com as estatísticas das métricas analisadas, bem como as tabelas completas, inclusive com as métricas que não foram investigadas (Anexo B), para eventuais consultas de quem assim desejar; e os arquivos ARFF, usados no sistema Weka, que poderão ser úteis a quem se interessar em usar esse recurso, seja para conhecê-lo, seja para aprender a usá-lo (Anexo C); as tabelas completas com os resultados da análise de cada um dos 28 textos obtidos através do processamento nas ferramentas Coh-Metrix e Coh-Metrix-Port.

1. REVISÃO DA LITERATURA (I)

Nesta primeira parte, é feita uma breve revisão de postulados teóricos linguísticos relevantes ao trabalho, partindo da concepção de linguagem e língua de Ferdinand de Saussure (1857-1913), considerado o fundador da Linguística Moderna, e da releitura desses conceitos feita por Simon Bouquet. Isso é feito para contextualizar, no âmbito dos Estudos da Linguagem, a nossa base de entendimento para conceitos conexos como os de tradução, leitura, etc. Assim, o objetivo é situar nossa perspectiva linguística para diálogo com áreas como o PLN – uma perspectiva radicalmente diferente da tradicional dialogante com os colegas da Ciência da Computação, associada ao gerativismo chomskyano.

Em seguida, é traçado um breve panorama dos estudos do léxico em Linguística de Corpus, Análise Multidimensional, Linguística Computacional e Processamento de Língua Natural, a fim de apresentar as propostas teóricas que embasam o trabalho.

Antes de prosseguir, é importante deixar claro desde já que esta primeira etapa da revisão teórica, apesar de sucinta, tem também o objetivo de situar o leitor cuja área de especialidade não seja a Linguística. Eis o motivo por que a discussão começa com Saussure e sua obra. Além disso, para que as propostas de Bouquet sejam entendidas, é preciso, antes, apresentar a obra saussuriana.

A revisão teórica das outras áreas mencionadas – Estudos do Léxico em Linguística de Corpus, Análise Multidimensional, Linguística Computacional e Processamento de Língua Natural – apontam para o caminho teórico que sustenta esta dissertação.

1.1 CONCEPÇÃO DE LINGUAGEM E LÍNGUA

1.1.1 SAUSSURE: UM PROJETO EPISTEMOLÓGICO PARA A LINGUÍSTICA

Com a publicação do *Curso de Linguística Geral* (doravante CLG), em 1916, os estudos da linguagem passaram por uma transformação radical e começaram a deixar de

concentrar-se somente na gramática, na filologia e na comparação entre línguas. O CLG foi compilado por três organizadores (Charles Bally e Albert Sechehaye, com colaboração de Albert Riedlinger) a partir dos cadernos de alunos de Saussure – uma empreitada que, sob retrospecto, "inaugurou" a Linguística tal como a conhecemos hoje. Desde a primeira edição, o CLG passou por reedições, edições críticas e releituras, num esforço exegético contínuo em busca de uma voz autoral inquestionável e reputável única e exclusivamente a Saussure. Contudo, ainda que se possa questionar a autoria do CLG, pois os organizadores sequer foram alunos de Saussure, as ideias apresentadas no livro conformaram o alicerce da Linguística, ciência incipiente na época.

Sobre essa polêmica, Claudine Normand (2009, p. 15) afirma que "o CLG é um texto chamado Saussure", sugerindo, com isso, que deixemos de lado a discussão sobre a autenticidade da obra para que possamos refletir sobre o conteúdo do CLG "tal como foi publicado em 1916". E será a este Saussure, num primeiro momento, a quem se referirá aqui, sobretudo para destacar a visão de linguagem e língua a partir da qual esta dissertação se fundamenta. No entanto, a trajetória do CLG e do pensamento saussuriano também é relevante à discussão sobre as leituras possíveis de um texto conforme a comunidade leitora desse texto se modifica ao longo do tempo e do espaço. Evidentemente, o *corpus* saussuriano, ainda que esparso e em larga medida apócrifo, é muito extenso para ser discutido aqui em detalhes[8], mas há que se ressaltar a opinião de vários autores e estudiosos da obra saussuriana, como, por exemplo, Gadet (1996), Trabant (2005) e Depecker (2009), de que a busca por uma "autenticidade" ou uma "univocidade" é o menos importante quando o assunto é Saussure. O conjunto da obra de Saussure impõe-se como fundamental para os Estudos da Linguagem, mesmo que não se possa tomar o CLG como uma obra fechada e acabada. Em outras palavras, isso equivale a dizer que não existe um "pensamento de Saussure", mas "pensamentos sobre Saussure".

[8] Além do *Cours*, o corpus saussuriano, conforme Bouquet (2004) e Depecker (2009), é constituído pelas publicações de Saussure (alguns artigos acadêmicos e uma tese de doutorado), pelos manuscritos de conferências proferidas em 1891, por um conjunto de notas sem data, por cartas, pelas notas preparatórias aos cursos de linguística geral de 1907 a 1911, pelos cadernos dos alunos (fontes usadas tanto para a redação do *Cours* original quanto para futuras edições críticas) e, enfim, por escritos encontrados em 1996 e publicados em 2002 sob o título *Écrits de Linguistique Générale* (a primeira edição brasileira é de 2004, sob o título *Escritos de Linguística Geral*.

1.1.2 O CURSO DE LINGUÍSTICA GERAL

O CLG é dividido em seis partes: introdução e cinco capítulos, os quais se subdividem em Princípios Gerais, Linguística Sincrônica, Linguística Diacrônica, Linguística Geográfica e Linguística Retrospectiva. Logo nas primeiras páginas da introdução (p. 13), após um breve histórico sobre a evolução dos estudos da linguagem, Saussure preocupa-se em definir a matéria da Linguística: "A matéria da Linguística é constituída inicialmente por todas as manifestações da linguagem humana, (...) não só a linguagem correta e a 'bela linguagem', mas todas as formas de expressão". E acrescenta, logo em seguida: "Como a linguagem escapa (...) à observação, o linguista deverá ter em conta os textos escritos, pois somente eles lhe farão conhecer os idiomas passados ou distantes". Como tarefa, a Linguística tem a descrição das línguas em termos sincrônicos e diacrônicos; a dedução de universalidades e leis particulares das línguas; e a delimitação de si própria como ciência a partir da definição do seu objeto, que é a língua. Para Saussure, é mais importante tomar a língua, que é manifestação da faculdade humana da linguagem, como objeto do que abordar a linguagem de forma abstrata:

> Mas o que é a língua? Para nós, ela não se confunde com a linguagem; é somente uma parte determinada, essencial dela (...). É (...) um produto social da faculdade de linguagem e um conjunto de convenções necessárias, adotadas pelo corpo social para permitir o exercício dessa faculdade nos indivíduos. [...] Não se deixa classificar em nenhuma categoria de fatos humanos, pois não se sabe como inferir sua unidade. A língua, ao contrário, é um todo por si e um princípio de classificação. (SAUSSURE, 2006, p. 17)

Assim, Saussure concebe a dimensão epistemológica do objeto da Linguística, deixando questões relacionadas ao método nas mãos do observador ao afirmar que "é o ponto de vista que cria o objeto" (p. 15), ou seja, é a perspectiva sobre o fato linguístico que determinará o método por meio do qual ele será investigado. A primeira característica desse objeto multifacetado é a sua natureza sistêmica: a língua é um sistema de signos que estabelecem relações de oposição e diferença – um signo é o que os outros não são ("na língua só existem diferenças" [SAUSSURE, 2006, p.139]). Em outras palavras, trata-se de um sistema de valores composto somente por termos

complexos que só fazem sentido quando em oposição e em relação aos outros elementos do sistema. É no jogo de relações intrassistêmicas que os valores se estabelecem.

Como anteriormente mencionado, a autora Claudine Normand considera o CLG "um texto chamado Saussure". Para ela, não se trata de ignorar os esforços filológicos sobre o processo de criação do livro, "mas de resguardar-lhes seu papel de complemento (...), recusando que eles sejam obstáculos a uma primeira reflexão sobre o *Cours* como texto" (id., p. 18). Ao escrever o capítulo "System, arbitrariness, value" para o livro *The Cambridge Companion to Saussure*, publicado pela Cambridge University Press em 2006, Normand, cuja língua materna é o francês, vê-se diante do impasse da tradução de *langue* para o inglês:

> To start with there is the problem of translating "la langue est un système" into English. (…) More problematic in English, however, is the term "language" itself since it does not differentiate, as Saussure did, between *le langage*, *une langue*, *la parole* and *la langue*. For Saussure *le langage* refers to the general human faculty of language. *Une langue* refers to any particular language (…). *La parole* refers to a particular utterance, to an example of individual speech (…). *La langue*, however, is a new technical term developed by Saussure, and is the essential object of his investigations. (2006, p. 89, itálicos da autora)[9]

Normand, partindo do pressuposto de que o objeto de investigação de Saussure é a língua, desvia-se do caminho trilhado por leituras estruturalistas do CLG ao ampliar a teoria saussuriana para além do signo. É o signo, e não o sistema aberto de relações de valor entre signos linguísticos, o elemento cuja definição proposta por Saussure é comumente apontado como o divisor de águas entre a linguística que se fazia até então e a linguística pós-CLG. Essas relações de valor estabelecem-se por meio de oposições, daí a afirmação de que "na língua só existem diferenças". Segundo a visão de Normand, Saussure considerava tarefa do linguista descrever as operações realizadas pelo falante ao

[9] Para começar, há o problema de traduzir "la langue est un système" para o inglês. (...) Ainda mais problemático em inglês, no entanto, é o próprio termo "language", uma vez que ele não diferencia, como fez Saussure, entre *le langage*, *une langue*, *la parole* e *la langue*. Para Saussure, *le langage* se refere à faculdade humana da linguagem como um todo. *Une langue* se refere a qualquer língua em particular (...). *La parole* se refere a uma elocução em particular, a um exemplo do discurso individual (...). *La langue*, contudo, é um novo termo técnico desenvolvido por Saussure, e é o objeto essencial de suas investigações. (Tradução minha.)

usar o sistema da língua, uma ideia em total oposição à pesquisa linguística realizada na época, fundamentalmente histórica. "Os princípios de uma teoria geral da linguagem precisam ser entendidos como um conjunto de traços abstratos a partir dos quais uma descrição correta e acessível se torna viável", prossegue Normand (2006, p. 92).

Dentre tais princípios, os seguintes constituem os fundamentos epistemológicos de Saussure: **a língua é um *sistema*; a língua é um *fato social*, enquanto a fala é um *ato individual*; os signos são *arbitrários* e devem ser tomados como *valores*; e a linguística pertence a uma ciência mais geral a ser desenvolvida (a semiologia)**.

A autora acrescenta que, na época, as concepções positivistas e empiristas dominantes nas ciências emergentes impossibilitaram a compreensão plena das propostas de Saussure. Rompendo com o extralinguístico – preocupação dos filósofos da linguagem –, Saussure assume uma posição peculiar entre os pensadores de seu tempo (id., p. 104).

Por fim, é importante salientar também que Saussure faz inúmeras referências ao processo tradutório, ainda que camufladas e indiretas. As mais expressivas estão, como não poderia deixar de ser, no capítulo "O valor linguístico" (2004, p. 130). O exemplo da comparação entre "*mouton*" e "*sheep*" é uma delas. Ainda que tenham a mesma significação, essas duas palavras têm valores diferentes, e tais valores são determinados por aquilo que elas *não são*, ou seja, pelas palavras que as rodeiam e pela identidade que cada palavra assume através de sua relação com as outras. "Se as palavras estivessem encarregadas de representar os conceitos de antemão, cada uma delas teria, de uma língua para outra, correspondentes exatos para o sentido" (SAUSSURE, 2004, p. 135), e todo tradutor sabe que não é assim. A tradução se opera no nível do *sentido*, e não no nível do sistema.

1.1.3 SIMON BOUQUET: A TRANSVERSALIDADE SEMÂNTICA E A LÍNGUA COMO ÁLGEBRA

Simon Bouquet, filósofo francês e estudioso da obra saussuriana, tem um ponto de vista divergente e mais radical que o de Normand. Embora Bouquet ressalte a importância da publicação do *Cours* como uma "síntese magistral da reflexão saussuriana", declara também que a obra é "um reflexo deformado do pensamento que

pretende divulgar" (2004, p. 13). A principal crítica do autor a Charles Bally e Albert Sechehaye, redatores do *Cours*, é a de terem organizado o livro "segundo a lógica de um sistema acabado" (id., p. 13). A proposta de Bouquet não se restringe a uma leitura do *Cours* "assessorada" ou "confirmada" pelos manuscritos. Na verdade, ele usa os manuscritos como uma contraprova da deformação do pensamento "original" e "autêntico" de Saussure, pensamento que Bouquet afirma inacabado e ao qual deu continuidade. Bouquet faz uma leitura do *Cours* a partir do *corpus* completo de textos saussurianos – incluindo em especial os documentos manuscritos encontrados em 1996 – e propõe uma gramática do sentido tomando a língua como álgebra (BOUQUET, 2004).

Essa gramática do sentido, segundo a leitura do filósofo, fundamenta-se na teoria da arbitrariedade do signo linguístico e na teoria do valor apresentadas por Saussure, concebidas em analogia com as leis da física: "opondo o princípio de movimento ao princípio de inércia" (idem, p. 220). O princípio de movimento refere-se à diacronia, às forças de transformação de uma língua ao longo do tempo; e o princípio de inércia refere-se à sincronia ou a um determinado estado de língua.

A definição do signo linguístico como "objeto de natureza concreta embora puramente espiritual" está contida na seguinte "glosa" do pensamento saussuriano, de acordo com Bouquet: "A unidade do sentido, contravalor da unidade fonológica, é, como esta última, uma unidade perfeitamente concreta". Caracterizando-se o signo, têm-se presentes e de modo concreto o fato psicológico e o fato fonológico, e, assim, "o concreto fonológico é a garantia do concreto semântico".

Bouquet (idem, p. 243) chama a atenção para um equívoco de Bally e Sechehaye, que não levaram em conta uma autocrítica de Saussure sobre a metáfora da folha de papel: "A língua é comparável a uma linha cujos elementos são cortados, e não *recortados* cada um com uma forma" (frase retirada, por Bouquet, das fontes manuscritas[10]). Em seguida, aponta que a divisibilidade fonológica em unidades discretas implica divisibilidade semântica e cita uma nota de Saussure, que diz: "O apossema é o

[10] No CLG, página 131: "A língua é também comparável a uma folha de papel: o pensamento é o anverso e o som o verso; não se pode cortar um sem cortar, ao mesmo tempo, o outro".

envoltório vocal do sema", apossema tido aqui como componente fonológico e invólucro do componente semântico.[11]

No entanto, para Bouquet, Saussure insiste "no fato de que a noção de 'unidade' é equivalente às de 'identidade', 'valor', 'realidade' ou de 'elemento concreto'" (p. 243). A partir disso, oferece uma definição de sentido: "é sentido o que é, tanto quanto a unidade fonológica, uma unidade concreta".

O termo *signo*, assim, "torna-se apto a refletir a transversalidade do fato semântico" (2004, p. 244). Bouquet aponta que vários aforismos de Saussure revelam a confusão entre unidade e o próprio fato semântico. O sentido é colocado como o que é traduzido e criado pela unidade. Fora da relação com o objeto fonológico, que o contém, o objeto semântico não existe. No entanto, essas considerações levantam um problema aparentemente paradoxal no que diz respeito à sintaxe ou aos fatos semânticos qualificados "de entidades abstratas da língua". Sobre isso, o autor afirma: "são as entidades sintáticas que são consideradas como entidades abstratas" (2004, p. 246).

Em uma nota que ocupa uma página e meia, Bouquet, contudo, faz digressões sobre a imprecisão terminológica do que vem a ser "palavra", "unidade" e "subunidade" no CLG e em todas as fontes originais de Saussure para concluir que, "se consideradas em relação às palavras, essas outras unidades podem ser chamadas de 'subunidades'; elas são, no entanto, do ponto de vista do valor semântico [...] *unidades*" (2004, p. 247). Assim, vê-se que, para Bouquet, o valor semântico é uma constante – seja no fato linguístico, na enunciação propriamente dita, na realização da língua no discurso, seja na língua como potencialidade na mente do falante. Aquilo que se fragmenta, nos estudos da linguagem, em morfologia, sintaxe, fonologia e semântica é, segundo Bouquet, indissociável, pois a semântica é transversal a todo e qualquer fenômeno linguístico. Desse modo, **uma abordagem lexicológica à língua sempre será uma abordagem também ao *sentido***, que é a alma da linguagem.

Em seguida, Bouquet chama a atenção para a tese de Saussure "segundo a qual a sintaxe é uma *sintagmática*, uma teoria de marcas concretas, fonológicas ou posicionais"

[11] Nos *Escritos de Linguística Geral* (SAUSSURE, 2004), página 94: "O apossema é o envoltório vocal do sema. E não o envoltório de uma significação. O sema não existe apenas por fonismo e significação, mas por correlação com outros semas."

e esclarece, citando as palavras do próprio CLG "revisitado", a aparente contradição do "agenciamento" de marcas semânticas *concretas* por unidades sintáticas *abstratas*:

> [...] Reservamos o termo *concreto* [para] o caso em que a ideia tem diretamente seu apoio numa unidade sonora. Sendo que *abstrato* tem seu apoio indiretamente, através de uma operação de sujeitos falantes. (2004, p. 248)

Bouquet antecipa o possível argumento de que não haveria diferença entre o fenômeno e as unidades, que o mais importante, no que diz respeito aos fenômenos semânticos, seriam as unidades concretas no sentido estrito. Porém, na visão de Bouquet, os textos originais respondem a essa objeção:

> "Na língua, assim como em outros sistemas semiológicos, não pode haver diferença entre o que caracteriza uma coisa e o que a constitui." (p. 249)[12]

> "Para o fato linguístico, *elemento* e *caráter* são eternamente a mesma coisa." (p. 249)[13]

Conforme o raciocínio de Bouquet, a existência de unidades semânticas abstratas é um aspecto fundamental da teoria que coloca essa natureza concreta, ou seja, a tese segundo a qual a globalidade do sentido é um fato homogêneo e concreto. E afirma:

> No plano da epistemologia programática da linguística, o axioma da transversalidade e da homogeneidade semânticas [...] põe em questão as divisões tradicionais da lexicologia, da morfologia e da sintaxe. (...) Da reivindicação de uma gramática global – transversal aos domínios da lexicologia, da morfologia e da sintaxe – não decorre que falte questionar por princípio todas as categorias tradicionais: o que importa [...] é considerar o fato semântico como algo que implica [...] uma abordagem global. (2004, p. 252)

Nessa concepção, na transversalidade da gramática do sentido está implicada a transversalidade de um "valor" semântico. O princípio do que Bouquet denomina uma "nova gramática geral" é o da generalidade do específico, segundo o qual o "sentido" é "um objeto construído pelo sistema da língua, um objeto específico a uma língua específica". A esse objeto corresponde a teoria do valor. Além disso, afirma que "a

[12] Ver também *Escritos de Linguística Geral*, p. 109 (SAUSSURE, 2004).

[13] Ver também *Escritos de Linguística Geral*, p. 224 (SAUSSURE, 2004).

língua, em sua face semântica, não tem nenhuma propriedade geral senão a de ser uma álgebra" (2004, p. 287). Isso significa dizer que **as unidades linguísticas, por si só, são nulas e que é somente no sistema total de uma determinada língua que passam a ter valor**.

Subdividindo os valores linguísticos em valores *in praesentia* (relações sintagmáticas) e valores *in absentia* (relações associativas ou paradigmáticas), ele propõe diversos graus de arbitrariedade entre os signos conforme as relações de valor que estabelecem entre si (ver Figura 1). Na primeira categoria de valores *in absentia* (primeiro grau de arbitrariedade) estão:

- a arbitrariedade da ligação de um significante a um significado; a arbitrariedade da ligação entre um significado e um significante; e a arbitrariedade da constituição do signo através da ligação entre significante e significado.

Na segunda categoria (ou no segundo grau de arbitrariedade) estão:

- a arbitrariedade do sistema fonológico; e a arbitrariedade do sistema semântico.

A última categoria proposta pelo autor é o valor *in praesentia*, que engloba o caráter linear do fato sintático, ou seja, a produção linguística ou enunciação dos valores *in absentia*. A conjugação dos valores internos e sistêmicos do signo forma o valor *in absentia*, o qual é, em essência, o sistema linguístico. O valor *in praesentia*, para Bouquet, "abrange tudo o que a sintaxe estuda na linguagem – sendo que a palavra sintaxe deve ser entendida no sentido mais geral de uma teoria do fato sintagmático" (2004, p. 269), e é, trocando em miúdos, a produção linguística. De acordo com Bouquet, esse conceito foi mal compreendido pelos redatores do CLG e ignorado pelas leituras estruturalistas da obra de Saussure. Assim, conclui que o valor semântico é dado a partir da integralidade da combinação entre valores *in absentia* e *in praesentia*, formando um todo coeso e harmônico no uso da língua pelos falantes.

A visão de Bouquet, ainda que questionável em alguns pontos, sobretudo no que se refere a suas afirmações categóricas a respeito das intenções de Saussure, tem implicações relevantes para os estudos da linguagem na medida em que apresenta uma

semântica cuja natureza fundamental é a transversalidade sistêmica. Com isso, Bouquet propõe uma noção de semântica vinculada às relações entre os valores internos e sistêmicos da língua, contrapondo-se a uma visão de semântica como estudo de significados universalizados, desconectados de suas inter-relações intrassistêmicas. A Figura 1, adiante, apresenta um esquema das propostas de Bouquet de acordo com os postulados saussurianos.

Outra implicação importante das propostas do autor é a possibilidade de se pensar a semântica em termos algébricos (ou seja, uma perspectiva *quantificável* da língua), um grande desafio para alguns linguistas dedicados ao processamento computacional da linguagem. Além disso, uma das críticas que avaliações de complexidade e inteligibilidade textual baseadas em contagens lexicais recebe é a de ser superficial. No entanto, considerando a noção de transversalidade semântica proposta por Bouquet, essa superficialidade é apenas aparente, pois jamais um signo estará desprovido de sentido, seja *in praesentia* ou *in absentia*.

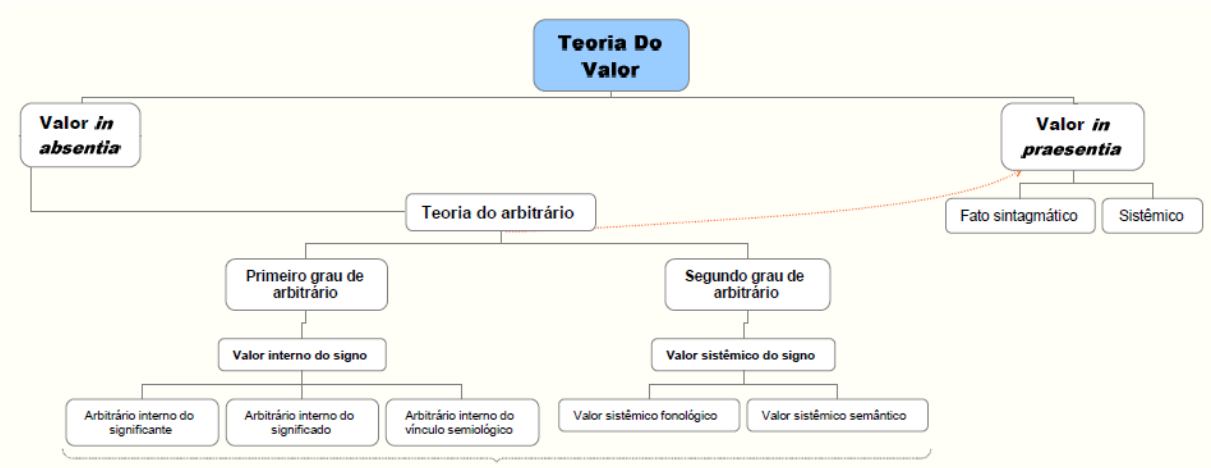

Figura 1. A teoria do valor saussuriana, segundo Simon Bouquet (2004).

Concluindo esta longa seção, que pode parecer bastante penosa para pesquisadores de PLN/Ciência da Computação, importa dizer, a título de síntese, que as ideias de Saussure sobre a língua e seu funcionamento e a perspectiva de Bouquet (para quem a semântica não existe sozinha e independentemente da sintaxe, da morfologia e da fonologia, mas, sim, constitui um elemento que atravessa essas dimensões), somadas, são catalisadoras para um enfoque formal diferenciado da língua – bastante afastado da

tradição gerativista. Esse atravessamento da semântica, conforme entendo, evoca a figura de uma teia de relações num sistema em que todos os elementos são complexos.

1.2 ESTUDOS DO LÉXICO EM LINGUÍSTICA DE CORPUS

Conforme a definição de Berber Sardinha, "a Linguística de Corpus ocupa-se da coleta e exploração de corpora, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística" (2004, p. 3). Dito de outra forma, a Linguística de Corpus (LC) pode ser considerada uma abordagem empirista que parte de uma perspectiva probabilística da linguagem. Além disso, através da análise de *corpora* e grandes coleções de textos, evidencia o uso colocacional da língua, em que a ordenação lexical não se dá de forma aleatória, sustentando a teoria saussuriana do valor linguístico.

Apesar de os avanços tecnológicos terem favorecido a coleta e a manipulação de *corpora*, os primeiros levantamentos e estudos nessa área, ao contrário do que se imagina, tiveram início na década de 20 do século passado, com Thorndike, que identificou as palavras mais frequentes da língua inglesa (BERBER SARDINHA, 2004).

A explosão tecnológica das últimas décadas consolidou e validou o status dos estudos em LC (TOGNINI BONELLI, 2010). Ao permitir a manipulação de grandes volumes de material linguístico, o uso do computador também permitiu a aplicação de metodologias sofisticadas de interpretação dos dados coletados pelo linguista. Tognini Bonelli faz o seguinte comentário sobre a particularidade do objeto e do método da LC (2010, p. 18):

> O que presenciamos no desenvolvimento da Linguística de Corpus como disciplina é que o nosso ponto de vista metodológico progressivamente passou a determinar tanto o objeto quanto o objetivo da pesquisa. Em outras palavras, nesse caso, a metodologia acabou definindo o domínio da disciplina.[14]

[14] Em inglês: What we have witnessed in the development of corpus linguistics as a discipline is that our chosen methodological standpoint has progressively determined both the object and the aim of the enquiry. In other words, in this instance, the methodology has ended up defining the domain of the discipline.

A quantidade de informação linguística coletada e disponível atualmente é tamanha que é possível perceber padrões imperceptíveis sem a tecnologia a que temos acesso hoje. O desafio para o linguista disposto a trabalhar com corpora, hoje, é "elaborar uma metodologia confiável para descrever"[15] toda essa informação (TOGNINI BONELLI, 2010). Sobre isso, Teixeira (2008, p. 154) afirma que "não se trata apenas de ter à disposição uma maior quantidade de dados linguísticos. As ferramentas computadorizadas de análise textual permitem observá-los de outra perspectiva, evidenciando novos fenômenos."

Por *corpus*, é bom lembrar, entende-se o conjunto suficientemente extenso de informações linguísticas (de origem oral ou escrita) que, submetido a critérios de sistematização, represente o uso da língua (total ou compartimentado) de uma determinada comunidade linguística em um determinado período (BERBER SARDINHA, 2004). As amostragens, a especificidade e o tamanho determinam a representatividade do *corpus* conforme a origem da população que o produziu. Indo além, a representatividade também está ligada ao caráter probabilístico da linguagem, uma vez que estabelece uma conexão direta entre particularidades mais comuns e menos comuns em dados contextos de uso da linguagem (idem, 2004).

Mas a discussão sobre o estatuto da LC como disciplina ou como método está longe de acabar, principalmente porque o objeto da LC não é delimitado como noutras áreas (BERBER SARDINHA, 2004); ela na verdade toma o objeto de áreas correlatas, como, por exemplo, o léxico, a sintaxe, etc. Se entendermos metodologia como instrumental, nada impede que se faça uso do instrumental da LC em investigações de outras áreas. Isso seria o que se chama *corpus-driven approach*. E a LC não necessariamente precisa ser classificada como disciplina, ela pode ser uma abordagem, uma perspectiva sobre a linguagem – e as pesquisas cuja abordagem se vincula a essa visão são consideradas *corpus-based*.

Berber Sardinha (2004, p. 38) agrupa os estudos em LC desenvolvidos nas últimas décadas em dois paradigmas: paradigma informal baseado em concordâncias (que concentra a maior parte das pesquisas e é caracterizado por descrições da

[15] Em inglês: (…) elaborating a reliable methodology to describe (…).

linguagem); e paradigma estatístico (subdividido em modelos de regressão linear e modelos ocultos de Markov). O primeiro é, evidentemente, mais qualitativo, ao passo que o segundo é quantitativo e parece mais em sintonia com a linguística computacional. Conforme Teixeira (2008, p. 153), grande parte dos trabalhos concentra-se na vertente aplicada da LC:

— Em Lexicografia e Terminologia: criação de dicionários, glossários e bases de dados.
— Em aprendizado e ensino de língua estrangeira: confecção de material didático a partir do uso real da língua.
— Em Linguística Computacional e Processamento de Língua Natural: produção de ferramentas de auxílio à fala e à escrita, tradução automática, automatização de tarefas linguísticas, reconhecimento de voz, extração e recuperação de informação, sumarização e simplificação textual, etc.
— Em tradutologia: observação contrastiva de textos-fonte e traduções, contextos definitórios, etc.

1.2.1 LINGUÍSTICA DE CORPUS E ANÁLISE MULTIDIMENSIONAL

A Análise Multidimensional (AMD) é uma metodologia de estudo da linguagem baseada em *corpus* criada por Douglas Biber, acadêmico da Universidade do Arizona, nos Estados Unidos. Com base em uma abordagem estatística multivariada conhecida como análise fatorial, o objetivo da AMD é permitir a comparação entre vários registros[16] a partir de diferentes parâmetros linguísticos – as "dimensões". Tais dimensões caracterizam-se de acordo com a função comunicativa dos fatores, ou seja, com o papel que os fatores desempenham no texto. Dois registros podem apresentar diferenças de maior ou menor grau em cada dimensão. Ao considerar todas as dimensões linguísticas, é possível descrever como um registro se diferencia de outro, bem como qual é o alcance dessa diferença; assim, pode-se, em última instância, descrever o padrão geral da variação do registro em uma língua (BIBER e CONRAD, 2009). O nome dessa

[16] Para uma definição do que Biber considera "registro", ver BIBER, Douglas; CONRAD, Susan. *Register, genre and style*. Nova York: Cambridge, 2009. Exemplos de registro são textos orais e textos escritos, é neste sentido que se está usando o termo aqui.

abordagem deriva do conceito de dimensão de variação. Dimensão, nesse caso, é entendida como um conjunto de traços de um *corpus* (BERBER SARDINHA, 2004, p. 300).

A partir da técnica estatística usada para identificar padrões de coocorrência (a análise fatorial), cada conjunto de traços comuns é chamado de fator. Para cada fator sob análise, o número de variáveis – ou seja, os traços linguísticos – é reduzido a um pequeno conjunto de variáveis subjacentes: os fatores ou dimensões de variação. E cada fator representa um grupo de traços linguísticos que tendem a coocorrer em textos (BIBER e CONRAD, 2009, p. 225). O enfoque da AMD inova por combinar análises de nível macro com análises de nível micro, ou seja, da macrodimensão do *corpus* chega-se à microdimensão do texto, e a microdescrição dos traços de cada texto revela macroagrupamentos textuais, que caracterizam os gêneros (FINATTO, 2011).

Os passos metodológicos da AMD são os seguintes:

1) Coleta e tratamento do *corpus*.
2) Identificação do conjunto de traços linguísticos a serem incluídos na análise a partir de estudos anteriores ou descrições disponíveis.
3) Análise automatizada do *corpus* para calcular as frequências de cada traço linguístico em cada texto.
4) Analisam-se os padrões de coocorrência dos traços usando-se análise fatorial das contagens de frequência.
5) Calculam-se os escores de cada texto de acordo com as dimensões; a média dos escores são então comparadas para analisar semelhanças e diferenças linguísticas.
6) Os "fatores" obtidos a partir da análise fatorial são interpretados conforme a função linguística que desempenham.

É importante ressaltar também que, no caso do sistema Coh-Metrix, a unidade de processamento é a unidade do texto, destacado em meio a um *corpus*. O movimento analítico *corpus*-texto-*corpus*, embutido no sistema, é realizado a partir de um *corpus* de treinamento que pode ter características bem diferentes dos textos literários aqui em foco, e essa é uma limitação deste trabalho.

Em um primeiro momento, considerei a AMD uma opção metodológica interessante para levantar traços linguísticos coocorrentes em textos originais e textos traduzidos, pois a possibilidade de identificar padrões particulares a um e outro tipo de texto seria uma forma de apontar eventuais desequilíbrios no nível de complexidade desses textos. No entanto, conforme a pesquisa progrediu, optei pela Aprendizagem de Máquina (AM) (ver seção 1.3.1 deste capítulo) em função da menor incidência de erro no levantamento e manuseio dos dados e também porque as métricas das ferramentas Coh-Metrix e Coh-Metrix-Port constituem, por si só, traços linguísticos relevantes para o contraste entre os textos.

1.3 A LINGUÍSTICA COMPUTACIONAL E O PLN

A Linguística Computacional e o Processamento de Língua Natural (PLN) cresceram e se expandiram em anos recentes. A partir da Inteligência Artificial e da Linguística, desenvolveram-se como disciplinas relativamente independentes e tornaram-se áreas importantes na pesquisa de sistemas, modelos e técnicas de processamento de línguas naturais (O'KEEFFE e McCARTHY, 2010; VOLPE NUNES, 2008). Renata Vieira (2004) afirma que a Linguística Computacional "preocupa-se com a compreensão da língua e de técnicas computacionais adequadas para o tratamento da língua escrita e falada, tanto para sua interpretação quanto sua geração", e que o PLN "tem o objetivo de reproduzir comportamentos inteligentes em sistemas computacionais, como a solução de problemas e automatização do raciocínio". Para Dias da Silva (2006), a Linguística Computacional nasceu com o foco voltado para o estudo de algoritmos para análises morfológicas e gramaticais, enquanto o PLN diferencia-se dela por ter a particularidade de:

> agregar uma heterogeneidade de objetivos: desde a meta de investigar meios de empregar o computador como uma simples ferramenta auxiliar para investigar material linguístico (por exemplo, a criação de programas de computador para calcular estatísticas de ocorrências de palavras em textos ou para identificar e indexar palavras e segmentos de texto), até a meta de criar uma inteligência artificial. (op. cit., p. 104)

O tratamento computacional de línguas naturais é o desafio assumido pelo PLN, que nasce com a tarefa de investigar e criar modelos de língua operacionalizáveis pelo computador (DIAS DA SILVA, 2006). Tendo a língua por matéria-prima, o PLN abarca tarefas tão heterogêneas quanto a própria linguagem – desde criar uma inteligência artificial com características comunicativas antropomórficas (robótica) até tarefas mais simples, como o aprimoramento do computador como ferramenta de auxílio à comunicação e à pesquisa linguística. No entanto, uma vez que a manipulação da fala apresenta problemas tecnológicos específicos, os estudos nessa área tendem a ser realizados de forma independente, fazendo com que o PLN seja praticamente "sinônimo de processamento de língua escrita" (VOLPE NUNES, 2008).

Dias da Silva (1996) propôs uma estratégia de três etapas para o desenvolvimento de projetos em PLN: a fase linguística, a fase representacional e a fase implementacional, como mostra a Figura 2. Na primeira, o objetivo é criar o modelo linguístico. Na segunda, o objetivo é representar o modelo da primeira fase de forma a ser "compreendido" pelo computador. Na fase final, o objetivo é implementar o modelo em uma aplicação codificada a partir das representações da segunda fase.

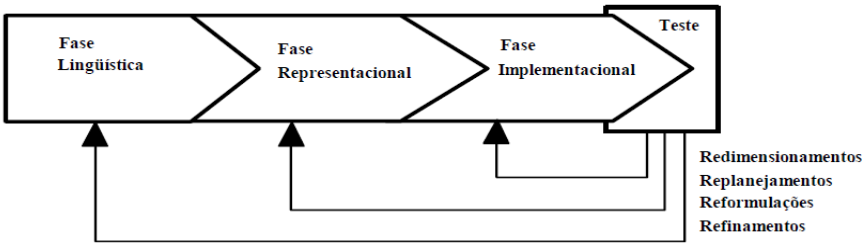

Figura 2. As três fases para o desenvolvimento de projetos em PLN. Fonte: Dias da Silva, 2006.

Para o autor (1996, p. 178), o equacionamento do domínio representacional do PLN envolve a discussão de questões em três níveis:

• **morfossintático**, que trata da representação das gramáticas e dos analisadores gramaticais, incluindo a representação das regras e das estruturas morfossintáticas e de léxicos enriquecidos com informações pragmático-discursivas;

• **semântico**, que trata da representação de estruturas semânticas, de domínios conceituais e de estratégias computacionais de interpretação dessas representações;

• **pragmático-discursivo**, que trata da representação da estrutura do discurso e dos contextos pragmático-discursivo e situacional.

Bento Dias da Silva (1996) afirma que essas representações precisam ser explícitas, consistentes e não ambíguas para serem transformadas em programas de computador. Além disso, não necessariamente precisam ocorrer de forma sucessiva, podem ser realizadas simultaneamente a partir da etapa linguística, que é a base para as outras.

A principal tarefa de PLN foi (e continua a ser) a tradução automática (TA), um esforço iniciado na década de 50, bem antes da explosão tecnológica do fim do século. A expectativa era a de que o computador produziria traduções perfeitas e completas entre duas ou mais línguas tal como um ser humano. Propunha-se a gerar traduções em um nível de sofisticação que talvez nem os tradutores humanos atinjam. Mesmo após o surgimento de computadores mais potentes, a TA continuou a encontrar barreiras e a frustrar os pesquisadores de PLN.

Uma das razões para esse fracasso, de acordo com Ronaldo Martins (2011), é o modelo antropomórfico de língua com que se insiste em trabalhar em PLN e em TA. Mais do que um acúmulo de conhecimentos que permitiriam a ampliação progressiva do desempenho dos sistemas de tradução automática, Martins acredita que o principal obstáculo à criação e à consolidação de um sistema robusto e escalável de IA são as premissas difusas do PLN:

> (O PLN), embora evidentemente profícuo na produção de aplicativos de utilidade incontestável, constitui principalmente uma dispersão, sem que possa ser observada, nitidamente, a hegemonia de um corpo teórico sobre os demais. Trata-se, na verdade, de uma coleção de posturas difusas e fragmentárias que orbitam um objetivo comum: ensinar a máquina a falar. (2011, p. 291)

A natureza interdisciplinar do PLN, em que participam matemáticos, cientistas da computação e linguistas, é, por si só, uma fonte de mal-entendidos entre os pesquisadores, sobretudo por agrupar pesquisadores de áreas com tradições tão diversas. Por um lado, os cientistas da computação esperam dos linguistas uma concepção pronta de língua, matematizável, formalizável e processável; os linguistas, por sua vez, esperam dos cientistas da computação soluções instantâneas para problemas encontrados na

manipulação de dados linguísticos. Além disso, tanto a Linguística quanto a Ciência da Computação são ciências que, por serem "jovens", ainda discutem sua matéria, seu objeto e suas tarefas. E o PLN, em que pesquisadores dessas disciplinas se unem, acaba herdando as indefinições das áreas que o compõem, tal como ocorre com os Estudos de Tradução em relação aos Estudos da Linguagem.

Para Martins (2011), ele mesmo linguista, é preciso desviar o foco do PLN da *langue* e direcioná-lo à *parole*, ou seja, em vez de tentar impor um modelo humano de língua ao computador, é preciso criar um modelo de aquisição de língua para a máquina, que poderá, então, *aprender* a língua e usá-la (*parole*), mas com base num modelo de *máquina*. *Langue* e *parole*, vale frisar, são conceitos de Saussure. A visão de Martins, no entanto, ainda pressupõe a mesma tarefa colossal do início dos estudos em processamento de línguas naturais, em especial em TA: a de fazer com que a máquina realize funções exclusivamente humanas.

Volpe Nunes (2008), a esse respeito, afirma que as expectativas do usuário, após um convívio mais intenso com o computador, são mais realistas – como também são mais realistas as expectativas dos envolvidos na execução das tarefas de PLN. A autora, ao contrário de Martins, acredita que a complexidade da tarefa foi inicialmente subestimada e que, por isso, as abordagens ao problema da TA foram, de certa forma, ingênuas – tanto da parte dos cientistas da computação quanto dos linguistas. Desse embate, fica claro que a expectativa irrealista de desempenho dos sistemas de TA são, antes de mais nada, consequência das expectativas irreais em relação ao que cada uma das disciplinas contribuiria para a concretização da tarefa.

Já Bento Dias da Silva (1996, p. 253) acredita que, "de um lado, é necessária a explicitação de um referencial teórico-metodológico mínimo que passe a servir de norte para pesquisas integradas, e, de outro, a difícil adoção de posturas científicas mais cooperativas". Afirma ainda que o trabalho em conjunto entre pesquisadores de áreas distintas em prol de um objetivo comum a fim de criar modelos de PLN aponta um caminho de reconciliação e de parceria que trará frutos a todos ao facilitar a comunicação entre o homem e a máquina.

É importante que essas questões sejam reconhecidas e discutidas, pois corremos o risco de perder de vista a motivação do PLN. No que diz respeito à TA, é fundamental

que se tenha claro a quem servirá um sistema automatizado de tradução. O que me parece irônico é o fato de que quem mais se beneficiaria seriam os tradutores profissionais – irônico porque, em vez de tentar auxiliar operações humanas de tradução, a TA parece se colocar numa posição de competição com os tradutores. A TA se beneficiaria imensamente se ampliasse os seus objetivos também às necessidades *reais* da atividade profissional de tradutores *reais*. Para o "público geral", os sistemas atuais de TA são razoavelmente aceitáveis. É quando um tratamento mais sofisticado do texto é necessário que os sistemas apresentam desempenho insatisfatório. Justamente nesse ponto é que a contribuição de tradutores humanos e de profissionais do texto é fundamental.

É também nesse espírito que esta dissertação, uma pesquisa circunscrita à Linguística, faz uso de ferramentas de PLN: com o intuito de extrapolar a ênfase nos resultados obtidos a fim de, futuramente, usar esses dados para projetar ferramentas mais adequadas às necessidades específicas de usuários específicos – e aqui me refiro, sobretudo, a tradutores e a revisores. Assim, como mencionei anteriormente, a motivação desta pesquisa foi também a de dialogar produtivamente com os pesquisadores de PLN, apontando problemas surgidos na prática profissional concreta de tradutores/revisores – neste caso, o fenômeno da complexidade textual em traduções literárias – e, por meio do uso de ferramentas criadas em PLN, contribuir para o aperfeiçoamento desses recursos.

1.3.1 APRENDIZAGEM DE MÁQUINA

Witten e Frank (2005) definem Aprendizagem de Máquina (AM) como um campo da Inteligência Artificial dedicado ao estudo de sistemas automáticos de aquisição e integração de conhecimento. Para os autores, AM é a aquisição de descrições estruturais a partir de exemplos, e as descrições podem usadas para vários fins, como predição, explicação e compreensão de um banco de dados segundo seus padrões. Algoritmos de AM têm valor inestimável em várias aplicações, como, por exemplo: problemas de mineração de dados, na busca de regularidades implícitas em grandes bancos de dados; em áreas em que as máquinas possam ter desempenho melhor que os humanos (p. ex., reconhecimento de expressões faciais em imagens); em áreas em que há necessidade de adaptação do programa (p. ex., a máquina "aprende" as preferências de

um indivíduo, adaptando-se a elas); e em áreas em que a aquisição ou sistematização manual do conhecimento é muito trabalhosa, como é o caso do PLN.

Souza ressalta que há vários sistemas de AM, os quais possuem características particulares e compartilhadas "que possibilitam sua classificação quanto à linguagem de descrição, modo, paradigma e formas de aprendizado" (2011, p. 42). As estratégias de aprendizado listadas pela autora são: aprendizado por hábito, por instrução, por dedução, por analogia e por indução. O aprendizado indutivo pode ser dividido em supervisionado, não supervisionado e semissupervisionado, modelo em que as duas técnicas são mescladas. Entretanto, seja qual for a estratégia, há modelos comuns a todos os métodos, de acordo com Monard e Baranauskas (2003, apud SOUZA, 2011, p. 42):

- Modelo simbólico: constitui-se em aprendizagem através de representações simbólicas de conceitos por meio de exemplos e contraexemplos. As representações simbólicas costumam estar na forma de alguma expressão lógica, árvore de decisão, regras ou rede semântica.
- Modelo estatístico: costuma usar modelos probabilísticos baseados no conhecimento anterior de um problema, combinado com exemplos extraídos de um conjunto de dados de treinamento para determinar a probabilidade de uma hipótese.
- Modelo baseado em exemplos: classificar um exemplo a partir de outro similar, cuja classe é conhecida, e assumir que esse novo exemplo terá a mesma classe.
- Modelo conexionista: são as Redes Neurais, construções matemáticas inspiradas no modelo neurológico humano. Envolve unidades altamente interconectadas e, assim, é denominado conexionismo.
- Modelo evolutivo: uma população de elementos de classificação que competem para fazer a predição.

Em PLN, métodos de AM têm sido usados em diversas tarefas de mineração de dados (p. ex., CANDIDO Jr. et al., 2009; ALUISIO et al., 2010), recuperação de informação, mineração de textos (p. ex., LOPES et al., 2009, 2010), tradução automática (p. ex., SPECIA, 2010), entre outras. Os algoritmos de AM comumente usados nessas tarefas são indutivos: podem ser supervisionados, ou seja, com foco na *extração* de

regularidades a partir dos padrões de um conjunto de dados de treinamento, ou não supervisionados, com enfoque na *descoberta* de propriedades, padrões e estruturas nos dados analisados. Em geral, as tarefas de PLN que fazem uso de AM podem ser tratadas como problemas supervisionados de classificação, em que, dado um determinado objeto em um determinado contexto, a classe a que esse objeto pertence pode ser determinada, como, por exemplo, em tarefas de etiquetagem morfossintática (DAELEMANS & HOSTE, 2002).

O método automático de classificação usado nesta pesquisa baseia-se no modelo estatístico supervisionado de AM, e a ferramenta usada é o Weka (Waikato Environment for Knowledge Analysis[17]). O Weka é uma coleção de algoritmos de AM que contém ferramentas para pré-processamento, classificação, regressão, agrupamento e associação de dados. O algoritmo escolhido foi a implementação J48 do algoritmo de classificação C4.5 para construção de árvores de decisão (ver o item 4.3 do Capítulo 4).

Alguns exemplos de aplicação de técnicas de AM incluem: previsão do tempo; seleção de embriões in vitro; previsão de comportamento de consumidores; diagnósticos médicos; mineração de textos para criação de glossários. Um exemplo bastante conhecido é o da rede americana WalMart. Ao procurar relações entre o volume de vendas e os dias da semana, o *software* de AM apontou que, às sextas-feiras, as vendas de cervejas cresciam na mesma proporção que as de fraldas, pois, ao comprar fraldas para os filhos, os pais aproveitavam para abastecer o estoque de cerveja para o final de semana.

[17] http://www.cs.waikato.ac.nz/ml/weka/

2. REVISÃO DA LITERATURA (II)

Uma vez que este trabalho trata sobre o tema da complexidade textual, é importante revisar conceitos sobre Leitura. A abertura desta parte traça, de forma resumida, um panorama de estudos textuais sobre a questão da complexidade, incluindo a história do índice Flesch. A seguir, é definido o perfil social do leitor a que nos referimos, trazendo dados das pesquisas mais recentes de instituições como o Indicador de Alfabetismo Funcional (INAF). É feita uma revisão, também, de perspectivas tradutológicas sobre leitura e leitor, levando em conta principalmente a posição de profissionais do texto diante da tarefa tradutória. Por fim, é delineado um brevíssimo panorama da Leitura segundo a visão da Crítica Literária e dos Estudos de Tradução, com ênfase no leitor diante do texto literário.

2.1 LEITURA

2.1.2 EM DIREÇÃO AO TEXTO: PESQUISAS EM COMPLEXIDADE TEXTUAL

As pesquisas sobre o tema da complexidade textual partem de pontos de vista diversos, especialmente pela ótica da leitura e do ensino de leitura. Em geral, esses estudos, entre diferentes possibilidades de realização e de aplicação de resultados, tendem a seguir três grandes grupos de encaminhamento:

a) centrar-se em características ou elementos presentes em determinados tipos de texto e associá-los a determinadas dificuldades ou facilidades de compreensão de uma determinada categoria de leitores;

b) inferir o *modus operandi* do processamento mental-cognitivo de diferentes tipos de leitor mediante aplicações de testes de compreensão após ou durante diferentes tipos de experimentos de leitura com diferentes tipos de texto;

c) reunir os indicativos dos itens *a* e *b* em prol da melhoria do ensino de leitura ou da produção materiais informativos com maior acessibilidade de compreensão leitores com proficiência de leitura reduzida ou em fase de desenvolvimento.

Na bibliografia estrangeira, há registros de pesquisas sobre *readability* ou inteligibilidade ou complexidade linguística pelo menos desde os anos 1920, conforme assinalam Davison e Green (1988, p. 121), que afirmam que esses estudos surgiram a partir da necessidade de adequar materiais de leitura a públicos específicos. Os estudos baseavam-se no pressuposto de que todos os problemas de leitura estão relacionados a traços textuais mensuráveis, os quais, após a sua identificação, são inseridos em fórmulas cujos resultados estimam a legibilidade de um texto. Os traços mais comumente mensurados, até hoje, são a dificuldade lexical, baseada na frequência e na extensão das palavras, e a dificuldade imposta pelo tamanho da sentença, a partir do cálculo do número de palavras por sentença.

Davidson e Green (1988, p. 122) criticam a superficialidade de fórmulas puramente lexicais – ou seja, fórmulas com base na medida de frequência e extensão de palavras e frases – e argumentam que a complexidade de um texto não é um traço que possa ser fisicamente isolado sem que se levem em conta outras variáveis como complexidade sintática, características discursivas, estrutura retórica e assim por diante. Além disso, para as pesquisadoras, as fórmulas não consideram aspectos importantes relativos ao leitor, como motivação e interesse, objetivo de leitura, etc. As autoras afirmam que uma visão mais holística da questão da complexidade textual desenvolveu-se a partir desses estudos iniciais, e essas novas perspectivas são provenientes de três fontes: do leitor, das características intrínsecas do texto e de teorias linguísticas.

Para Hoey (1991), hoje uma referência entre os linguistas de *corpus*, no âmbito dos estudos sistêmico-funcionais de Linguística Textual de Halliday e Hassan iniciados na década de 60 e 70, o texto escrito só é "ativado" em todos os níveis – sintático, fonológico, semântico, pragmático – por meio da leitura e da interação com um leitor real. O texto oferece conexões semânticas potenciais tanto no nível da palavra quanto no nível da oração, mas é preciso que o leitor ative esses recursos e selecione as conexões mais relevantes.

O papel do leitor, de acordo com Hoey (op. cit., pp. 221-225), é ativo e criativo, e a sua interação com o texto é livre. Essa liberdade de interação com o texto manifesta-se de várias maneiras, sendo a mais significativa a liberdade de reconhecer ou ignorar conexões lexicais e oracionais do texto. Se o leitor não reconhece as conexões, ele não

tratará todas as frases potencialmente conectadas como conectadas de fato, deixando de perceber a relação semântica entre elas. É na apreensão de probabilidades organizacionais (através, evidentemente, da leitura) que o leitor encontra pistas de conexão entre palavras, frases e estruturas do texto. O leitor passa a antecipá-las assim que as apreende. Além disso, Hoey (op. cit.) acredita que não há sentido a ser "extraído" ou "encontrado" – para ele, o leitor é quem dá sentido ao texto.

Hoey (op. cit.) afirma também que o objetivo da leitura é essencial na avaliação da compreensão do texto pelo leitor. Dependendo do propósito da leitura, é possível que o leitor faça conexões não antecipadas pelo autor do texto. Entretanto, apesar não explicar o papel do autor nas intenções de produção textual, Hoey entende o texto como um conjunto de possibilidades interpretativas independentes das intenções do autor, ou seja, o texto existe por si só, como entidade autônoma, e são as estratégias de leitura, determinadas pelo objetivo do leitor no momento da leitura, que estabelecem as conexões de sentido do texto.

Para Graesser e colaboradores (2004), as fórmulas de legibilidade e de avaliação de complexidade ignoram componentes linguísticos e discursivos que influenciam na dificuldade de compreensão textual. Os autores apontam para o fato de que, apesar de os parâmetros de tamanho das sentenças e das palavras terem alguma validade, tais parâmetros não revelam, por si só, a complexidade de um texto. Assim, propõem uma análise da coesão e da coerência textual em múltiplos níveis. De acordo com Graesser (2004), coesão textual é uma propriedade objetiva do texto, e coerência é a representação mental do conteúdo do texto feita pelo leitor através das palavras, sentenças e frases que orientam a leitura e conectam as ideias umas às outras. O desafio, segundo os autores, é automatizar esses níveis mais profundos de análise textual. Essa foi a motivação da criação da ferramenta Coh-Metrix.

Entretanto, a questão da complexidade textual extrapola os limites da discussão acadêmica. Nos Estados Unidos, por exemplo, as pesquisas sobre avaliação de inteligibilidade textual são uma questão de saúde pública. Dubay (2004) mostra que, de acordo com estimativas do National Center for Health Statistics, o maior fator de risco para lesões em acidentes de trânsito é o uso inadequado de assentos infantis. Especialistas norte-americanos em saúde pública acreditam que uma das principais

causas desse problema é a compreensão insatisfatória dos manuais de instrução dos assentos para crianças. De acordo com os órgãos governamentais responsáveis por estatísticas de alfabetização, o americano médio tem proficiência de leitura proporcional à de alunos de sétima série do Ensino Fundamental brasileiro. Ao analisarem 107 manuais fornecidos pelos fabricantes de assentos infantis, os pesquisadores constataram que o nível de proficiência requerido era, em média, compatível com o nível de proficiência de leitura de alunos do segundo ano do Ensino Médio (conforme o sistema brasileiro). Assim, ficou clara, para esses especialistas, a urgência da adoção de medidas de simplificação de documentações de instrução voltadas ao público em geral.

Para Dubay (2004), inteligibilidade (*readability*) é o que faz com que alguns textos sejam mais fáceis de ler do que outros, e legibilidade (*legibility*) são aspectos físicos do texto, como design, tipo de fonte, etc. Segundo o autor, as variáveis usadas nas fórmulas de avaliação de complexidade textual correspondem ao esqueleto de um texto (op. cit., p. 57). Assim como Davison e Green (1982), Dubay (op. cit., p. 3) lista momentos distintos na evolução das pesquisas sobre complexidade textual nos Estados Unidos:

- Primeira fase: **estudos sobre alfabetização**, que revelaram as diferenças entre os níveis de proficiência de leitura dos adultos e sua implicação social.
- Segunda fase: **estudos clássicos sobre complexidade textual**, iniciados no fim do século XIX e concluídos na década de 40, com a publicação do índice Flesch e de Dale-Chall. O foco, nessa fase, era fazer com que os textos fossem adequados à proficiência do leitor.
- Terceira fase: **novos estudos sobre complexidade textual** a partir da década de 50. Com o desenvolvimento de testes de compreensão de leitura e com a contribuição da linguística e da psicologia cognitivas, os pesquisadores passaram a explorar o modo como o interesse, a motivação e os conhecimentos prévios do leitor afetam a compreensão de um texto. Esses estudos levaram à criação de fórmulas novas e mais precisas.

As fórmulas, embora sempre muito criticadas, sobreviveram a mais de oitenta anos de aplicações, investigações e polêmicas, sem nunca deixar de ser usadas para gerar

textos de acesso mais facilitado para uma grande fatia da população leitora (DUBAY, 2004) nos Estados Unidos. Naturalmente, há aqui, nessa ideia de "facilitação", implicações históricas, interesses políticos, ideológicos e econômicos associadas ao acesso à escolarização, à cultura letrada e ao incremento da produção e do consumo para faixas maiores de população. Esses aspectos não serão discutidos nesta dissertação. Aqui serão considerados somente aspectos mais estritamente relacionados a padrões de uso da língua no que tange à recepção de textos escritos.

No Brasil, um dos primeiros linguistas a tratar do tema da leitura funcional e da maior ou menor habilidade de leitura foi Perini (1982) com o trabalho Tópicos discursivos e legibilidade (*apud* FULGÊNCIO, LIBERATO, 2004, p. 9). Propunha o autor, então, que os estudantes brasileiros tivessem acesso a materiais de leitura graduados de acordo com o seu nível de escolaridade e nível de dificuldade de compreensão. A partir do legado de trabalhos fundadores tais como o de Perini, surgiram os trabalhos de Neis (1982), de Angela Kleiman (1987, 1995 e 1997) e Kato (1982), entre outros, produzidos especialmente ao longo dos anos 80 e 90, e temos hoje no Brasil alicerce multifacetado de estudos sobre o tema da Leitura. Esse corpo de conhecimento permite distinguir especificidades das noções de leitura, alfabetização, letramento, competência textual e competência leitora. Isso sem mencionarmos os inúmeros trabalhos sobre o tema da Leitura na área da Educação, Ensino de Língua Portuguesa e de Línguas Estrangeiras ou de Psicolinguística (FINATTO, 2011, p. 32 e seguintes).

No livro *Ler e Compreender* (2007), Ingedore Koch e Vanda Maria Elias apresentam uma visão de leitura a que denominam "sociocognitivo-interacional", uma abordagem que privilegia a interação entre sujeitos e seus conhecimentos. A interação entre autor-leitor-texto pressupõe a operação de estratégias de leitura usadas pelo leitor e que o conduzirão à significação do texto, ou seja, à produção de sentidos mobilizados pela leitura. Esse processo é dirigido e regulado pelo leitor em todas as instâncias de leitura ao estabelecer relações entre conhecimentos prévios e as novas informações contidas no texto.

As autoras apontam a pluralidade de leituras e sentidos:

> "Considerar o leitor e seus conhecimentos e que esses conhecimentos são diferentes de um leitor para outro

> implica aceitar uma pluralidade de leituras e de sentidos em relação a um mesmo texto." (2007, p. 21)

Mais adiante, afirmam o seguinte:

> "Se, do lado do autor, foi mobilizado um conjunto de conhecimentos para a produção do texto, espera-se, da parte do leitor, que considere esses conhecimentos (de língua, de gênero textual, de mundo) no processo de leitura e construção de sentido." (2007, p. 27)

Segundo Koch e Elias (2007), é nos conhecimentos mobilizados pelo autor durante a produção do texto que o perfil de um leitor "ideal", ou seja, um leitor-modelo, forma-se. Nessa acepção, o contexto sociocognitivo é condicionante para que autor-texto-leitor interajam de forma satisfatória, isto é, que o contato do leitor com o texto crie condições de produção de sentido.

No entanto, ainda que nessa visão teórica o leitor seja considerado em todas as suas capacidades, o que fica a desejar é a noção do autor (como também em HOUEY, 1991). O autor é a soma das intenções de um texto; entretanto, a figura do autor não é necessariamente a do indivíduo que o escreveu originalmente – como é o caso das traduções, em que o tradutor assume o papel do autor. Outro ponto importante a ser ressaltado é que, se é o autor quem determina o seu leitor ideal, então o leitor ideal é, em última instância, uma função da intenção do texto. E a complexidade de um texto só poderá ser avaliada de acordo com a relação estabelecida entre autor (intenção) e leitor.

É isso que Leffa (1996) já apontava como uma descrição completa do processo da compreensão, que deve levar em conta três aspectos essenciais: o texto, o leitor e as circunstâncias em que se dá o encontro entre ambos. Ao tratar do papel do texto, Leffa observa que, nos estudos atuais, ainda há uma preocupação centrada no léxico e na estrutura sintática das frases. Porém, ao contrário de estudos desenvolvidos durante as décadas de 50 e 60, as análises do texto evoluíram da micro para a macroestrutura, levando aos estudos de gênero textual.

2.1.2.1 ÍNDICE FLESCH PARA AVALIAÇÃO DE COMPLEXIDADE TEXTUAL

No Brasil, foi só mais recentemente que pesquisadores de Linguística Computacional e Processamento de Língua Natural se interessaram por fórmulas e

medidas de complexidade textual, adaptando-as ao português. É o caso do índice Flesch, adaptado para o português brasileiro por pesquisadores do Instituto de Ciências Matemáticas e da Computação da Universidade de São Paulo (MARTINS et al., 1996). Por se tratar da única fórmula de complexidade textual adaptada para o português do Brasil, é importante narrar, em poucas palavras, a sua trajetória.

Rudolf Flesch nasceu na Áustria e formou-se em Direito na Universidade de Viena em 1933. Trabalhou como advogado até 1938, quando imigrou para os Estados Unidos, fugindo do regime nazista em ascensão na Europa. Como o diploma austríaco não fora aceito nos Estados Unidos, Flesch trabalhou em diversas funções, inclusive no setor de despachos de uma gráfica em Nova York. Como imigrante, ao perceber a dificuldade de estrangeiros em compreender textos oficiais do governo americano, Flesch interessou-se pelo assunto e passou a dedicar-se aos estudos após ganhar uma bolsa, em 1939, na Universidade de Columbia. Em 1940, formou-se em Biblioteconomia e, em 1942, completou seu mestrado em Educação de Adultos. No ano seguinte, recebeu o grau de doutor em pesquisa em Educação com a dissertação "Marks of a Readable Style". Nos anos seguintes, escreveu *The Art of Plain Talk* (1946), *The Art of Readable Writing* (1949), *The Art of Clear Thinking* (1951), *Why Johnny Can't Read – And What You Can Do About It* (1955), *The ABC of Style: A Guide to Plain English* (1964), *How to Write in Plain English: A Book for Lawyers and Consumers* (1979) (DUBAY, 2004).

Em "Marks of a Readable Style", a primeira fórmula para estimação da complexidade textual de materiais voltados para adultos foi publicada. E as editoras, ao aplicarem a fórmula proposta por Flesch, logo perceberam que o número de leitores aumentou entre 40 a 60%. A partir daí, pesquisadores de diversas áreas passaram a usá-la em seus textos, com o objetivo de tornar os seus textos mais acessíveis (DUBAY, 2004).

No ano de 1948, Flesch publicou uma segunda fórmula, em duas partes. Na primeira, a Reading Ease, são usadas duas variáveis: o número de sílabas e o número de sentenças a cada amostra de 100 palavras. A complexidade é estimada em uma escala de 1 a 100, sendo 1 equivalente a muito difícil e 100 a muito fácil. A segunda parte da fórmula estima o "interesse humano" ao contar o número de palavras pessoais, como pronomes e nomes, e marcas de personalização, como citações, exclamações e frases incompletas (DUBAY, 2004).

A fórmula é a seguinte:

Escore = 206.835 – (1.015 x TMS) – (84.6 x MSP)

Onde:

Escore = posição numa escala de 0 (difícil) a 100 (fácil), com até 30 = muito difícil e acima de 70 = adequado para todos os públicos.

TMS = tamanho médio das sentenças (o número de palavras dividido pelo número de sentenças).

MSP = número médio de sílabas por palavras (o número de sílabas dividido pelo número de palavras).

Em 1949, Flesch publicou os resultados de um estudo conduzido ao longo de dez anos sobre o conteúdo do editorial de diversas revistas e jornais de prestígio nos Estados Unidos. O que ele revelou foi o seguinte:

- Cerca de 45% da população entendia o editorial do jornal The Saturday Evening Post.
- Cerca de 50% da população entendia os editoriais das revistas McCall's, Ladies Home Journal e Woman's Home Companion.
- Pouco mais de 50% da população entendia o editorial da revista American Magazine.
- Mais de 80% da população conseguia entender o editorial da revista Modern Screen, Photoplay.

Esses resultados revolucionaram a imprensa dos Estados Unidos, e tanto Flesch quanto colegas e colaboradores de suas pesquisas prestaram assessoria para as agências de notícias United Press e a Associated Press, que baixaram o índice de complexidade de leitura dos seus editoriais (DUBAY, 2004). Outro elemento de suma importância indicado pelas pesquisas de Flesch foi a perda de interesse do leitor quando o texto se mostra muito complexo, o que também foi corroborado por pesquisas posteriores (p. ex., McCNAMARA et al., 2002; GRAESSER et al., 2004).

2.1.3 A LEITURA E OS LEITORES

"Conhecimento." "Tudo na vida do homem." Eis algumas das respostas espontâneas à pergunta "O que é leitura?", conforme revela a pesquisa Retratos da Leitura no Brasil (AMORIM, 2008). De acordo com essa pesquisa, a leitura tem significado positivo no imaginário de três em cada quatro brasileiros, e uma em cada quatro pessoas não faz a menor ideia sobre o papel da leitura. O que chama a atenção, no entanto, é que nenhum dos entrevistados afirmou considerar a leitura um direito.

O Indicador de Alfabetismo Funcional (INAF), programa de pesquisa empreendido pela Ação Educativa e pelo Instituto Paulo Montenegro desde 2001 (RIBEIRO, 2010), subdivide em quatro categorias os níveis de proficiência de leitura: analfabetismo (incapacidade de realizar tarefas simples de leitura), alfabetismo em nível rudimentar (capacidade de localizar informações em textos curtos e familiares), alfabetismo em nível básico (capacidade de ler e compreender textos de média extensão e localizar informações fazendo algumas inferências) e alfabetismo em nível pleno (capacidade de ler textos longos e relacionar suas partes, comparar informações e fazer inferências e sínteses). Analfabetismo e alfabetismo rudimentar são considerados índices de analfabetismo funcional, enquanto as categorias de alfabetismo básico e pleno são consideradas índices de alfabetismo funcional. A partir de 2005, os testes cognitivos aplicados aos entrevistados passaram a envolver também a resolução de operações matemáticas (para a avaliação do numeramento), além de leitura e escrita (para a avaliação do letramento) (RIBEIRO, 2010).

De acordo com os resultados mais recentes do INAF (2009), o analfabetismo funcional no Brasil diminuiu de 39% para 27% entre 2001 e 2009, e o índice de alfabetismo funcional aumentou de 60% para 73% no mesmo período. A diferença mais notável é o aumento de indivíduos na faixa do alfabetismo básico: de 34% em 2001 para 46% em 2009, compondo a maioria da população entre 15 e 64 anos. Um dado alarmante da pesquisa, no entanto, mostra que **60% dos indivíduos que têm da 5ª. à 8ª. série do Ensino Fundamental e 56% dos indivíduos que cursaram o Ensino Médio são considerados alfabetizados em nível básico**. O perfil socioeconômico dos entrevistados revela que 51% dos que ganham entre dois a cinco salários mínimos e 48% dos que ganham entre um e dois salários mínimos estão também na faixa de alfabetização básica.

ESCOLARIZAÇÃO	NÍVEL DE LETRAMENTO (%)			
	ANALFABETO	RUDIMENTAR	BÁSICO	PLENO
NENHUMA	66	29	4	1
1ª A 4ª SÉRIE	10	44	41	6
5ª A 8ª SÉRIE	0	24	61	15
ENSINO MÉDIO	0	6	**56**	31
ENSINO SUPERIOR	0	1	**31**	68

Tabela 1. Nível de letramento da população brasileira de acordo com o INAF (2009).

Com o objetivo de divulgar um perfil detalhado da leitura no país, o Instituto Pró-Livro, em parceria com a Imprensa Oficial do Estado de São Paulo, publicou a última edição da pesquisa "Retratos da Leitura no Brasil" (RLB), mencionada anteriormente (AMORIM, 2008). Os resultados traçam o perfil não só daqueles que se declaram leitores (ou que leram ao menos um livro nos últimos três meses), mas também daqueles que se declaram não leitores. De acordo com a pesquisa, 55% dos entrevistados são leitores, e os 45% restantes são não leitores. Dentre os leitores, 45% têm até a segunda etapa do Ensino Fundamental ou o Ensino Médio, 61% têm mais de 18 anos e 78% pertencem às classes C e D. Já o grupo dos não leitores é composto por indivíduos entre 18 e 59 anos (61%), das classes C e D, sendo que 35% têm até a 4ª. série do Ensino Fundamental e 36% cursaram da 5ª. até a 8ª. série do Ensino Fundamental ou o Ensino Médio. A pesquisa mostra também que as classes C e D são responsáveis por 69% das compras de livros. Vê-se, portanto, que todos esses dados confirmam os indicativos do INAF – e acrescentam uma informação importante: os maiores consumidores de livros no Brasil são também as pessoas mais pobres.

No entanto, a pesquisa RLB traz informações também sobre quem não lê. No que diz respeito aos não leitores, ao serem indagados sobre limitações pessoais à leitura (resposta estimulada com múltipla escolha), 16% declararam ler muito devagar, 7% declaram não compreender o que leem, 11% declararam não ter paciência para ler e 7% declararam que não conseguem concentrar-se. Além disso, 83% dos não leitores pertencem às classes C e D – as que mais consomem livros no país. E a tendência atual é a de que a classe C cresça ainda mais, recebendo populações saídas das classes D e E, conforme apontam dados levantados pela Fundação Getúlio Vargas (FOLHA DE SÃO PAULO, 2011).

Em síntese, ambas as pesquisas sugerem que o contingente da maioria dos leitores brasileiros é composto por indivíduos entre 15 e 64 anos, das classes C e D, em nível básico de alfabetização (mas não necessariamente com baixa escolaridade). São brasileiros que não têm condições de exercer plenamente o direito à leitura por terem proficiência limitada de letramento e também por restrições socioeconômicas. Ainda assim, são as pessoas que mais leem no país. Uma especificidade desses leitores é a de que eles compõem o grupo dos chamados neoleitores – leitores adultos, com experiência de vida e domínio da oralidade, porém com experiência de leitura em níveis iniciantes (TIEPOLO, 2008). É este o leitor que será o foco desta dissertação.

O INAF e a pesquisa RLB mostram que os índices de analfabetismo no Brasil vêm diminuindo e que a maioria da população é composta por pessoas que se declaram leitoras. Contudo, fica claro também que o nível de letramento da maioria dos indivíduos é baixo. Assim, o que leem esses leitores? Entre os dez livros mais importantes na vida dos leitores (resposta espontânea e com uma única opção), seis são livros infantis ou infanto-juvenis, e os três escritores brasileiros mais admirados (resposta espontânea e com uma única opção) são Monteiro Lobato, Paulo Coelho e Jorge Amado, cujos textos, bem sabemos, são altamente acessíveis. Machado de Assis aparece em quarto lugar, e os quatro autores mais votados receberam quase metade das indicações. Percebe-se, então, que os leitores declaram preferir autores de linguagem "fácil", ainda que se leve em conta que o universo demográfico da amostra inclui crianças (13% dos entrevistados têm de cinco a dez anos).

Podemos especular também que o fato de leitores adultos terem citado livros infantis e infanto-juvenis como livros marcantes indica que o encontro com o livro provavelmente se deu em uma situação ideal: livro e leitor estavam em pé de igualdade – seja em termos de faixa etária, seja em termos do nível de inteligibilidade do texto. Os gêneros infantil e infanto-juvenil têm um público bem definido e um projeto editorial voltado para esse público, dois elementos que, em conjunto, criam condições favoráveis à leitura, ou seja, o livro estabelece uma espécie de intimidade com o leitor.

2.1.4 LEITURA, CRÍTICA LITERÁRIA E ESTUDOS DE TRADUÇÃO

Antoine Compagnon, em *O demônio da teoria* (1999), traça o percurso da noção de "leitor" no campo dos Estudos Literários e descreve as principais posições e perspectivas sobre o papel da leitura da obra literária para a Crítica a partir do século XIX. O primeiro embate entre visões sobre o leitor, conforme descreve Compagnon (1999, p. 140), se dá entre impressionistas e positivistas. Enquanto os representantes do impressionismo, como, por exemplo, Anatole France, colocavam a leitura como uma experiência cultural subjetiva, os positivistas defendiam que a obra deveria escapar aos caprichos do leitor e sustentar-se por si mesma e o mais objetivamente possível. Entretanto, ainda que aparentemente antagônicas, ambas as posições sustentam que o exercício da leitura atenta e passional (impressionista) e da leitura objetiva e disciplinada (positivista) levam a uma interpretação fiel dos textos literários.

Outra corrente de pensamento que proclamava a autossuficiência da obra literária foi a Neocrítica, surgida nos Estados Unidos na década de 20. Para eles, a leitura deveria ser fechada (*close reading*), descritiva, sem levar em conta a produção da obra nem a sua recepção. I. A. Richards, um dos fundadores do movimento neocrítico, tinha a convicção de que os obstáculos que se impunham entre a obra e seu efeito poderiam ser eliminados por meio da educação do leitor. Educado rigorosamente, o leitor teria acesso a uma "compreensão plena e perfeita" da obra literária e corrigiria os erros comuns de leitura. Conforme Compagnon, a opinião de Richards é de que o problema está com o leitor, é ele quem tem "limitações individuais e culturais" que levam a leitura "a fracassar diante do texto" (op. cit., 1999, pp. 142-143). De certa forma, os neocríticos viam o sentido na obra literária como algo a ser dissecado e extraído em laboratório.

Para o pensamento estruturalista do pós-guerra, o leitor, quando chega a ser considerado, é também um "intruso" (op. cit., 1999, p. 142), e uma noção de que o leitor é na verdade uma função do texto afasta o leitor real daquilo a que se denominou *arquileitor*, um leitor omnisciente, ideal. Compagnon (1999, p. 143) afirma o seguinte:

> A leitura real é negligenciada em proveito de uma teoria da leitura, isto é, da definição de um leitor competente ou ideal, o leitor que pede o texto e que se curva à expectativa do texto. (...) Assim, a desconfiança em relação ao leitor é

> (...) uma atitude amplamente compartilhada nos estudos literários.

É na esteira dos estudos da hermenêutica fenomenológica (representados, por exemplo, por Sartre, entre outros) que posições teóricas de revalorização da leitura surgem, como a Estética da Recepção. É com a Estética da Recepção que a ideia de leitor começa a se separar da ideia de autor, uma vez que, até então, o leitor só seria levado a sério se se tornasse, ele próprio, autor – com *autoridade* para comentar a obra de outros autores. Além disso, essa nova perspectiva sobre o leitor buscou compreender o impacto da obra literária em termos de uma leitura não só individual como também coletiva. Interessados na impressão causada pela obra literária no leitor, duas correntes de estudo se formaram: uma focada na fenomenologia da leitura individual (com Roman Ingarden e Wolfgang Iser, inicialmente), e outra voltada à resposta coletiva ao texto (principalmente com Gadamer e Hans Robert Jauss).

Iser (1995, p. 149) analisa o processo de leitura como a realização do efeito potencial do texto. A obra literária teria dois polos complementares: o polo artístico, que é o texto do autor[18], e o polo estético, que é a leitura. O sentido, ou o efeito da obra, não é um objeto definido e não existe *antes* da leitura. Para Iser (1995, p. 54), a obra literária não é o texto nem a experiência subjetiva, mas "o esquema virtual feito de lacunas e de indeterminações" e a parceria entre o leitor e o texto no processo comunicativo da leitura. É a partir dessas ideias que Iser propõe a noção de "leitor implícito", que constitui o papel assumido pelo leitor real diante das instruções fornecidas pelo texto. É a suposição de que a obra contém em si a expectativa de um leitor – o que significa dizer que o leitor, nessa visão, é uma estrutura textual e que a leitura é um ato estruturado (COMPAGNON, 1999, p. 151). Mas o leitor implícito deve, além disso, trazer para a leitura o que Iser denomina "repertório", ou seja, suas experiências prévias e as normas sociais de seu tempo, e esse repertório, para que a leitura se realize, deve fazer intersecção com o repertório da obra. Assim, o leitor implícito parece ser também um leitor idealizado, que só sentirá o efeito do texto se tiver características específicas que o qualifiquem a tanto.

[18] Há muitas teorias e concepções acerca da problemática em torno do autor literário; ver, por exemplo, Foucault (1994), cuja noção de autor é a de "autor-função", sem relação com o indivíduo real, mas com o prestígio e a consagração literária conquistados pelo autor, um "constructo discursivo" cujas qualidades tendem a ser repetidas por outros autores e constitui um gênero textual em si mesmo.

O esquema de Iser baseia-se em grande medida nos romances realistas do século XIX e está atrelado a uma "escola literária" e à expectativa de leitura dos tipos de obras pertencentes a essa escola literária específica (COMPAGNON, 1999, p. 154).

Na visão de leitura como experiência coletiva de recepção, Hans Robert Jauss propõe o conceito de *horizonte de perspectiva*, que seria, segundo Compagnon (1999, p.156), o equivalente à ideia de *repertório* de Iser: "o conjunto de convenções que constituem a competência de um leitor". Mas leitor e leitura ganham destaque no palco dos estudos de literatura, aprofundando a dimensão coletiva da experiência literária, no radicalismo das ideias de Stanley Fish (1980), que chega a afirmar que "literatura é o que acontece quando lemos", desconstruindo o estatuto científico dos estudos literários ao tomar seu objeto como fenômeno.

A crítica literária de tradição formalista pressupõe que o texto contém em si significados que o leitor deve compreender, conforme afirma Fish (1976). Para ele, depender da consulta a dicionários e gramáticas é pressupor que os sentidos podem ser especificados independentemente da atividade de leitura. É a experiência de leitura do leitor, e não as estruturas dos textos, que deveria ser o objeto de investigação. Fish contrapõe-se à postura formalista que considera o texto autossuficiente, uma visão predominante também em relação aos estudos de tradução. A questão da temporalidade, então, ganha importância quando se centra a análise no processo interpretativo engendrado pelo leitor, uma vez que épocas diferentes produzirão sentidos diferentes.

Além disso, Fish (1976) caracteriza o ato de leitura como uma atividade não só de produção de sentidos, mas de determinação de intenções. O leitor busca a intenção do autor, e, se o universo de conhecimentos de ambos coincidir, o leitor produzirá significados condizentes com essa suposta intenção, que é, ela mesma, fruto do processo de criação de sentidos promovido pela leitura. Esse leitor seria, na análise de Fish, o leitor ideal, ou seja, um leitor cuja vivência, competência linguística e compreensão de mundo lhe permitam acesso à experiência que o autor quis proporcionar.

Percebe-se que, para Fish, interpretação cria intenção. Assim, não é o texto que precede o leitor, mas as estratégias de leitura que precedem os textos. Tais estratégias não são nem naturais, nem universais, são aprendidas, o que confere às comunidades interpretativas um caráter essencialmente instável e mutável. Aqui, Fish não se refere à

educação formal do leitor, como Richards e Iser (COMPAGNON, 1999), mas sim à aprendizagem de estratégias por meio do convívio social da comunidade interpretativa à qual o leitor pertence. Outro elemento importante das ideias de Fish é a natureza impermanente das leituras possíveis de um texto em função das mudanças nas estratégias de leitura das comunidades interpretativas. As leituras de um texto são as leituras possíveis em uma determinada época por um determinado grupo. Sartre, em *O que é literatura?* (2004, p. 56), antecipou a importância da coletividade e da convergência das experiências compartilhadas entre autor e leitor no efeito da obra:

> Os indivíduos de uma mesma época e de uma mesma coletividade, que viveram os mesmos eventos, que se colocam ou eludem às mesmas questões, têm um mesmo gosto na boca, têm uns com os outros a mesma cumplicidade e há entre eles os mesmos cadáveres.

O que parece ser consenso é o fato de que o autor encontrará mais condições de efeito e impacto de sua obra entre leitores da sua própria comunidade do que entre leitores que não compartilham das mesmas vivências culturais e históricas do autor do texto. No entanto, isso levanta uma questão da máxima relevância e, de certo modo, paradoxal: sem entrar na discussão sobre o que torna uma obra um clássico, mas considerando uma obra cujo status seja o de cânone literário, como, por exemplo, o *corpus* homérico, de que forma poderíamos equacionar seu efeito (entendido aqui como produção de sentido) no leitor contemporâneo, que muito provavelmente não conhece nem sequer o alfabeto grego? E não precisaríamos voltar tão longe no tempo. Consideremos os cânones da literatura brasileira do fim do século XIX e do início do século XX: qual é a intersecção entre a experiência relatada na ficção dos autores desse período com a experiência cultural do leitor de hoje?

Essas questões são especialmente relevantes quando o assunto é tradução, que herda os conflitos e incertezas dos Estudos Literários e da própria Linguística. A questão da fidelidade do tradutor ao texto-fonte – que é, no fundo, uma expectativa de que há uma "leitura correta" –, a busca por equivalências lexicais perfeitas entre línguas e a utopia da invisibilidade do tradutor ilustram alguns dilemas herdados pelos estudos de tradução. Além disso, traz à baila a posição do tradutor, antes de mais nada, como *leitor*.

2.1.5 ESTUDOS DE TRADUÇÃO E LEITURA

Partindo do pressuposto de que um texto não contém em si significados preexistentes à leitura e de que o ato da leitura é tanto produção de sentidos (e não do sentido) quanto atividade interpretativa (FISH, 1976; KOCH, 2007), pode-se considerar a prática da tradução uma atividade essencialmente criativa (RODRIGUES, 2000). Em outras palavras, a tradução pode ser entendida como a atividade interpretativa de um leitor-tradutor a fim de produzir significados aceitáveis para uma comunidade leitora. Nessa perspectiva, o texto na língua de chegada passa a ter a importância e receber a atenção normalmente dispensada ao texto-fonte, que tradicionalmente é tido como um texto fechado e com significados a serem "extraídos" pelos leitores, numa visão elitista que pressupõe uma espécie de monopólio interpretativo daqueles que se proclamam especialistas seja no autor, no assunto ou na obra. Essa postura estabelece que há um significado hegemônico desvinculado de uma prática livre de leitura aberta a múltiplas significações.

Isso é relevante sobretudo para tradutores que se aventuram nos mares turbulentos da tradução literária, onde costumam encontrar inúmeras dificuldades e incontáveis situações diante das quais lhes é exigido um posicionamento em relação ao texto. Além de ver-se perdido entre teorias de tradução irremediavelmente antagônicas, o tradutor vê-se também cobrado pelo mercado de trabalho, que lhe impõe prazos e condições nem sempre razoáveis quando se considera o esforço de pesquisa dispendido. Há também a crença subjacente de que a norma culta é a única aceitável em traduções literárias.

Um dos dilemas com que o tradutor comumente se depara é a tradução de uma obra literária canônica. Com a tarefa de estabelecer possibilidades de sentido entre uma obra literária consagrada e uma comunidade interpretativa, é imprescindível que o tradutor se arme não só de uma fundamentação teórica que o auxilie na abordagem ao texto e na escolha de estratégias, mas também de ferramentas que o guiem e que lhe forneçam dados a partir dos quais ele possa manipular a estratégia tradutória escolhida, sem mencionar o fato de que a concepção de tradução corrente no mercado é a noção leiga de que traduzir é transportar significados de uma língua para outra.

Considerando-se que o tradutor ocupa a posição do autor ao definir estratégias tradutórias, conforme afirma Rodrigues (2000), é vital que, em se tratando de um texto literário dirigido a uma comunidade interpretativa específica (FISH, 1976), seu ponto de referência seja o leitor-modelo pertencente a essa comunidade. Além disso, o tradutor é também o primeiro leitor do texto (tanto o texto-fonte quanto o texto de chegada pré-revisão), e as operações interpretativas que ele coloca em ação dependem em larga escala do perfil dos leitores a quem o texto traduzido se dirige. Em outras palavras, é útil ao tradutor conhecer as características dos leitores a fim de adequar e ajustar suas escolhas de acordo com eles.

Susan Bassnett, em *Estudos de Tradução* (2005), comenta os trabalhos e posturas dos principais teóricos da tradução desde os estudos dos primeiros gramáticos até os estudos modernos do século XX. A autora aponta que só recentemente os estudos de tradução passaram a ser uma disciplina de fato, tendo em vista o status historicamente baixo que os críticos literários, em especial, atribuíram às traduções e aos estudos que delas se ocuparam. A tradução tem sido vista como uma área submissa a outras, julgadas maiores e mais dignas de atenção, como as teorias literárias e linguísticas tradicionais, e o tradutor, nessa lógica, costuma ser visto como um autor desqualificado cujas traduções nunca estarão à altura dos textos "originais".

A maior ironia, segundo Bassnett (2005), dos debates em torno da tradução é

> "que os mesmos especialistas que rejeitam a necessidade de investigar a tradução cientificamente por causa de seu status tradicionalmente baixo no mundo acadêmico não hesitam em ensinar um número considerável de textos traduzidos a estudantes monolíngues." (p. 26)

Essa postura revela que, por mais que se negue à tradução a posição de objeto digno de investigação científica e que se insista em considerá-la uma prática "inferior", a tradução continuará a ser necessária e continuará a existir no plano da "vida real", que parece ter sido deixado de lado por alguns setores da academia. É evidente que, em se tratando de textos de partida de caráter literário, o debate sobre a tradução sempre andará de mãos dadas com o debate em torno da arte, e, na verdade, o debate sobre tradução beneficia-se de toda e qualquer discussão que envolva comunicação, cultura, linguística e literatura. No entanto, é preciso também que os estudos de tradução saiam da

obscuridade e rompam com as posições tradicionais que os relegam a uma subcategoria desprezada dos estudos da linguagem.

Assim, dessa necessidade, surgem posições que desafiam as noções tradicionais, como Rodrigues (2000), citando Rosemary Arrojo, afirma:

> "É impossível encontrarmos 'um nível de apreensão neutra de significados, que possa ocorrer fora de um contexto e independentemente da interferência de um sujeito. (...) Assim, a compreensão, num plano humano e não-divino, será, sempre, também interpretação, uma produção – e não um resgate – de significados que impomos aos objetos, à realidade e aos textos.'" (p. 213)

Logo, na visão de Rodrigues (2000) e Arrojo, a tradução, agora não mais vista como a tentativa de buscar equivalências e transferir significados, ganha a autonomia de uma atividade que privilegia a produção de sentidos a partir da leitura feita por um tradutor, que é, ele mesmo, um leitor que recria um texto de língua estrangeira na língua de chegada para leitores monolíngues que, não fosse o trabalho do tradutor, não acessariam aquele texto.

A abordagem funcionalista de tradução de Reiss e Vermeer (1984) propõe uma autonomia semelhante ao tradutor, que é quem julga o "escopo" da tradução. Nessa visão, a tradução deve ter uma finalidade, deve apresentar coesão interna para a comunidade interpretativa a que se destina e deve manter coerência com o texto-fonte. No que diz respeito à avaliação da qualidade da tradução, o fator preponderante são as regras da comunidade interpretativa, e não da língua de partida, e a função do texto traduzido dentro dessa comunidade. A noção de "função", no entanto, é bastante difusa e pouco clara, uma vez que estabelecer o impacto de uma tradução é também estabelecer o perfil linguístico e cultural da comunidade interpretativa à qual a tradução se dirige, com nuances sociológicas que fogem à competência do tradutor. Outra crítica a essa abordagem é o foco no texto de chegada, em detrimento do texto de partida (HURTADO ALBIR, 2008).

Christiane Nord (2006), levando adiante a perspectiva funcionalista de Reiss e Vermeer, considera o tradutor apenas mais um leitor entre outros leitores de um texto e que sua tradução será também apenas uma entre outras tantas possíveis. Além disso, o tradutor nem sequer costuma fazer parte do público pretendido pelo texto-fonte, o que o

coloca numa posição pouco privilegiada para decidir o que é relevante para o leitor pretendido pela tradução. Nord então propõe uma visão de tradução como atividade essencialmente comunicativa: "Toda tradução tem a intenção de atingir um objetivo comunicativo específico no público-alvo[19]" (2006, p. 133). Com isso, a autora dá ênfase ao texto-final, e não ao texto-fonte. Essa inversão implica analisar a função do texto e o perfil dos leitores a quem o texto se destina, suas expectativas e necessidades, a fim de que a tradução preencha esses requisitos.

Nord (2006) classifica as funções comunicativas em quatro grupos, seguindo o modelo de Jakobson: função representativa, função expressiva, função apelativa e função fática. A função referencial envolve referência a objetos e fenômenos da realidade e pode ser analisada de acordo com a natureza do objeto em questão. Se o referente é desconhecido pelo receptor, a função do texto pode ser a de descrever o objeto; se for uma língua ou um uso específico de língua, a função pode ser metalinguística; se o referente for o uso de um eletrodoméstico, a função pode ser instrutiva. A função referencial depende da compreensibilidade do texto, a qual é presumida a partir das características do público-alvo. Na função expressiva, o foco está no emissor, em suas atitudes e opiniões a respeito do referente (por exemplo, interjeições). Na função apelativa, o foco está em persuadir o receptor (por exemplo, uso de imperativos). A função fática tem o objetivo de abrir e fechar o canal de comunicação entre emissor e receptor e tem características pragmáticas (por exemplo, modos de tratamento).

Em resumo, de acordo com Nord (2006), o propósito de uma tradução determina a escolha das estratégias e modelos de tradução, e o objetivo do tradutor é produzir um texto cujas funções sejam reconhecidas pelo leitor. E a função do texto-fonte pode ser diferente da função do texto-alvo, contanto que a intenção comunicativa de ambos os textos não seja incompatível.

Com exceção da função fática, a classificação de Nord corresponde à classificação tipológica proposta por Reiss, citada por Hurtado Albir (2008, p. 474). A função representativa concentra-se no conteúdo; a função expressiva, na forma; e a função apelativa, na persuasão. Cada uma dessas funções caracteriza tipo textuais

[19] No artigo em inglês: "Every translation is intended to achieve a particular communicative purpose in the target audience."

específicos, com características específicas, que devem ser levados em conta pelo tradutor.

Na prática, estabelecer a função de um texto equivale a determinar a que registro, gênero e estilo a tradução se conforma, ainda que, em se tratando de traduções literárias, possam ser identificados múltiplos gêneros, estilos e registros em um único texto (BIBER & CONRAD, 2009). Além disso, nem sempre o gênero do texto de partida será o mesmo no texto de chegada, como, por exemplo, no caso dos textos homéricos, hoje traduzidos em prosa, e não em versos – o que é reflexo também das mudanças operadas nas comunidades interpretativas (FISH, 1980).

2.1.5.1 TRADUZIR OU ADAPTAR?

Definir adaptação é, também, definir tradução. Afirmar que um texto é uma "adaptação" presume que exista uma "tradução" mais legítima desse texto, mas "fiel" ou mais adequada (por exemplo, MILTON, 2009). Tratando especificamente de obras de caráter estético, Eco (2007) evita usar o termo "adaptação", usando-o como sinônimo de "transmutação". Para Eco, uma transmutação ocorre quando há uma mudança semiótica na tradução, e uma "reelaboração" ocorre quando o tradutor toma licenças radicais que não permitem que se reconheça o texto fonte: "Se uma máquina tradutora qualquer, mesmo que de modo perfeito, vertesse novamente o texto de destino para outro texto da língua-fonte, seria difícil reconhecer o original" (op. cit., p. 353). Assim, Eco considera uma tradução uma "adaptação" quando há mudança semiótica (por exemplo, uma novela publicada em forma de livro levada para as telas de cinema), e reelaboração quando o texto torna-se irreconhecível na língua de chegada quando comparado com o texto-fonte.

Estabelecer o que constitui e caracteriza uma adaptação é uma das eternas indefinições em estudos do texto no campo da Tradutologia. As opiniões e categorizações são tão diversas quanto as teorias e visões de tradução em que se fundamentam. Tendo em conta a perspectiva teórica de tradução adotada neste trabalho, a funcionalista (VERMEER e REISS, 1984; NORD, 2006), a questão da classificação dos tipos e gêneros textuais é importante para determinar se um texto literário fruto de tradução é uma adaptação ou não, ainda que não seja possível estabelecer essa distinção clara e definitivamente para todos os tipos, gêneros e registros textuais existentes.

O estudo de gêneros textuais é heterogêneo e, nas palavras de Marcuschi (2008), está "na moda". Sem a pretensão de cobrir todas as dimensões dos estudos sobre gêneros, Marcuschi faz o seguinte esquema das principais vertentes teóricas:

- Perspectiva sócio-histórica e dialógica, influenciada por Bakhtin.
- Perspectiva comunicativa de Steger, Güllich, Bergmann e Berkenkotter.
- Perspectiva sistêmico-funcional, a partir da teoria sistêmico-funcionalista de Halliday.
- Perspectiva sociorretórica e etnográfica no ensino de língua adicional, influenciada pelos estudos de John Swales e Vijay K. Bhatia.
- Perspectiva interacionista e sociodiscursiva de natura psicolinguística com ênfase no ensino de língua materna; influenciada por Bronckart, Dolz e Schneuwly.
- Perspectiva de análise crítica, de Fairclough e Kress.
- Perspectiva sociorretórica/ sócio-histórica e cultural, influenciada por Bakhtin e pesquisadores de outras áreas do conhecimento, como antropólogos, sociólogos e etnógrafos.

Poderíamos acrescentar também a perspectiva cognitiva de Graesser, Gernsbacher e Goldman (1997), com uma proposta de modelos cognitivos de compreensão de gêneros textuais a partir de inferências de leitura. Graesser e colaboradores (2004) apontam para interações "intrigantes" entre a tessitura coesiva de um texto e o conhecimento de mundo do leitor ao construir e usar modelos mentais subjacentes a, por exemplo, textos científicos. Leitores com menos conhecimento prévio a respeito da área em questão beneficiam-se de textos com maior coesão, ao passo que leitores conhecedores do assunto tratado no texto beneficiam-se mais de textos menos coesos. Uma menor coesão textual permite que o leitor que domina o assunto faça inferências e, consequentemente, estabeleça mais conexões entre as ideias do texto e o seu conhecimento sobre o assunto. Esse processo resulta em uma representação mental mais coerente e sugere que nem sempre um texto com coesão homogênea é o texto ideal para todos os tipos de leitores.

Os estudos de gênero, como regra geral, abrangem o texto literário, mas poucas pesquisas em linguística concentram-se exclusivamente nele. Tendo em vista que os estudos de literatura têm sua própria concepção de gênero, as pesquisas linguísticas

tendem a evitar, com propriedade, uma apreciação estética do texto literário. Fazem, no mais das vezes, apenas comentários sobre estilo ou classificações genéricas a fim de diferenciar gêneros "não literários" e "não ficcionais" (por exemplo, BIBER, 2009; BEGTHOL, 2001; BHATIA, 1993; SWALES, 1990), situando o texto literário no domínio cuja instância discursiva é a mimese ou a representação ficcional da realidade (por exemplo, BAKHTIN, 2010; MARCUSCHI, 2008).

No entanto, em Tradutologia, caracterizar gêneros textuais é apenas parte do processo de escolhas de estratégias realizado pelo tradutor. Reiss, citada por Hurtado Albir (2008, p. 475), propõe uma classificação textual baseada nas três funções da linguagem: representativa, expressiva e apelativa. Distingue três dimensões textuais (com predomínio do conteúdo, com predomínio da forma ou com predomínio da persuasão), aos quais correspondem várias classes de textos classificadas de acordo com as características ou convenções linguísticas. Tais classes correspondem ao que chamamos *gêneros* textuais. Além disso, acrescenta a noção de modalidade textual, que corresponde à distinção semiótica entre textos de sistemas diferentes (texto escrito, oral, etc.), e a noção de âmbito textual, que corresponde a pelo menos um traço distintivo em comum entre os textos (textos de ficção, textos técnicos, poéticos, etc.).

Nesse sentido, seguindo postulados e perspectivas funcionalistas em Tradutologia, tanto os textos-fonte quanto os textos traduzidos analisados neste trabalho podem ser classificados da seguinte forma:

- Função: expressiva.
- Dimensão: estética.
- Modalidade semiótica: escrita.
- Âmbito: ficção.
- Classe ou gênero textual: conto literário.

No processo tradutório do conjunto de textos analisado nesta dissertação, não houve alteração de função, dimensão, modalidade, âmbito e gênero entre os textos, o que, a meu ver, sugere fortemente que os textos traduzidos não são adaptações, mas traduções. Milton (2009) comenta o caso da tradução de livros infantis, questionando se seriam adaptações ou não. Nesse caso, o projeto editorial típico voltado para o público infantil envolve alterações semióticas, pois acrescentam-se imagens ao texto; logo, livros infantis

podem ser considerados adaptações, caso o texto-fonte não seja dirigido a esse público. Quanto ao critério da reconhecibilidade de Eco, apesar de subjetivo, provavelmente é um parâmetro aceitável, porém dificilmente mensurável.

Uma proposta de tradução que leve em conta a proficiência de leitura dos leitores a quem o texto final se destina tampouco altera a natureza do texto produzido em língua-alvo: sua função permanece a mesma, assim como a sua dimensão estética, a sua natureza semiótica, o seu âmbito ficcional e o seu gênero literário. Trata-se, considerando-se o exposto acima, de uma visão de tradução em que o texto em língua de chegada tem como parâmetro o leitor, sem deixar de lado as características essenciais do texto de partida. Assim, ao propor uma reflexão sobre medidas de avaliação de complexidade textual entre originais e traduções, entendemos que a simplificação de um texto em tradução que se traduz pode ser parte integrante do processo tradutório.

3. POSICIONAMENTO DO TRABALHO

Como previamente mencionado, durante o meu trabalho como revisora e tradutora de textos literários, percebi um fenômeno difícil de nomear e localizar objetivamente nos textos. Na tentativa de descrever e entender esse fenômeno, empreendi o trabalho narrado nesta dissertação. O primeiro desafio foi delimitar o terreno teórico, em vista da multiplicidade de abordagens viáveis. As seções anteriores situaram teoricamente as ideias e estudos que contribuíram para a construção desta dissertação, as quais são bastante amplas e recebem influências de várias áreas – o que é natural, em se tratando de um estudo cujo objeto é o texto em tradução. Retomemos, agora, algumas dessas ideias.

Entre os **Estudos de Tradução**, a escola funcionalista (NORD, 2006; HURTADO ALBIR, 2008) propõe que a tradução deve ter uma finalidade, apresentando coesão interna para a comunidade a que se destina e mantendo coerência com o texto-fonte. Nessa perspectiva, no que diz respeito à avaliação da qualidade da tradução, um fator preponderante são as regras e necessidades da comunidade leitora, e não as da língua de partida, e a função do texto traduzido dentro dessa comunidade. A noção de função, nessa ótica, está ligada principalmente ao perfil linguístico e cultural da comunidade à qual a tradução se destina.

O tradutor é o responsável pela ressignificação do texto-fonte em outro sistema linguístico, e é nessa medida que o tradutor assume o papel de "autor" do texto de chegada: é ele quem deve acomodar a transversalidade semântica da língua do texto em língua-fonte ao sistema da língua-alvo. E uma das atribuições do tradutor-autor é definir a quem a tradução se destina a fim de priorizar os elementos textuais relevantes para aquele público e para aquela situação comunicativa específica. É nesse sentido que a determinação das características de tipologia e gênero textual são importantes, pois norteiam o tradutor em suas estratégias de tradução do texto que tem em mãos.

Partindo de Saussure (2004), com uma concepção de língua como um sistema de signos que se definem por serem o que os outros não são, **chegamos a Bouquet** (2009) e sua visão algébrica de língua, em que todos os termos são complexos e em que o fenômeno do sentido dá-se pela transversalidade semântica. O fenômeno do sentido não

se dá de forma isolada, mas em todas as instâncias da língua. Esse ponto de partida é fundamental, pois, não só firma o alicerce de **uma visão semântica de tradução**, como determina a impossibilidade de uma tradução sistêmica, ou seja, não se traduz o sistema de uma língua, mas os sentidos dela. Além disso, uma pesquisa com enfoque lexical é aprofundada com essa visão semântica transversal, pois tende a enriquecer o plano lexical.

Com os estudos do léxico em **Linguística de Corpus**, beneficiados pelo crescente aparato tecnológico moderno, essas instâncias de sentido são analisadas em *corpora* reais a partir de uma visão empírica e probabilística de língua, em que os "termos complexos" saussurianos estão explicitamente expostos na língua em uso com frequências distintas de acordo com uma multiplicidade de fatores. Com a AMD de Biber, traços globais e genéricos de usos específicos da língua em diferentes registros e gêneros são agrupados, partindo da microdimensão do texto para a macrodimensão do *corpus*, mantendo a identidade e a unidade textual em meio a um *corpus*. Apesar de a AMD não ser a metodologia escolhida para esta pesquisa, a ideia de que é possível manter a percepção da individualidade de cada texto dentro de um macroconjunto com traços distintivos e comparáveis foi uma contribuição importante para o trabalho com *corpus*.

Já a contribuição deste trabalho **ao PLN** se dá de duas formas. Em primeiro lugar, nesta pesquisa testaram-se ferramentas geradas com recursos de PLN para fins diferentes dos originalmente pretendidos pelas ferramentas em questão (Coh-Metrix e Coh-Metrix-Port), as quais não têm a finalidade de comparar textos traduzidos. Em segundo lugar, esta pesquisa propõe uma avaliação de complexidade textual de textos-fonte e de suas traduções sob a ótica de uma profissional do texto, tradutora e revisora, partindo de problemas reais, encontrados na prática do ofício, e não de abstrações acadêmicas. O inverso também é verdadeiro: foi através do contato e da troca com pesquisadores de PLN que a abordagem por técnicas de aprendizagem de máquina foi introduzida neste trabalho, contribuição fundamental a esta dissertação.

Uma vez que o objetivo é comparar textos completos, e considerando a minha experiência de trabalho com textos literários, optar pelo gênero "conto de ficção" foi uma decisão natural: trata-se de textos integrais, com início, meio e fim, e também de um

gênero popular entre os leitores brasileiros. Além disso, as ferramentas usadas (Coh-Metrix e Coh-Metrix-Port) impuseram um limite operacional de 15 mil caracteres, como será visto mais adiante (ver capítulos 4 e 6), o que restringiu o tamanho dos textos a selecionar para a composição do *corpus* de estudo.

As **pesquisas sobre letramento** mostram a situação do contingente de leitores no Brasil: a maioria dos leitores tem proficiência de leitura em nível básico ou rudimentar, mesmo leitores com escolaridade alta (56% dos leitores com Ensino Médio completo têm letramento básico, 31% dos leitores com curso superior têm proficiência de leitura em nível básico; ver Tabela 1). Esses números sugerem a possibilidade de que esses leitores não estejam compreendendo plenamente os materiais de leitura a que são expostos. Fish (DUBAY, 2004) mostrou que o interesse de leitura diminui caso o leitor julgue o texto difícil, o que equivale a dizer que um texto considerado fácil é também um texto potencialmente mais lido. Por outro lado, independentemente do nível de letramento do leitor, a hipótese aqui aventada é a de que as traduções para o português brasileiro são mais complexas do que os textos-fonte em inglês. Colocam-se, assim, dois problemas: o baixo nível de letramento dos leitores e o maior nível de complexidade das traduções quando comparadas com seus textos de origem.

No entanto, os tradutores – assim como outros profissionais do texto, incluindo revisores – dispõem de poucos recursos que os auxiliem na tarefa de caracterizar linguisticamente os perfis das comunidades a que se dirigem, sabendo quais formulações textuais lhes seriam mais ou menos inteligíveis, restringindo-se ao que lhes ofereçam dicionários, gramáticas e sua intuição linguística. Carecem, portanto, de ferramentas que os auxiliem a analisar, como um todo, o texto original e a tradução feita. Nesse sentido, tratar de sistemas que contemplem medidas de inteligibilidade e aplicá-los a traduções e originais, ainda que haja uma série de limitações a superar, poderia representar uma importante contribuição para ajudar a ponderar-se sobre a maior ou menor acessibilidade de determinados grupos de leitores brasileiros a esses textos.

Vale lembrar ainda que não considero uma tradução voltada para um público leitor específico uma tradução "adaptada". Uma adaptação, segundo a lógica comunicativa e funcionalista da tipologia proposta por Reiss (HURTADO ALBIR, 2008), implicaria mudanças semióticas na tradução da obra, o que não é o caso aqui.

Além disso, reiteramos, o objetivo principal deste trabalho não é o de propor um modelo de tradução de para leitores com proficiência de leitura baixa, mas avaliar se o nível de complexidade das traduções dos textos selecionados para o português brasileiro é compatível com o nível de complexidade dos textos-fonte.

Sabemos que a dificuldade de um texto não é óbvia. Portanto, consciente de que dificuldade não é uma qualidade somente do texto, mas também do leitor e de suas proficiências, não pretendo estabelecer uma definição categórica de complexidade textual. A partir de um recorte que incide apenas sobre o que está concretamente posto de modo explícito em um texto ou conjunto de textos, este trabalho visa empreender uma comparação entre medidas de complexidade textual de textos-fonte e suas traduções obtidas a partir do uso de ferramentas computacionais, como as aqui utilizadas: as ferramentas Coh-Metrix e Coh-Metrix-Port.

4. MATERIAIS E MÉTODOS

Partindo da microperspectiva estrutural do texto, isto é, considerando apenas sua tessitura coesiva, a pesquisa empreendida aqui é um estudo quantitativo e qualitativo sobre métricas para estimação de complexidade textual em um pequeno *corpus*, dividido em dois blocos: bloco 1, 14 contos literários em inglês e suas traduções para o português brasileiro, totalizando 28 textos; e bloco 2, 14 contos da literatura brasileira e suas traduções para o inglês, totalizando 28 textos. As métricas referidas são provenientes das ferramentas Coh-Metrix e Coh-Metrix-Port.

4.1 COH-METRIX

Coh-Metrix (GRAESSER et al., 2004), que significa *cohesion metrics*, é uma ferramenta para análise de textos em inglês, disponível gratuitamente *on-line*. Elaborada por pesquisadores da Universidade de Memphis, nos Estados Unidos[20], tem como propósito calcular índices de coesão e coerência textual num amplo espectro de medidas lexicais, sintáticas, semânticas e referenciais a fim de indicar a adequação de um texto a seu público-alvo (a *demanda cognitiva* e a legibilidade do texto) e de apontar problemas textuais de ordem estrutural.

Até o momento, mais de 500 métricas estão disponíveis em uma versão restrita do Coh-Metrix. Dessas 500, apenas sessenta estão disponíveis na versão gratuita *on-line* no *site* do projeto. Para todas essas métricas, vários recursos e ferramentas de Processamento de Língua Natural são utilizados.

A versão livre Coh-Metrix 2.0 opera com índices que vão desde métricas simples (como contagem de palavras) até medidas mais complexas, envolvendo algoritmos de resolução anafórica. As sessenta métricas estão divididas em seis blocos que avaliam a complexidade de um texto a partir da mensuração dos seguintes elementos:

- Identificação geral e informação de referência, índices de inteligibilidade, palavras gerais e informação do texto, índices sintáticos, índices

[20] O site do projeto é < http://cohmetrix.memphis.edu/cohmetrixpr/index.html>. A documentação está totalmente disponível para o público, bem como o acesso à ferramenta.

referenciais e semânticos e dimensões do modelo de situações. Essa primeira classe corresponde às informações que referenciam o texto, como título, gênero entre outros.

- Índices de inteligibilidade calculados com as fórmulas *Flesch Reading Ease* e *Flesch Kincaid Grade Level*. Essas fórmulas consideram tamanho de sentença, número de palavras por sentença e número de palavras diferentes por sentença.
- Verificação de quatro subclasses: contagens básicas, frequências, concretude, hiperônimos.
- Verificação de cinco subclasses: constituintes, pronomes, tipos e *tokens*, conectivos, operadores lógicos e similaridade sintática de sentenças.
- Verificação de três subclasses: anáfora, correferência e análise semântica latente.
- Verificação de quatro subclasses: dimensão causal, dimensão intencional, dimensão temporal e dimensão espacial.

A Figura 3 (abaixo) apresenta a tela principal do Coh-Metrix.

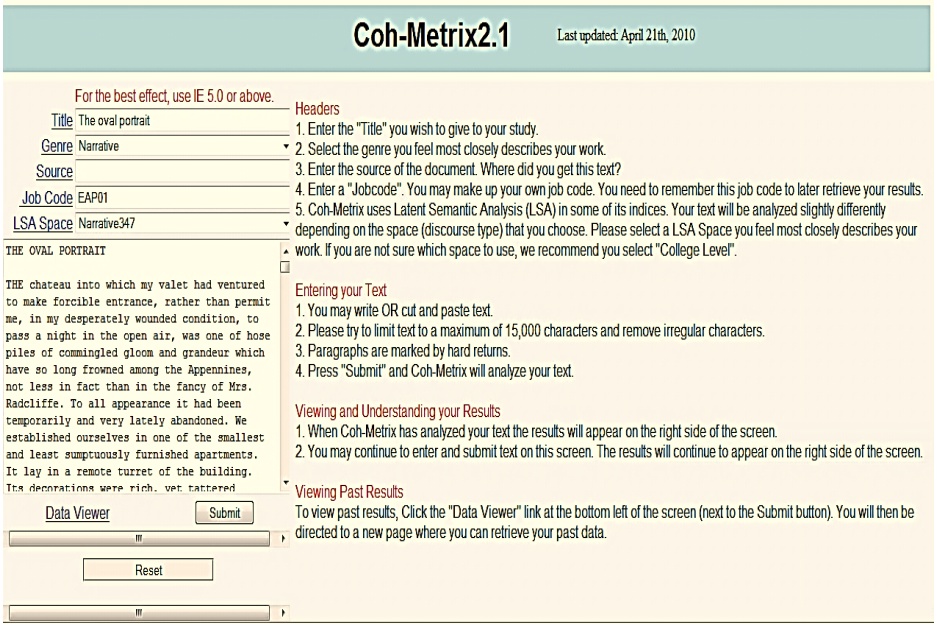

Figura 3. Interface do Coh-Metrix.

4.2 COH-METRIX-PORT

A partir do Coh-Metrix em inglês, no âmbito do Projeto PorSimples, surgiu uma iniciativa de adaptação para o português brasileiro das sessenta métricas oferecidas gratuitamente. O objetivo dessa iniciativa foi o de identificar índices de complexidade textual para simplificação de textos e facilitação do acesso à informação para analfabetos funcionais e para pessoas com deficiências cognitivas. O nome da ferramenta em português é Coh-Metrix-Port e está disponível no *site* do PorSimples[21].

A ferramenta Coh-Metrix-Port (SCARTON e ALUÍSIO, 2010) adapta o sistema para o português, contando com 48 métricas. Entretanto, nem todas as métricas de ambos os recursos podem ser comparadas, pois há métricas próprias de cada recurso e métricas incompatíveis devido aos recursos utilizados (*wordnet*, por exemplo). As métricas que puderam ser diretamente comparadas e, portanto, puderam ser utilizadas neste trabalho, são:

- Contagens básicas: número de palavras, número de sentenças, número de parágrafos, sentenças por parágrafos, palavras por sentenças e sílabas por palavras.
- O índice Flesch.
- Constituintes: incidência de sintagmas nominais, modificadores por sintagmas nominais e palavras antes de verbos principais.
- Conectivos: incidência de todos os conectivos, incidência de conectivos aditivos positivos, incidência de conectivos aditivos negativos, incidência de conectivos temporais positivos, incidência de conectivos temporais negativos, incidência de conectivos causais positivos, incidência de conectivos causais negativos, incidência de conectivos lógicos positivos e incidência de conectivos lógicos negativos.
- Operadores lógicos: incidência de operadores lógicos e número de negações.

[21] http://www.nilc.icmc.usp.br/cohmetrixport.

- Pronomes, tipos e *tokens*: incidência de pronomes pessoais, pronomes por sintagmas nominais e relação tipo/*token*.
- Correferências: sobreposição do argumento em sentenças adjacentes, sobreposição de argumento, sobreposição do radical de palavras em sentenças adjacentes, sobreposição do radical de palavras, sobreposição de palavras de conteúdo em sentenças adjacentes.
- Anáforas: referência anafórica em sentenças adjacentes e referência anafórica.

A Figura 4 adiante apresenta a interface do Coh-Metrix-Port.

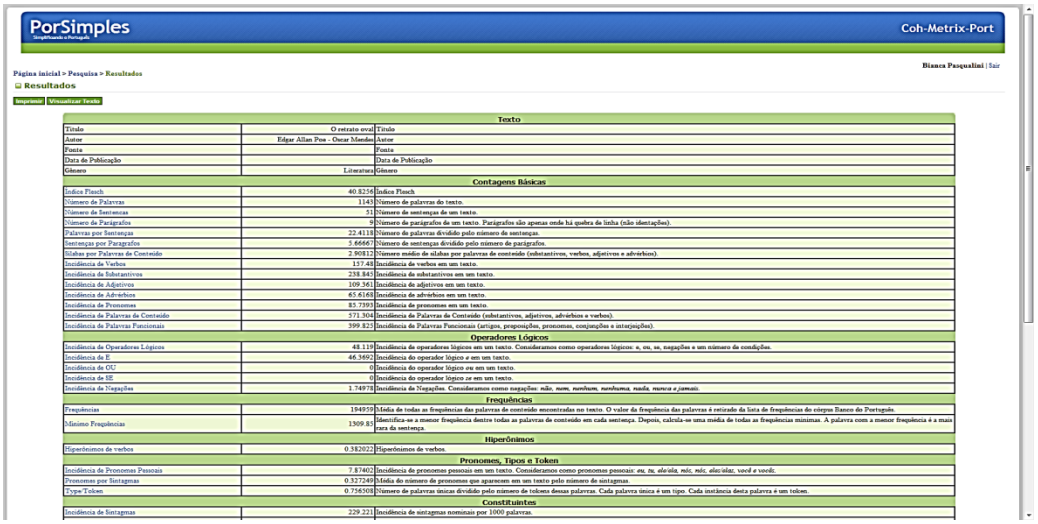

Figura 4. Interface do Coh-Metrix-Port.

Essas ferramentas, ainda que não tenham sido criadas com o intuito de serem usadas na análise de traduções ou comparações de textos, abrem um universo de possibilidades para os pesquisadores de Linguística Aplicada. Afinal, tratam de uma dimensão explorada entre nós de um modo bastante diferente, apresentando-a sob uma forma bastante objetiva.

Um item de destaque, nesse sistema de medidas, é o índice Flesch (ver capítulos 2, 6 e 7). É uma das diferentes medidas de complexidade do texto associada à sua inteligibilidade para diferentes tipos de leitores. O resultado é um número de 0 a 100 que é assim mensurado (com a devida adaptação para o sistema escolar brasileiro feita pela equipe PorSimples [MARTINS et al., 1996]):

- **Muito fáceis**: índice entre 90 a 100, textos adequados para leitores com nível de escolaridade até a 4ª. série do Ensino Fundamental.
- **Fáceis**: índice entre 80 a 89, textos adequados a alunos com escolaridade até a 8ª. série do ensino fundamental.
- **Razoavelmente fáceis**: índice entre 70 a 79, textos adequados a alunos com escolaridade até a 8ª. série do Ensino Fundamental.
- **Padrão**: índice entre 60 e 69, textos adequados a alunos com escolaridade até a 8ª. série do Ensino Fundamental.
- **Razoavelmente difíceis**: índice entre 50 a 59, textos adequados para alunos cursando o Ensino Médio ou universitário.
- **Difíceis**: índice entre 30 a 49, textos adequados para leitores com Ensino Médio ou universitário.
- **Muitos difíceis**: índice entre 0 a 29, textos adequados apenas para áreas acadêmicas específicas.

Como se verá adiante, no Capítulo 6, as métricas calculadas pelas ferramentas Coh-Metrix e Coh-Metrix-Port, por si só, não indicam o nível de complexidade de um texto. É na inter-relação entre as métricas que se encontra o melhor caminho na avaliação da complexidade. Ainda assim, é preciso mais uma vez lembrar que, nesta dissertação, o objetivo é *comparar* os resultados de textos originais aos de suas traduções (tanto na direção tradutória inglês-português quanto português-inglês) e analisar as *diferenças* entre os resultados obtidos. É também na relação estatística entre os resultados obtidos que se pode caracterizar cada conjunto de textos de acordo com as métricas que mais os discriminam. Para isso, usamos a ferramenta Weka, descrita a seguir.

4.3 WEKA

Como dito anteriormente, o método automático de classificação usado nesta pesquisa baseia-se no modelo estatístico supervisionado de AM, e a ferramenta usada é o Weka (Waikato Environment for Knowledge Analysis[22]). O Weka é uma coleção de

[22] O software é gratuito e está disponível, com toda a documentação, no site http://www.cs.waikato.ac.nz/ml/weka/.

algoritmos de AM que contém ferramentas para pré-processamento, classificação, regressão, agrupamento e associação de dados. O Weka opera a partir de arquivos em extensão ARFF (Attribute-Relation File Format), um arquivo de texto em ASCII que descreve uma lista de instâncias que compartilham um conjunto de atributos.

Arquivos com extensão ARFF têm duas seções distintas. A primeira é o cabeçalho (*header*), que contém o nome da relação a ser analisada, a lista de atributos e o tipo de atributo (se é numérico, nominal ou sequência de caracteres [*string*]); a segunda seção é composta pelos dados (*data*), com os valores de cada atributo listado.

O algoritmo escolhido foi a implementação J48 do algoritmo de classificação C4.5 para construção de árvores de decisão. A árvore de decisão mostra qual são as relações discriminativas entre os atributos (no caso, as métricas do Coh-Metrix e do Coh-Metrix-Port) do total das instâncias (cada um dos textos analisados) em cada classe (ou seja, classe de textos originais e classe de textos traduzidos). Em outras palavras, a estrutura em árvore de decisão mostra visualmente quais são as métricas estatisticamente mais características e distintivas em cada bloco de textos estudado, de acordo com a natureza do texto (a classe, na terminologia de AM): original em inglês, tradução para o português, original em português ou tradução para o inglês (ver Capítulo 6). A Figura 5 apresenta um exemplo de arquivo com extensão ARFF, e a Figura 6 mostra a interface principal do Weka.

```
@relation 'originais x traducoes'

@attribute title string
@attribute argument_overlap_adjacent numeric
@attribute stem_overlap_adjacent numeric
@attribute anaphor_reference_adjacent numeric
@attribute argument_overlap numeric
@attribute stem_overlap numeric
@attribute anaphor_reference numeric
@attribute noun_phrase_incidence_score numeric
@attribute ratio_of_pronouns_to_noun_phrases numeric
@attribute personal_pronoun_incidence_score numeric
@attribute number_of_paragraphs numeric
@attribute number_of_sentences numeric
@attribute number_of_words numeric
@attribute average_sentences_per_paragraph numeric
@attribute average_words_per_sentence numeric
@attribute average_syllables_per_word numeric
@attribute flesch_reading_ease_score numeric
@attribute mean_number_of_modifiers_per_noun_phrase numeric
@attribute mean_number_of_words_before_the_main_verb numeric
@attribute type_token_ratio numeric
@attribute proportion_of_content_words_that_overlap_between_adjacent_sentences numeric
@attribute class {ORIG,TRAD}

@data
text1-
poe_en,0.551,0.163,0.633,0.418,0.178,0.298,243.808,0.352,85.913,7,50,1292,7.143,25.84,1.413,
61.068,0.867,6.84,0.711,0.07,ORIG
text2-
poe_en,0.43,0.365,0.28,0.3325,0.245,0.1565,281.952,0.272,76.6135,38,108,1872.5,2.8435,17.341
5,1.586,55.0575,0.8025,4.0995,0.563,0.1015,ORIG
text3-
poe_en,0.5415,0.105,0.6355,0.5195,0.0945,0.332,278.5845,0.4015,111.879,13.5,91.5,1934.5,6.78
55,21.1435,1.473,60.7585,0.7705,4.29,0.732,0.0825,ORIG
text4-
poe_en,0.562,0.14,0.678,0.477,0.085,0.368,276.19,0.403,111.387,16,122,2415,7.625,19.795,1.55
7,55.021,0.727,4.066,0.707,0.108,ORIG
text5-
poe_en,0.51,0.134,0.618,0.446,0.069,0.453,292.556,0.468,137.032,16,158,2109,9.875,13.348,1.3
24,81.276,0.645,2.532,0.574,0.136,ORIG
text6-
poe_en,0.498,0.14,0.5355,0.413,0.1035,0.276,267.831,0.353,95.2565,12,66.5,1606,5.7785,24.419
5,1.5515,50.7925,0.8755,3.764,0.7415,0.0975,ORIG
text7-
poe_en,0.581,0.149,0.635,0.446,0.184,0.317,277.089,0.31,86.022,19,75,2418,3.947,32.24,1.425,
53.556,0.843,7.373,0.662,0.066,ORIG
text8-
poe_en,0.311,0.204,0.301,0.201,0.142,0.107,256.188,0.156,40.017,15,104,2424,6.933,23.308,1.4
09,63.976,0.969,4.481,0.654,0.066,ORIG
text9-
poe_en,0.321,0.04,0.438,0.261,0.034,0.263,320.924,0.479,153.616,90,250,2337,2.778,9.348,1.40
1,78.822,0.635,1.856,0.649,0.155,ORIG
text10-
```

Figura 5. Exemplo de arquivo ARFF.

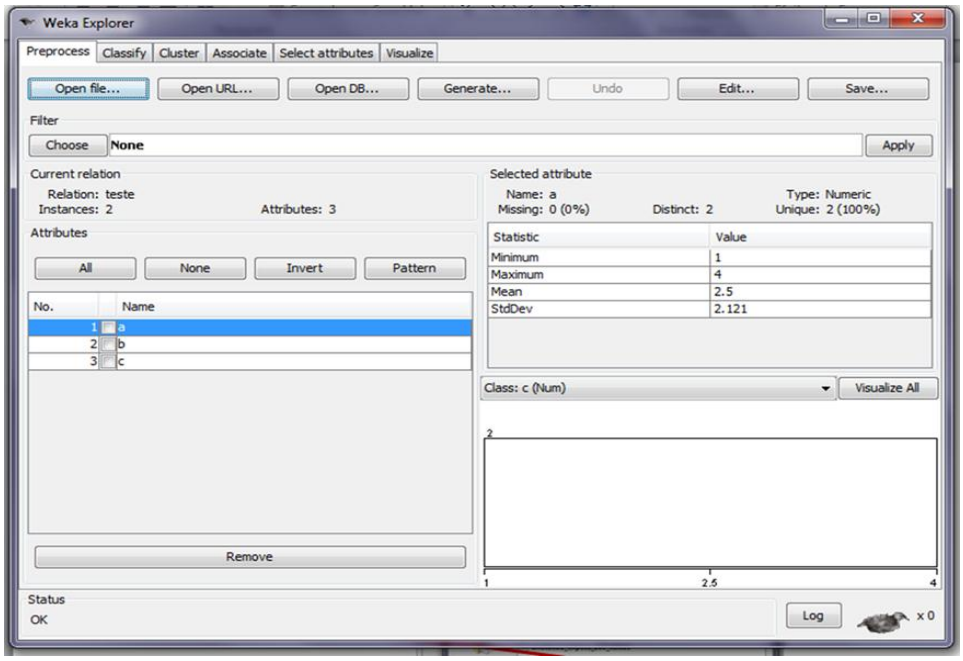

Figura 6. Interface do Weka.

4.4. *CORPUS*

Esta dissertação foi construída a partir da ampliação de trabalhos anteriores. Consequentemente, o *corpus* inicial foi o *corpus* norteador dos trabalhos posteriores. No estudo inicial (PASQUALINI, 2009), foi feita uma comparação de padrões de vocabulário entre um conto de Edgar Allan Poe ("The oval portrait") e duas traduções desse conto para o português brasileiro, tendo como parâmetro um leitor com pouca experiência de leitura, ou "neoleitor". A escolha do autor Edgar Allan Poe não foi aleatória. Como foi dito previamente, a motivação desta dissertação surgiu na minha prática profissional de tradução e revisão. Foi durante a tradução do conto "The mystery of Marie Rogêt", publicado na coletânea de contos de Edgar Allan Poe intitulada *O escaravelho de ouro & outras histórias*, obra revisada por mim para a editora L&PM[23], que as dúvidas e hipóteses levantadas nos trabalhos preliminares mencionados surgiram.

Em parceria com Finatto e Evers (PASQUALINI et al., 2010), decidimos ampliar o *corpus* e dividi-lo em três blocos, compostos por 10 textos cada um, totalizando 30 textos:

- **Bloco 1**: 5 contos de Edgar Allan Poe e suas respectivas traduções para o português brasileiro feitas por um mesmo tradutor brasileiro, totalizando 10 textos;
- **Bloco 2**: 5 contos da literatura brasileira e suas respectivas traduções para o inglês, feitas por um mesmo tradutor britânico, totalizando 10 textos;
- **Bloco 3**: 5 artigos científicos da área de Pediatria e suas respectivas traduções para o inglês, como contraponto de análise, totalizando 10 textos. A autoria individual das traduções não está indicada nos materiais, mas sua produção esteve a cargo da empresa brasileira *Scientific Linguagem*, prestadora de serviço para o periódico *Jornal de Pediatria*, publicação da Associação Brasileira de Pediatria.

E em parceria com Scarton e Finatto (PASQUALINI et al., 2011), optamos por eliminar o conjunto de textos científicos e concentrar a análise no contraste entre os

[23] POE, Edgar Allan. *O escaravelho de ouro & outras histórias*. Tradução de Bianca Pasqualini e Rodrigo Breunig. Porto Alegre: L&PM, 2011. (Tradução e revisão realizadas em 2009.)

textos do *corpus* de contos literários, focando na comparação entre a complexidade de originais e traduções de um mesmo tipo e gênero. Entretanto, acrescentamos novos textos ao *corpus*, que ficou composto da seguinte forma:

- **Bloco 1**: 14 contos literários em inglês e respectivas traduções para o português brasileiro, totalizando 28 textos.
- **Bloco 2**: 14 contos da literatura brasileira e respectivas traduções para o inglês, totalizando 28 textos.

A montagem do *corpus* e a coleta dos textos foram guiadas pelos seguintes critérios:

- Autoria: tendo Poe como referência, buscamos por autores de obras do mesmo gênero e com a mesma popularidade.
- Tamanho do conto: contos curtos, em virtude da limitação operacional de 15 mil caracteres da ferramenta Coh-Metrix em inglês.
- Data de publicação do texto original: estabelecemos o período inicial o das publicações de Poe, a partir de 1840, até o pós-guerra, entre 1945-1950, cobrindo aproximadamente cem anos.
- Data de publicação da tradução: a partir da década de 1980.
- Experiência do tradutor: desde novatos até profissionais (tanto no Bloco 1 quanto no Bloco 2).

O *corpus* estudado nesta dissertação é o mesmo do trabalho citado (2011). Equacionar todos esses elementos, alguns bastante subjetivos, sem dúvida foi um desafio no processo de montagem do *corpus* – sobretudo na montagem do *corpus* de contos da literatura brasileira traduzidos para o inglês, pois foi difícil encontrar traduções que se encaixassem nos critérios que buscávamos. Assim, o número de tradutores de contos para o inglês ficou menor do que o número de tradutores para o português. O maior problema, entretanto, foi o limite de 15 mil caracteres, pois é um tamanho muito reduzido. A solução foi aumentar para 30 mil caracteres o tamanho limite de cada texto e processar em duas partes, separadamente, os textos cuja extensão excedesse esse limite. Após o processamento em separado, os resultados foram somados individualmente e divididos por 2.

É importante deixar claro que a coleção textual analisada não se pretende exaustiva nem completa e que os resultados, as conclusões e as perspectivas suscitados nesta dissertação dizem respeito somente ao *corpus* estudado, e não se pretende estendê-los à totalidade de textos com as mesmas características (contos literários em tradução).

4.4.1 Bloco 1: Contos literários em inglês e respectivas traduções para o português brasileiro

O Bloco 1, de originais em inglês e respectivas traduções para o português do Brasil, é composto por 14 contos dos seguintes autores: Edgar Allan Poe (10); Nathaniel Hawthorne (01), O. Henry (01), Virginia Woolf (01) e James Joyce (01). Os tradutores dos contos de Edgar Allan Poe são: Marcelo Bueno (01), Oscar Mendes (04), Bernardo Carvalho (02), Celina Portocarrero (01), Rodrigo Breunig (01) e Dorothée de Bruchard (01). Os tradutores dos contos restantes são: Roberto Schmitt-Prym (01), Bianca Pasqualini (01) e Zaida Maldonado (01). Os textos originais têm uma média aproximada de 1.800 palavras (*tokens*) cada. (Ver tabelas 2 e 3.)

4.4.2 Bloco 2: Contos da literatura brasileira e respectivas traduções para o inglês

O Bloco 2, de originais em português e respectivas traduções para o inglês, é composto por 14 contos dos seguintes autores: Machado de Assis (06), Coelho Neto (02), Humberto de Campos (03) e Lima Barreto (03). Os tradutores são: Isaac Goldberg (02), Francis Johnson (10) e Gregory Rabassa (02). Os textos originais têm uma média aproximada de 1.600 palavras (*tokens*) cada. (Ver tabelas 4 e 5.)

BLOCO 1.1 – EDGAR A. POE E TRADUÇÕES PARA O PORTUGUÊS BRASILEIRO						
TÍTULO ORIGINAL	**NÚMERO DE CARACTERES**	**FONTE**	**TÍTULO TRADUÇÃO**	**NÚMERO DE CARACTERES**	**TRADUTOR**	**FONTE**
The oval portrait	7.204	Projeto Gutenberg	O retrato oval	7.081	Marcelo Bueno	http://www.bestiario.com.br/18_arquivos/poe.html
Mesmeric revelation	21.615 10.529 11.071		Revelação mesmeriana	22.262 11.058 11.143	Oscar Mendes	POE, EDGAR ALLAN POE. FICÇÃO COMPLETA. 7ª. ED. RIO DE JANEIRO: NOVA FRONTEIRA, 1981.
The black cat	21.749 10.798 10.610		O gato preto	21.749 10.991 10.645	Bernardo Carvalho	SÃO PAULO: COSAC & NAIFY, 2004.
The imp of the perverse	13.741		O demônio da impulsividade	14.135	Rodrigo Breunig	O ESCARAVELHO DE OURO & OUTRAS HISTÓRIAS (L&PM, 2011)
The tell-tale heart	11.147		O coração delator	11.780	Celina Portocarrero	HTTP://WWW.RELEITURAS.COM/EAPOE_CORACAO.ASP
Berenice	18.635 9.978 8.545		Berenice	20.300 9.887 10.227	Oscar Mendes	POE, EDGAR ALLAN POE. FICÇÃO COMPLETA. 7ª. ED. RIO DE JANEIRO: NOVA FRONTEIRA, 1981.
Eleonora	13.375		Eleonora	13.400		
The masque of the red death	13.683		A máscara da morte rubra	13.213		
The cask of amontillado	12.757		O barril de amontillado	12.800	Bernardo Carvalho	SÃO PAULO: COSAC & NAIFY, 2004
The Man of the Crowd	20.375 9.928 10.440	http://classiclit.about.com	O homem da multidão	20.697 10.165 10.532	Dorothée de Bruchard	HTTP://WWW.BESTIARIO.COM.BR/12_ARQUIVOS/O%20HOMEM%20DA%20MULTIDAO.HTML

Tabela 2. Bloco 1.1 – Contos de Edgar A. Poe e traduções para o português brasileiro.

BLOCO 1.2 – CONTOS DE LITERATURA EM LÍNGUA INGLESA E TRADUÇÕES PARA O PORTUGUÊS BRASILEIRO							
AUTOR	TÍTULO ORIGINAL	NÚMERO DE CARACTERES	FONTE	TÍTULO TRADUÇÃO	NÚMERO DE CARACTERES	TRADUTOR	FONTE
James Joyce	Araby	12.240	http://fiction.eserver.org/short/araby.html	Arábia	12.182	Roberto Schmitt-Prym	http://www.bestiario.com.br/18_arquivos/arabia.html
Virginia Woolf	Monday or Tuesday	1.884	http://www.bartleby.com/85/3.html	Segunda ou terça-feira	1.998		http://www.bestiario.com.br/1_arquivos/woolf.html
O. Henry	The ransom of the Red Chief	21.778 (12.376 / 9.402)	http://fiction.eserver.org/short/ransom_of_red_chief.html	O resgate do Chefe Vermelho	22.960 (13.133 / 9.827)	Bianca Pasqualini	Arte e Letra, Estórias K, 2010
Nathaniel Hawthorne	Wakefield	19.737 (10.544 / 9.193)	http://classiclit.about.com	Wakefield	19.693 (10.556 / 9.146)	Zaida Maldonado	http://www.bestiario.com.br/5_arquivos/Wakefield.html

Tabela 3. Bloco 1.2 – Contos de literatura em língua inglesa e traduções para o português brasileiro.

BLOCO 2.1 – MACHADO DE ASSIS E TRADUÇÕES PARA O INGLÊS						
TÍTULO ORIGINAL	NÚMERO DE CARACTERES	FONTE	TÍTULO TRADUÇÃO	NÚMERO DE CARACTERES	TRADUTOR	FONTE
A cartomante	18.203 (10.362 / 7.617)	http://machado.mec.gov.br	The fortune-teller	22.170 (12.675 / 9.338)	Isaac Goldberg	Projeto Gutenberg
Viver!	14.209 (8.204 / 5.788)		Life	15.450 (8.939 / 6.384)		
Cantiga de esponsais	8.066		Wedding song	8.844	Gregory Rabassa	Oxford Anthology of the Brazilian Short Story
O enfermeiro	17.911 (8.797 / 9.022)	http://www.brazilianstories.com	Looking after	19.914 (9.627 / 10.157)	Francis Johnson	http://www.brazilianstories.com
Marcha fúnebre	14.431 (7.552 / 6.811)		With muffled drum	15.415 (8.116 / 7.202)		
A vida eterna	27.737 (14.796 / 12.881)		Life eternal	28.622 (14.780 / 13.721)		

Tabela 4. Bloco 2.1 – Machado de Assis e traduções para o inglês.

BLOCO 2.2 – CONTOS BRASILEIROS E TRADUÇÕES PARA O INGLÊS							
AUTOR	**TÍTULO ORIGINAL**	**NÚMERO DE CARACTERES**	**FONTE**	**TÍTULO TRADUÇÃO**	**NÚMERO DE CARACTERES**	**TRADUTOR**	**FONTE**
Lima Barreto	O único assassinato de Cazuza	9.800	http://pt.wikisource.org/wiki/O_%C3%BAnico_assassinato_de_Cazuza	Killer	10.300	Francis Johnson	http://www.brazilianstories.com
Humberto de Campos	A promessa	14.930 6.587 / 8.343	http://pt.wikisource.org/wiki/A_Promessa	Light of my life	15.200 6.500 / 8.500		
Coelho Netto	Firmo, o vaqueiro	8.478	http://peregrinacultural.wordpress.com	Christmas corral	8.971		
Lima Barreto	O homem que sabia javanês	18.251 9.505 / 8.640	Projeto Gutenberg	The man who spoke javanese	18.387 9.401 / 8.892	Gregory Rabassa	Oxford Anthology of the Brazilian Short Story
Humberto de Campos	O Diálogo das Caveiras	7.311	http://www.brazilianstories.com	Fish and Filossofy	7.393	Francis Johnson	http://www.brazilianstories.com
Coelho Netto	O duplo	7.774		Me Too	8.268		
Lima Barreto	O número da sepultura	18.067 10.320 / 7.747		Late bet	18.684 10.603 / 8.081		
Humberto de Campos	Vingança	12.556		In the forests of the night	12.525		

Tabela 5. Bloco 2.2 – Contos brasileiros e traduções para o inglês.

5. PROCEDIMENTOS

Diferentemente do procedimento comum a trabalhos cuja metodologia segue os preceitos da Linguística de Corpus, os textos analisados para este estudo foram tratados individualmente em blocos, conforme foi exposto nas seções e capítulos anteriores. Uma particularidade da preparação dos textos para uso nas ferramentas foi a necessidade de corrigir eventuais marcas de parágrafo, letras maiúsculas e pontuação, uma vez que interferem diretamente no processamento textual do Coh-Metrix e do Coh-Metrix-Port e, consequentemente, nos resultados. Desse modo, os textos foram salvos em arquivos individuais com extensão DOC, com um cabeçalho contendo informações tais como, por exemplo, título, autor, gênero, fonte e número de caracteres. Após a divisão em dois blocos, um de contos em inglês e respectivas traduções para o português brasileiro e outro de contos em português brasileiro traduzido para o inglês, cada um deles foi processado individualmente nas ferramentas Coh-Metrix (textos em inglês) e Coh-Metrix-Port (textos em português).

Das 48 métricas adaptadas do Coh-Metrix para o Coh-Metrix-Port, apenas 31 são comparáveis, pois há métricas próprias de cada recurso e métricas incompatíveis devido aos recursos utilizados.[24] Num primeiro momento, selecionamos as métricas a serem analisadas, englobando todas as categorias de análise (lexicais, sintáticas e semânticas, tendo em vista que a categoria de medidas do tipo referencial ainda está em construção). Então, ao compararmos a descrição de medidas do Coh-Metrix com as do Coh-Metrix-Port, verificamos que nem todas têm grandezas equiparáveis, como mostra a Tabela 6:

Métricas	**COH-METRIX-PORT**	**EQUIVALÊNCIA DOS ÍNDICES**	**COH-METRIX-ING**
Métricas Lexicais	Índice Flesch	**EQUIVALENTE**	Flesch Reading Ease
	Número de palavras	**EQUIVALENTE**	Number of words
	Incidência de palavras de conteúdo	**NÃO EQUIVALENTE**	Concreteness content words
	Frequências	**NÃO EQUIVALENTE**	Raw frequency
	Mínimo de frequências	**NÃO EQUIVALENTE**	Min. Raw frequency
Métricas	Operadores lógicos	**EQUIVALENTE**	Logic Operators

[24] Por exemplo, a *Wordnet*, que é uma base de dados lexicais exclusiva do inglês.Ver G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. 1990. WordNet: An online lexical database. *Int. J. Lexicograph*. 3, 4, pp. 235–244.

Sintáticas	Palavras antes de verbos principais	**EQUIVALENTE**	Words before main verb
	Types/*token*	**EQUIVALENTE**	Types/token
	Incidência de sintagmas nominais	**EQUIVALENTE**	NP incidence
	Incidência de conectivos	**EQUIVALENTE**	All conectives
Métricas Semânticas	Anáforas	**EQUIVALENTE**	Anaphor reference
	Sobreposição de argumentos adjacentes	**EQUIVALENTE**	Adjacent argument overlap
	Sobreposição de argumentos	**EQUIVALENTE**	Argument overlap

Tabela 6. Comparação da equivalência de medidas entre as ferramentas Coh-Metrix e Coh-Metrix-Port.

No caso da "incidência de palavras de conteúdo", a métrica é, para fins contrastivos, não comparável à métrica "concreteness content words", pois esta usa um banco de dados de concretude de palavras[25] de conteúdo indisponível em português. As métricas de frequências também não são equiparáveis, tendo em vista que o Coh-Metrix usa o banco de dados CELEX e o Coh-Metrix-Port usa o Banco de Português.

MÉTRICAS LEXICAIS	MÉTRICAS SINTÁTICAS	MÉTRICAS SEMÂNTICAS
Número palavras Número sentenças Número parágrafos Palavras por sentenças Sentenças por parágrafos Sílabas por palavras Índice Flesch	Conectivos (todas as métricas) Negações Operadores lógicos Sintagmas nominais Modificadores por sintagma Pronomes por sintagma Pronomes pessoais Tipo/ *token* Palavras antes de verbos	Referência anafórica Referência anafórica (adjacente) Sobreposição de palavras de conteúdo (adjacente) Sobreposição de argumentos Sobreposição de argumentos (adjacente) Sobreposição de radical de palavras Sobreposição de radical de palavras (adjacente)

Tabela 7. Métricas contrastáveis entre as ferramentas Coh-Metrix e Coh-Metrix-Port.

O primeiro passo foi calcular a média dos resultados das métricas para os textos com mais de 15 mil caracteres, que foram processados em duas etapas cada um. O segundo passo foi realizar o teste estatístico *t-Student*[26], para cada métrica e entre os blocos de textos, para avaliar se a diferença entre as médias dos resultados obtidos eram significativamente diferentes com confiança de 95% (*p-value* < 0,05). Assim, das 31

[25] A base de dados é a MCR Psycholinguistic Database; ver http://www.psych.rl.ac.uk/MRC_Psych_Db.html.

[26] O teste *t-Student* consiste em usar os dados de uma amostra para calcular a estatística t e contrastá-la com a distribuição *t-Student* a fim de determinar a probabilidade de se ter obtido o resultado observado, caso a hipótese nula seja verdadeira. Uma hipótese nula geralmente afirma que não existe relação entre dois fenômenos medidos. Ver FRIES, Stephan. Useful statistics for corpus linguistics. In SÁNCHEZ, Aquilino; ALMELA, Moisés (eds.). *A mosaic of corpus linguistics: selected approaches*. pp. 269-291. Frankfurt: Peter Lang, 2010.

métricas comparáveis, 18 apresentam resultados com diferenças estatisticamente significativas, como mostra a Tabela 8, a seguir.

		Português - Inglês				Inglês - Português			
		PORTUGUÊS TEXTOS-FONTE		INGLÊS TRADUÇÃO		PORTUGUÊS TRADUÇÃO		INGLÊS TEXTOS-FONTE	
		Média	Dsvp	Média	Dsvp	Média	Dsvp	Média	Dsvp
Métricas lexicais	**1) Sílabas por palavras**	2,64	0,11	1,40	0,08	2,85	0,14	1,44	0,09
	2) Índice Flesch	62,67	5,48	74,6	7,19	48,2	10,29	64,37	11,61
Métricas sintáticas	**3) Conectivos temporais positivos**	14,88	2,35	11,03	2,65	14,06	3,86	10,29	2,57
	4) Conectivos causais positivos	34,85	5,85	21,10	5,04	38,42	5,05	22,27	5,86
	5) Conectivos lógicos positivos	28,39	5,22	18,63	4,39	31,54	4,44	20,93	6,92
	6) Conectivos lógicos negativos	4,63	1,41	11,06	3,16	4,24	2,07	14,97	3,53
	7) Negações	5,29	3,55	7,78	2,78	3,65	1,57	11,34	3,94
	8) Sintagmas nominais	248,88	14,39	298,74	12,44	234,35	23,83	278,53	19,36
	9) Modificadores por sintagmas	0,53	0,05	0,74	0,16	0,59	0,06	0,81	0,11
	10) Pronomes por sintagmas	0,23	0,07	0,37	0,08	0,21	0,05	0,35	0,11
	11) Pronomes pessoais	14,75	7,36	111,74	26,74	13,23	8,59	97,99	33,73
Métricas semânticas	**12) Referência anafórica**	0,44	0,22	0,24	0,08	0,49	0,22	0,27	0,11
	13) Referência anafórica adjacente	0,31	0,16	0,40	0,12	0,34	0,17	0,49	0,18
	14) Sobreposição de argumentos	0,15	0,05	0,29	0,07	0,19	0,11	0,37	0,12
	15) Sobreposição de argumentos (adjacentes)	0,20	0,09	0,36	0,10	0,25	0,13	0,44	0,14
	16) Sobreposição de palavras de conteúdo (adjacentes)	0,18	0,07	0,08	0,02	0,21	0,08	0,08	0,04
	17) Sobreposição de radical de palavras	0,26	0,08	0,10	0,03	0,30	0,14	0,12	0,05
	18) Sobreposição de radical de palavras (adjacentes)	0,33	0,11	0,13	0,05	0,40	0,13	0,14	0,08

Tabela 8. As métricas das ferramentas Coh-Metrix e Coh-Metrix-Port com diferenças estatisticamente significativas.

De posse desses resultados, a ferramenta Weka foi usada para apurar as relações discriminativas entre as métricas, tendo como referência quatro classes: textos-fonte em inglês, traduções para o português, textos-fonte em português e traduções para o inglês.

6. RESULTADOS E DESCRIÇÃO DOS DADOS OBTIDOS

6.1 RESULTADOS DAS FERRAMENTAS COH-METRIX E COH-METRIX-PORT

Nesta parte, descrevem-se e comentam-se os resultados obtidos após o processamento dos textos nas ferramentas Coh-Metrix e Coh-Metrix-Port. Para facilitar a leitura do capítulo, as descrições foram subdivididas em três seções: a primeira apresenta os resultados das métricas lexicais (sílabas por palavras e índice Flesch); a segunda, os resultados das métricas sintáticas (conectivos, negações, incidência de pronomes pessoais, pronomes por sintagmas, sintagmas nominais e modificadores por sintagma); e a terceira, das métricas semânticas (referências anafóricas, referências anafóricas adjacentes, sobreposição de argumentos, sobreposição de argumentos adjacentes, sobreposição de palavras de conteúdo adjacentes, sobreposição de radical de palavras e sobreposição de radical de palavras adjacentes). Cada um dos itens é ilustrado por um gráfico com os resultados de cada conjunto de textos: textos-fonte em português, traduções para o inglês, textos-fonte em inglês e traduções para o português.

Em seguida, na parte final deste capítulo, são apresentados os resultados obtidos a partir da classificação das métricas por AM com a ferramenta Weka. Após uma breve introdução, em que alguns aspectos essenciais ao entendimento da descrição dos resultados são expostos, as análises são relatadas, acompanhando a seguinte ordem: métricas dos textos em português e métricas dos textos em inglês (Análise 1); métricas dos textos originais e métricas dos textos traduzidos (Análise 2); métricas dos textos originais em português e métricas dos textos traduzidos para o português (Análise 3); e métricas dos textos originais em inglês e métricas dos textos traduzidos para o inglês (Análise 4). Para cada uma das análises, há uma figura ilustrando a árvore de decisão obtida com o Weka, com tabelas listando informações sobre a precisão do processamento, obtidas, também, com a ferramenta mencionada. Passemos, então, à descrição dos resultados.

6.1.1 MÉTRICAS LEXICAIS

6.1.1.1 Sílabas por palavras e índice Flesch

O índice Flesch engloba a métrica "sílabas por palavras", a qual, portanto, não foi analisada. Houve diminuição considerável no índice Flesch dos textos-fonte em português para as traduções em inglês, como mostra o Gráfico 1, na página a seguir: de 62,7 (primeira coluna do Gráfico 1), valor que indica textos fáceis, de acordo com a escala de dificuldade do índice Flesch, ilustrada no Quadro 1, para 74,6 (segunda coluna), valor que indica textos razoavelmente fáceis. Nos textos-fonte em inglês e nas traduções para o português, ocorre o contrário: parte-se, em inglês, de 64,37 (quarta coluna), que indica textos dentro da faixa padrão, e a média dos índices Flesch das traduções para o português é 48,2 (terceira coluna), que indica textos difíceis. Nesse quesito, quando traduzidos para o português, temos textos que exigem mais esforço de compreensão e que são indicados para alunos do Ensino Médio ou superior, ou seja, leitores com letramento em nível pleno.

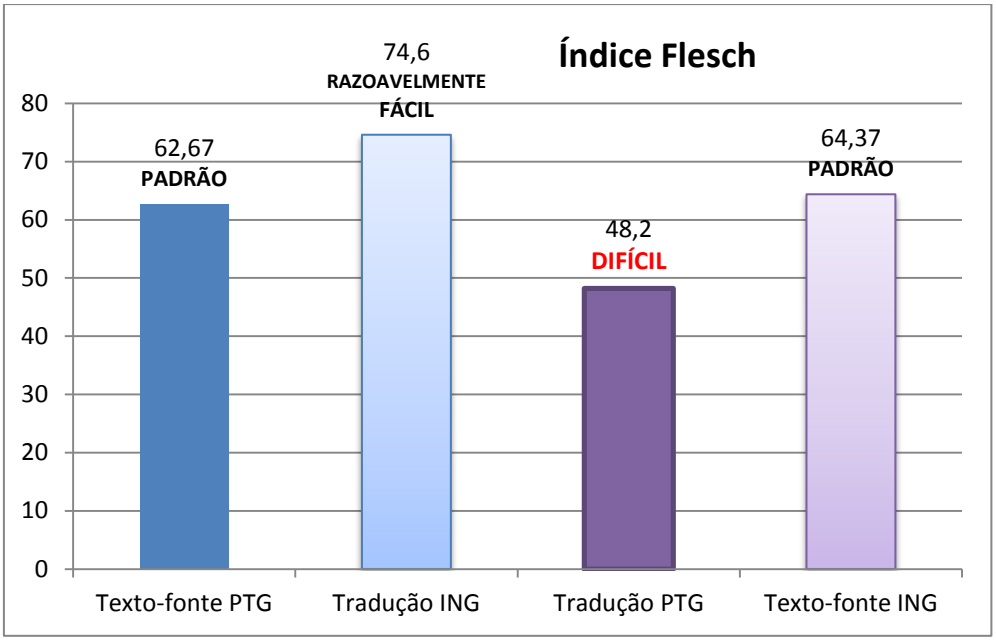

Gráfico 1. Índice Flesch.

Escala expandida de dificuldade textual de acordo com o índice Flesch[27]:
• **Muito fáceis**: índice entre 90 a 100, textos adequados para leitores com nível de escolaridade até a 4ª. série do Ensino Fundamental. • **Fáceis**: índice entre 80 a 89, textos adequados a alunos com escolaridade até a 8ª. série do Ensino Fundamental. • **Razoavelmente fáceis**: índice entre 70 a 79, textos adequados a alunos com escolaridade até a 8ª. série do Ensino Eundamental. • **Padrão**: índice entre 60 e 69, textos adequados a alunos com escolaridade até a 8ª. série do Ensino Eundamental. • **Razoavelmente difíceis**: índice entre 50 a 59, textos adequados para alunos cursando o Ensino Médio ou universitário. • **Difíceis**: índice entre 30 a 49, textos adequados para leitores com Ensino Médio ou universitário. • **Muitos difíceis**: índice entre 0 a 29, textos adequados apenas para áreas acadêmicas específicas.

Quadro 1. Escala expandida de dificuldade textual de acordo com o índice Flesch.

6.1.2 MÉTRICAS SINTÁTICAS

De acordo com Crossley e McNamara (2007), define-se complexidade sintática como interposição de orações, estrutura frasal densa, ambiguidade sintática ou agramaticalidade. A partir dessa definição é que o Coh-Metrix mede a complexidade textual de um texto. Sentenças complexas têm maior proporção de constituintes por palavra e por sintagma do que sentenças simples.

6.1.2.1 Conectivos (temporais positivos; causais positivos; lógicos positivos; lógicos negativos)

Conectivos compõem a tessitura do texto, sendo, em tese, facilitadores da leitura. Assim, os resultados apontam para um maior índice de coesão em português e, por conseguinte, maior legibilidade. No entanto, é preciso fazer algumas considerações – bastante breves – sobre essas métricas.

Em primeiro lugar, os conectivos foram traduzidos a partir da lista de conectivos em inglês (ver Anexo A para a listagem completa dos conectivos em inglês e em português). Assim, entram em jogo questões tradutórias, como, por exemplo, o fato de

[27] Conforme documentação dos desenvolvedores da ferramenta Coh-Metrix-Port, em http://caravelas.icmc.usp.br/wiki/images/9/91/Coh_Metrix_2.0.pdf.

que um conectivo em inglês, ao ser traduzido para o português, não necessariamente cumpre a mesma função coesiva. Um exemplo disso seria o conectivo lógico "enable", traduzido como "habilita" para o português, item que não parece encaixar-se no perfil de conectivos da língua portuguesa. Há também casos como o de "even though", traduzido como "mesmo embora", quando uma tradução mais apropriada seria "ainda que". Além disso, há várias possibilidades e opções de tradução para os diferentes conectivos, as quais não estão listadas na lista de conectivos em português, como, por exemplo, "ademais", "no mínimo" e "além do mais" (conectivos aditivos), "apenas se", "contanto que" e "com a finalidade de" (conectivos causais), "nesse meio-tempo", "ínterim" (conectivos temporais), só para citar alguns.

Em segundo lugar, alguns conectivos poderiam estar categorizados em categorias diferentes, como o conectivo "com", listado entre os conectivos causais, e não entre os aditivos. Por tratar-se de um item lexical de alta frequência, sem dúvida os resultados foram fortemente influenciados por essa categorização.

Sugere-se que as métricas para o cálculo da incidência de conectivos sejam revisadas por um linguista e que os conectivos sejam reanalisados e retraduzidos. Levando todas essas questões em consideração, as métricas de incidência de conectivos não serão incluídas entre as métricas selecionadas para a análise por aprendizagem de máquina.

6.1.2.2 Negações

Esta métrica apresentou diferenças bastante expressivas entre os textos, principalmente entre os textos-fonte em inglês (11,34) e as traduções para o português (3,65) (ver Gráfico 3). Assim, julgamos necessário conferir a incidência de negações diretamente nos textos para averiguar a discrepância entre os resultados. Como o número de textos é elevado, contar manualmente as negações em cada um deles não seria viável. A estratégia, então, foi selecionar um texto curto, o conto "The oval portrait", e sua tradução para o português, "O retrato oval", e marcar as negações em ambos. Em inglês, consideram-se negações: *no, not, never, none/neither, nothing;* em português, consideram-se negações: não, nem, nenhum, nenhuma, nada, nunca e jamais.

De acordo com o manual do Coh-Metrix-Port[28], a incidência de negações é calculada da seguinte maneira: número de negações/(número de palavras/1000). O número de palavras do texto em inglês é 1.292; o número de palavras da tradução é 1.138. Foram encontradas 14 negações no texto em inglês e 13 no texto em português. Fazendo-se o cálculo para o texto em inglês: 14/(1.292/1000), temos o resultado 10,8, que corresponde ao resultado do Coh-Metrix. No texto em português, foram encontradas 13 negações: 13/(1.138/1000), e o resultado é 12,3, o que não corresponde ao resultado do Coh-Metrix-Port, que foi de 1,75.

Para confirmar a suspeita de mau funcionamento da métrica, um segundo texto foi testado: "Monday Tuesday", e sua tradução para o português, "Segunda ou terça-feira". O número de palavras do primeiro é 306; do segundo, 305. A incidência de negações, segundo o Coh-Metrix, é 6,5; e a incidência de negações, segundo o Coh-Metrix-Port, para o texto em português, é zero. Ao fazer a leitura do texto em inglês, encontramos 2 negações, o que é condizente com o resultado obtido (6,5); no texto em português, encontramos 1 negação. Portanto, a incidência de negações no texto em português é 3,3, o que aponta para um erro no processamento do Coh-Metrix-Port. O erro nos resultados do Coh-Metrix-Port para esta métrica foi comunicado aos desenvolvedores da ferramenta, que estão averiguando a causa do problema. Por conseguinte, a incidência de negações não poderá ser usada nesta análise.

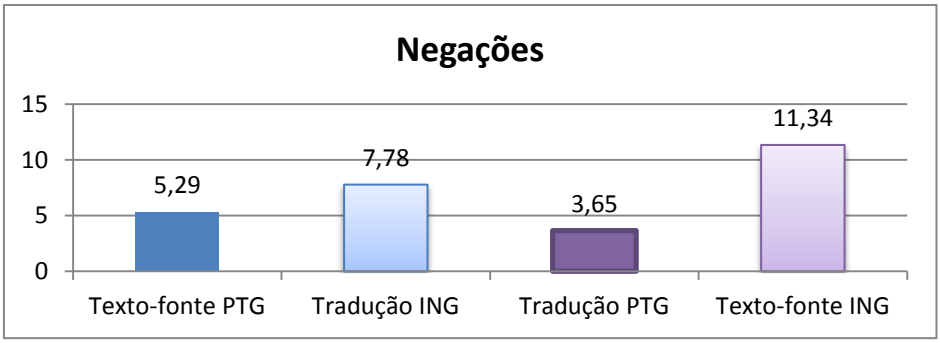

Gráfico 2. Incidência de negações.

[28] Disponível no *site* da ferramenta: http://caravelas.icmc.usp.br:3000/index/info#sub.

6.1.2.3 Incidência de pronomes pessoais, pronomes por sintagmas, sintagmas nominais e modificadores por sintagma

Em português, em função da desinência verbal, não é necessário usar pronomes pessoais com tanta frequência, o que explica a baixa incidência encontrada nos textos em português: 14,75 nos textos-fonte e 13,23 nas traduções (primeira e terceira colunas do Gráfico 4). Por outro lado, não é surpresa o número elevado de pronomes pessoais em inglês, visto que precisam ser expressos em função da gramática dessa língua: 97,99 nos textos-fonte (quarta coluna do Gráfico 4) e 111,74 nas traduções (segunda coluna do Gráfico 4). Essa relação também fica aparente no índice maior de pronomes por sintagmas em inglês, em torno de 0,35 nos textos-fonte e nas traduções para essa língua (segunda e quarta coluna do Gráfico 5), pois essa métrica contempla também os pronomes pessoais. A Tabela 9 ilustra essas diferenças, com exemplos retirados do *corpus*. As palavras em negrito indicam os pronomes pessoais, e os rasurados os pronomes elipsados.

Português	Inglês
Aos cinquenta e três anos, ~~**(ele)**~~ não tinha mais um parente próximo junto de si.	At fifty-three years of age, **he** was alone, with no close relatives.
~~**(ele)**~~ Andava pelos oitenta anos, mas quem **o** visse a cavalo, no campo, não lhe daria tanta idade.	**He** was about eighty years old, but if **you** saw **him** on a horse out in the country **you** wouldn't have thought so.

Tabela 9. Exemplos de ocorrências de pronomes pessoais e pronomes por sintagmas.

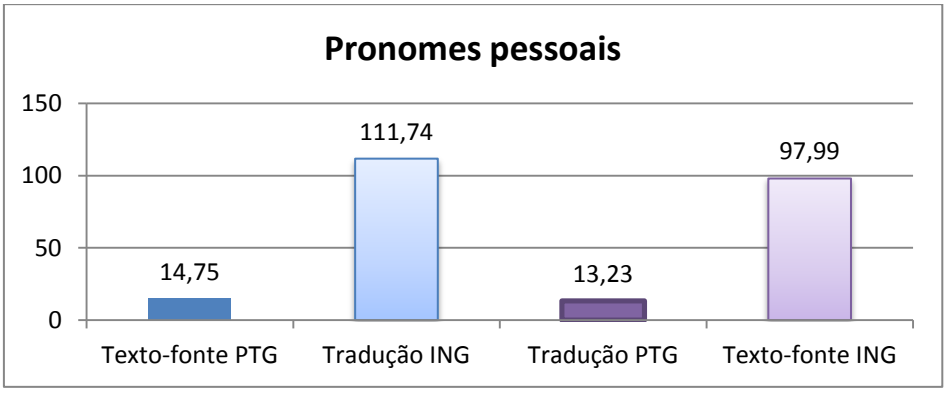

Gráfico 3. Incidência de pronomes pessoais.

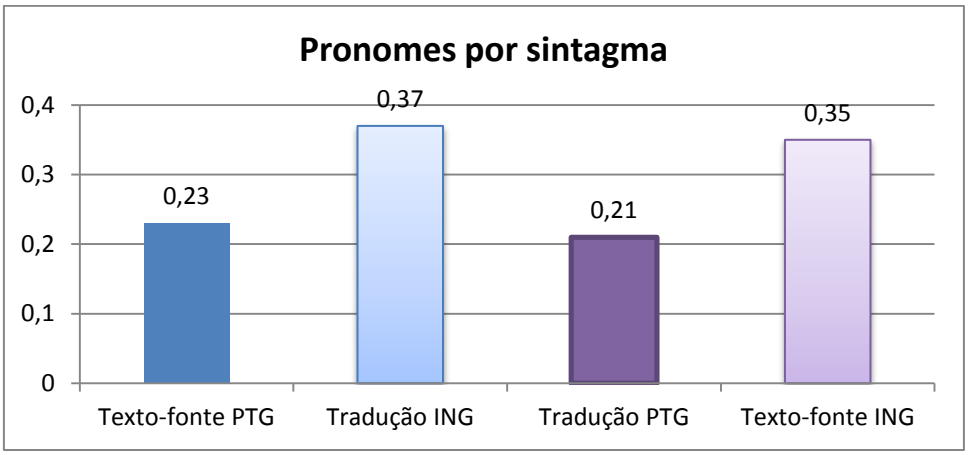

Gráfico 4. Média de pronomes por sintagma.

Os resultados apontam também que o número maior de sintagmas nominais encontrados nos textos em inglês, tanto de partida quanto de chegada (primeira e terceira colunas do Gráfico 6), também sofre a influência do maior uso de pronomes pessoais em inglês (ver Gráfico 4, acima). No entanto, o número de modificadores por sintagma – métrica que não envolve pronomes, mas adjetivos, advérbios e artigos – é inferior em português, em comparação ao inglês: os textos-fonte em português têm 0,53 modificadores por sintagma (primeira coluna do Gráfico 7), e as traduções para o inglês 0,74 (segunda coluna do Gráfico 7); as traduções para o português têm 0,59 modificadores por sintagma (terceira coluna do Gráfico 7), ao passo que os textos-fonte em inglês têm 0,81 (quarta coluna do Gráfico 7). Esses resultados indicam legibilidade maior neste quesito em português.

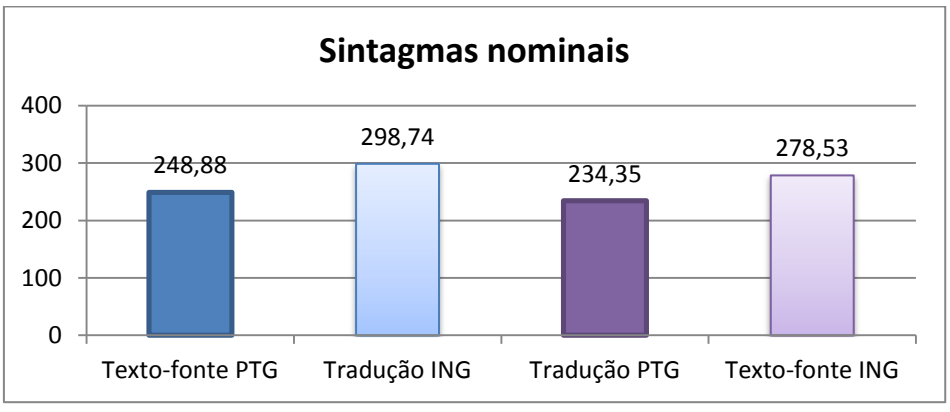

Gráfico 5. Incidência de sintagmas nominais.

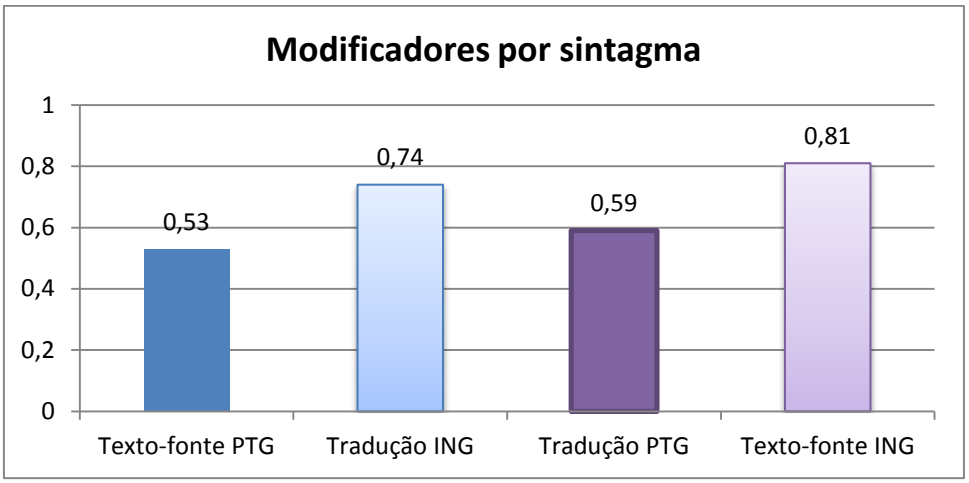

Gráfico 6. Média do número de modificadores por sintagma.

6.1.3 MÉTRICAS SEMÂNTICAS

6.1.3.1 Referência anafórica e referência anafórica adjacente

A referência anafórica é estabelecida por meio do uso de anáforas, repetição de sentenças, artigos definidos, pronomes demonstrativos, etc., e ocorre quando um substantivo, pronome ou sintagma nominal se refere a outro constituinte no texto (GRAESSER et al., 2001). A métrica referência anafórica calcula a proporção de referências anafóricas que se referem a um constituinte presente em até cinco sentenças anteriores; e a métrica referência anafórica adjacente calcula a proporção de referências anafóricas entre sentenças adjacentes.

Exemplo de referência anafórica adjacente (SCARTON, 2009):

"Dentro do lago, existem peixes, como a traíra e o dourado, além da palometa, um tipo de piranha. Ela é uma espécie carnívora que se alimenta de peixes." Nesse exemplo, os “candidatos” a resolver a anáfora pronominal são traíra, palometa e piranha. Como há três "candidatos" e uma sentença adjacente, o valor final da métrica é 3/1 = 3.

Nos textos traduzidos para o inglês, a incidência de anáforas diminui, o que sugere menos ambiguidade e menor exigência de processamento de leitura: a incidência de anáforas nos textos-fonte em português é de 0,44 (primeira coluna do Gráfico 8), ao passo que, nos textos traduzidos para o inglês, é de 0,24 (segunda coluna do Gráfico 8). Nos textos traduzidos para o português, o inverso ocorre, e a incidência de anáforas aumenta, aumentando também a potencial complexidade dos textos: a incidência de anáforas nos textos-fonte em inglês é de 0,27 (quarta coluna do Gráfico 8), enquanto nas

traduções para o português é de 0,49 (terceira coluna do Gráfico 8). Contudo, a incidência de referências anafóricas em sentenças adjacentes segue o caminho inverso: aumenta nas traduções para o inglês, com incidência de 0,4 (segunda coluna do Gráfico 8), enquanto nos textos-fonte em português é de 0,31 (primeira coluna do Gráfico 8). Já nos textos traduzidos para o português, a incidência de referências anafóricas adjacentes diminui: a incidência é de 0,34 (terceira coluna do Gráfico 8) nas traduções para o português, e de 0,49 nos textos-fonte em inglês (quarta coluna do Gráfico 8). Isso indica, consequentemente, que o uso de anáforas, nos textos em inglês, é mais restrito às sentenças adjacentes, o que possivelmente exige menos memória do leitor.

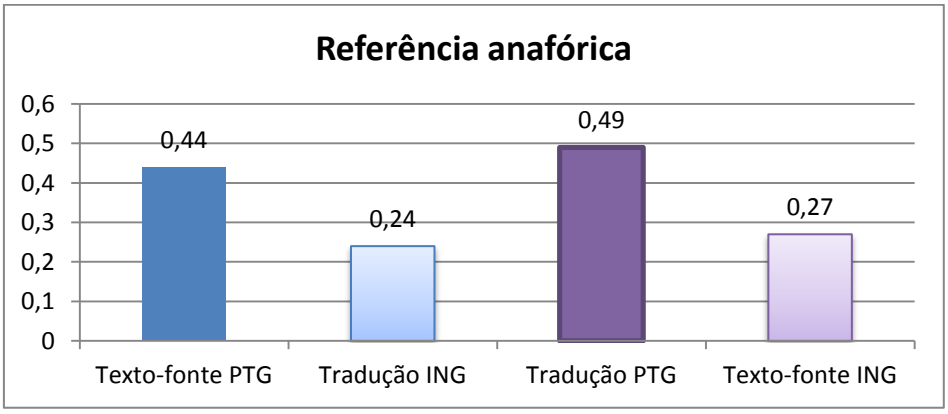

Gráfico 7. Referências anafóricas a constituintes até cinco sentenças anteriores.

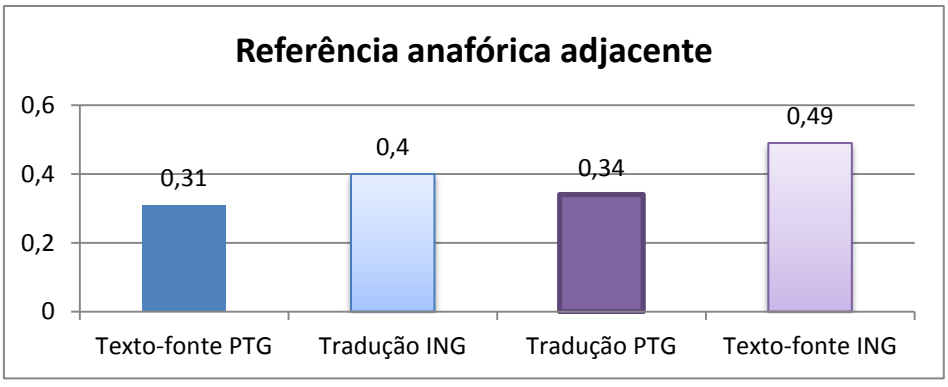

Gráfico 8. Proporção de referências anafóricas em sentenças adjacentes.

6.1.3.2 Correferência

6.1.3.2.1 Sobreposição de argumentos e sobreposição de argumentos (adjacente)

A métrica sobreposição de argumentos é o cálculo da proporção de todos os pares de sentenças que compartilham um ou mais argumentos. A métrica sobreposição de

argumentos adjacentes é o cálculo da proporção de sentenças adjacentes que compartilham um ou mais argumentos (substantivos, pronomes ou sintagmas nominais).

Exemplo de sobreposição de argumentos (SCARTON, 2009):

"(1) Dentro do lago, existem **peixes**, como a traíra e o dourado, além da palometa, um tipo de piranha. (2) Ela é uma espécie carnívora que se alimenta de **peixes**. (3) No verão, elas ficam mais próximas das margens da barragem, atraídas pela movimentação das pessoas e por restos de comida que alguns turistas deixam na água quando lavam os pratos."

As duas métricas aumentam nas traduções do português para o inglês e diminuem nas traduções do inglês para o português, como mostram os gráficos 10 e 11. A sobreposição de argumentos pode gerar dificuldade de leitura, pois pode induzir ao erro na interpretação de ambiguidades, sobretudo no processamento de leitura de pronomes e sintagmas nominais. A incidência de pronomes por sintagma e modificadores por sintagma também é maior nos textos em inglês, como mostrado na seção 6.1.2.3. Logo, os textos em inglês, tanto os textos-fonte quanto os traduzidos, podem ser considerados mais complexos.

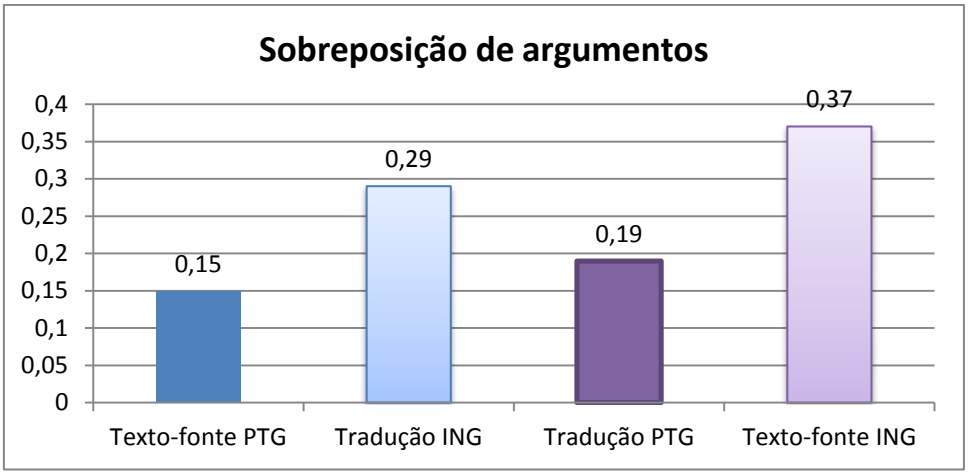

Gráfico 9. Proporção de todos os pares de sentenças que compartilham um ou mais argumentos.

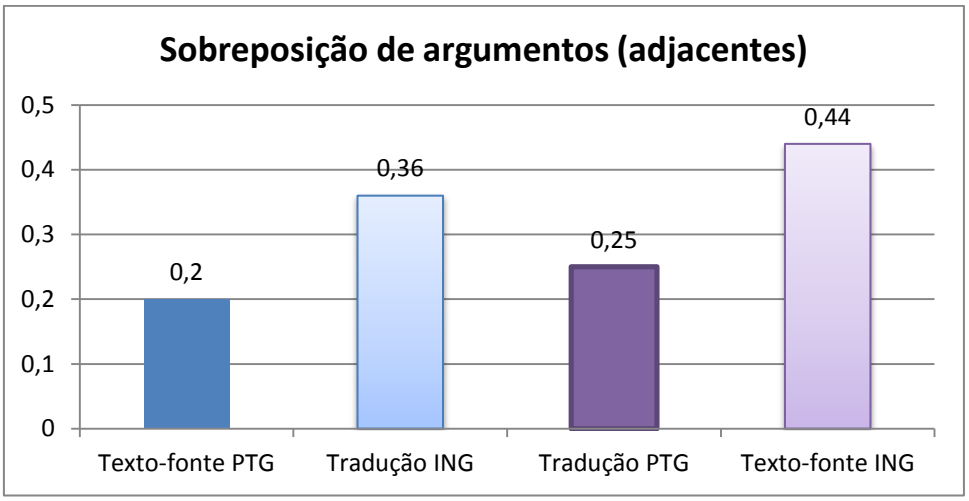

Gráfico 10. Proporção de sentenças adjacentes que compartilham um ou mais argumentos.

6.1.3.2.2 Sobreposição de palavras de conteúdo (adjacente), sobreposição de radical de palavras e sobreposição de radical de palavras (adjacente)

A métrica sobreposição de palavras de conteúdo (adjacente) calcula a proporção de sentenças adjacentes que compartilham palavras de conteúdo. A métrica sobreposição de radical de palavras calcula a proporção de todos os pares de sentenças que compartilham radicais. A métrica sobreposição de radical de palavras (adjacente) calcula a proporção de sentenças adjacentes que compartilham radicais.

Exemplo de sobreposição de palavras de conteúdo:

"As tapeçarias caíam em pesadas dobras de **tapetes** do mesmo material e da mesma **cor**. Mas somente nesta sala a cor das janelas não correspondia à das decorações. As vidraças e **tapetes** menores ali eram escarlates, da **cor** de sangue vivo."

Exemplo de sobreposição de radical de palavras:

"As **tape**çarias caíam em pesadas dobras **tape**tes do mesmo material e da mesma cor. Mas somente nesta sala a cor das janelas não correspondia à das decorações. As vidraças e **tape**tes menores ali eram escarlates, da cor de sangue vivo."

Essas métricas dizem respeito ao fluxo do texto e à manutenção de tópicos. Todos os textos em inglês, tanto traduzidos como originais, têm índices menores nessas três métricas (ver os gráficos 12, 13 e 14 adiante), o que indica menor repetição de palavras e de radicais e uma possível maior dificuldade de leitura, pois as palavras de conteúdo, em vez de serem repetidas, provavelmente estão sendo referenciadas por pronomes e expressões anafóricas. Esse resultado, que revela uma variação lexical maior nos textos em inglês, contraria a expectativa de que textos nessa língua são mais repetitivos.

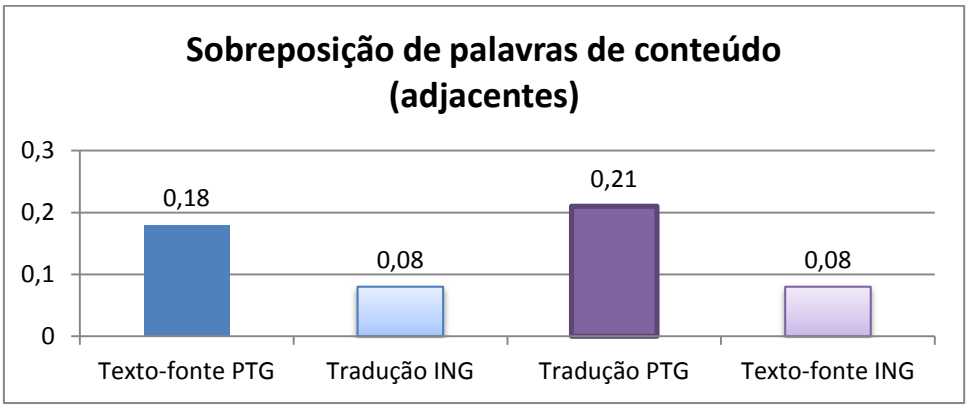

Gráfico 11. Proporção de sentenças adjacentes que compartilham palavras de conteúdo.

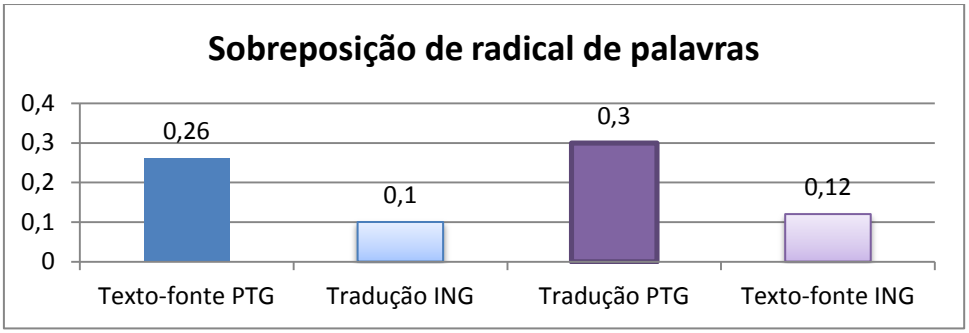

Gráfico 12. Proporção de todos os pares de sentenças que compartilham radicais.

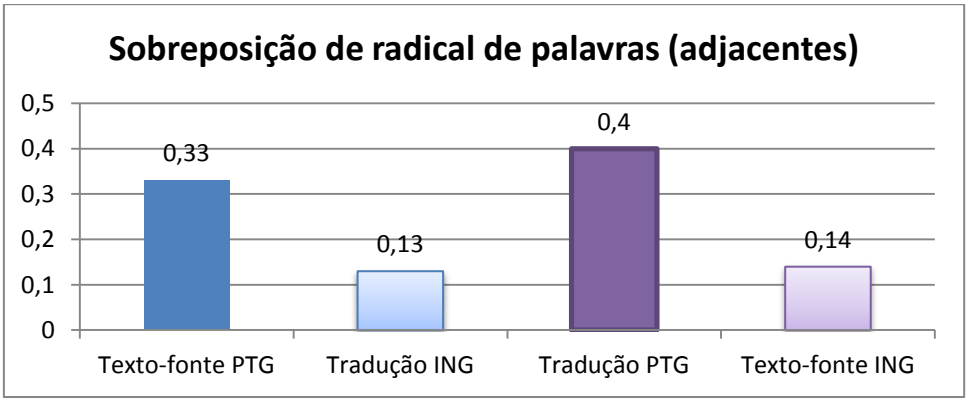

Gráfico 13. Proporção de sentenças adjacentes que compartilham radicais.

6.2 RESULTADOS DA CLASSIFICAÇÃO POR APRENDIZAGEM DE MÁQUINA

Como mencionado anteriormente, o algoritmo escolhido para classificação das métricas por AM foi a implementação J48 do algoritmo classificador C4.5 para a construção de árvores de decisão.

Uma árvore de decisão mostra quais são as relações discriminativas entre os atributos (no caso, as métricas do Coh-Metrix e do Coh-Metrix-Port) para cada classe de textos, ou seja, uma árvore de decisão é formada por um conjunto de nós de decisão, os quais permitem a classificação de cada caso. Em outras palavras, para fins desta pesquisa, a estrutura em árvore de decisão mostra visualmente quais são as métricas (os atributos) estatisticamente mais características e distintivas em cada classe, definida de acordo com a natureza do texto: textos em português (todos), textos em inglês (todos), textos originais (todos), textos traduzidos (todos), textos originais em inglês, textos traduzidos para o português, textos originais em português e textos traduzidos para o inglês (ver Figura 7). No Anexo C estão apresentados os arquivos ARFF usados nas análises, caso alguém tenha interesse em testá-los no Weka.

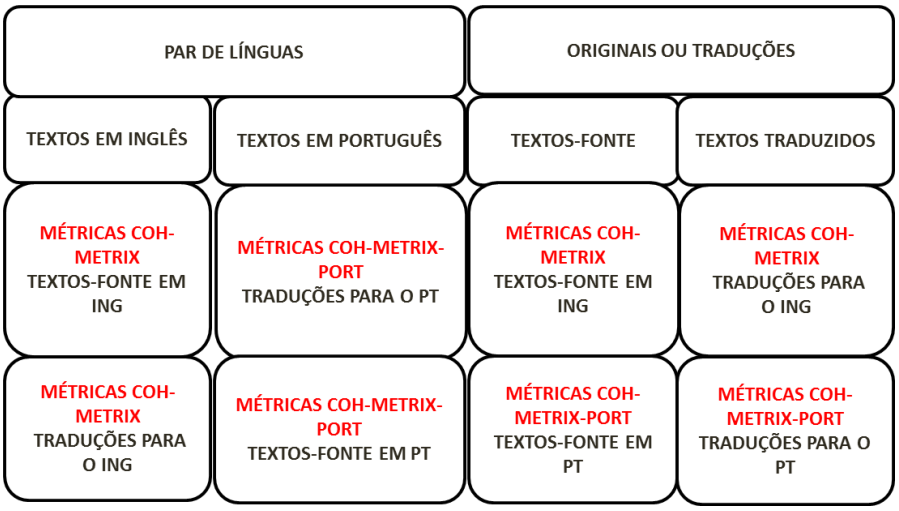

Figura 7. Organização dos textos para classificação das métricas estatisticamente significativas por AM.

As métricas processadas foram aquelas com diferenças estatisticamente significativas, exceto: número de sílabas por palavras (pois está incluída no índice Flesch), incidência de negações (pois essa métrica está produzindo resultados inexatos, como foi mencionado na seção 6.1.2.2) e incidências de conectivos (pois essas métricas precisam de revisão, como foi sugerido na seção 6.1.2.1). O total de métricas processadas é doze (ver Tabela 10).

MÉTRICAS LEXICAIS	MÉTRICAS SINTÁTICAS	MÉTRICAS SEMÂNTICAS
Índice Flesch	Sintagmas nominais Modificadores por sintagma nominal Pronomes por sintagma nominal Incidência de pronomes pessoais	Referência anafórica Referência anafórica (adjacente) Sobreposição de argumentos Sobreposição de argumentos (adjacente) Sobreposição de palavras de conteúdo (adjacente) Sobreposição de radicais Sobreposição de radicais (adjacente)

Tabela 10. Métricas processadas pelo classificador de AM.

As análises foram processadas da seguinte forma:

- **Análise 1**: Métricas dos textos em português x métricas dos textos em inglês; classificação pelo par de línguas (inglês ou português).
- **Análise 2**: Métricas dos textos originais x métricas dos textos traduzidos; classificação pela natureza do texto (original ou traduzido), mesclando as línguas.
- **Análise 3**: Métricas dos textos originais em português x métricas dos textos traduzidos para o português; classificação pela natureza do texto (original ou traduzido), somente em português.
- **Análise 4**: Métricas dos textos originais em inglês x métricas dos textos traduzidos para o inglês; classificação pela natureza do texto (original ou traduzido), somente em inglês.

Antes de passarmos aos resultados, é importante definir algumas medidas de validação dos dados: a **cobertura**, a **precisão** e a **medida-f**. A cobertura indica as classificações corretamente identificadas em relação a tudo que *deveria* ser identificado; a precisão indica as classificações corretas em relação a tudo que *foi* identificado; e a medida-f é a média entre cobertura e precisão. Em outras palavras, a cobertura é a porcentagem de todos os atributos pertencentes à classe em questão que conseguiram ser classificados; a precisão é a porcentagem dos atributos que foram corretamente classificados como pertencentes à classe; e a medida-f é a média dessas duas porcentagens.

As árvores de decisão de cada uma das análises estão ilustradas nas figuras 8 a 11, e a legenda para a leitura dos resultados está na Tabela 11, a seguir.

MÉTRICA EM INGLÊS	MÉTRICA EM PORTUGUÊS
argument_overlap_adjacent	Sobreposição de argumentos em sentenças adjacentes
stem_overlap_adjacent	Sobreposição de radicais de palavras em sentenças adjacentes
anaphor_reference_adjacent	Referências anafóricas em sentenças adjacentes
argument_overlap	Sobreposição de argumentos
stem_overlap	Sobreposição de radicais de palavras
anaphor_reference	Referências anafóricas
noun_phrase_incidence_score	Incidência de sintagmas nominais
ratio_of_pronouns_to_noun_phrases	Proporção de pronomes por sintagmas nominais
personal_pronoun_incidence_score	Incidência de pronomes pessoais
flesch_reading_ease_score	Índice Flesch
mean_number_of_modifiers_per_noun_phrase	Modificadores por sintagma nominal
proportion_of_content_words_that_overlap_between_adjacent_sentences	Proporção de sobreposição de palavras de conteúdo em sentenças adjacentes

Tabela 11. Legenda das figuras 8 a 11.

6.2.1 ANÁLISE 1: MÉTRICAS DOS TEXTOS EM PORTUGUÊS X MÉTRICAS DOS TEXTOS EM INGLÊS

Os resultados dessa análise foram os seguintes:

a. Incidência de pronomes pessoais ≤ 34,08935: classe de textos em português.
b. Incidência de pronomes pessoais > 34,08935: classe de textos em inglês.

Essa análise foi, na verdade, uma comparação entre as línguas inglesa e portuguesa, e mostra que, quando uma incidência de pronomes pessoais menor do que 34 é mais característica de textos em português, enquanto que uma incidência maior do 34 é mais característica de textos em inglês. Não é surpreendente o fato de a incidência de pronomes pessoais ser a raiz da árvore de decisão: o uso frequente de pronomes pessoais é uma propriedade sistêmica intrínseca do inglês, ao passo que, em português, o uso dos pronomes pessoais é integrado à desinência verbal. Os resultados indicam também precisão, cobertura e medida-f excelentes.

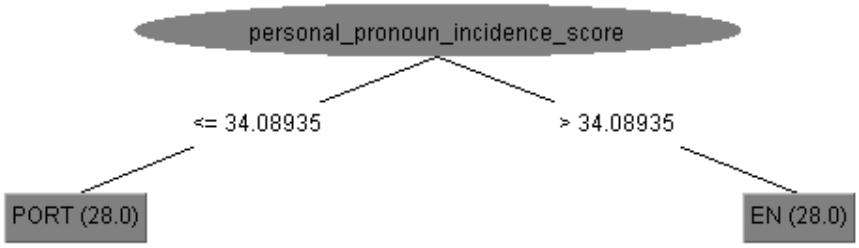

Figura 8. Árvore de decisão da classificação entre textos em português x textos em inglês.

Precisão	Cobertura	Medida-f	Classe
0,966	1	0,982	Inglês
1	0,964	0,982	Português

Tabela 12. Precisão, cobertura e medida-f da classificação entre textos em português x textos em inglês.

6.2.2 ANÁLISE 2: MÉTRICAS DOS TEXTOS ORIGINAIS X MÉTRICAS DOS TEXTOS TRADUZIDOS

Os resultados dessa análise foram os seguintes:

a. Índice Flesch ≤ 46,0269: classe de textos traduzidos
b. Índice Flesch > 46,0269
    a.1. Incidência de sintagmas nominais ≤ 284,526: classe de textos originais
    a.2. Incidência de sintagmas nominais > 284,526:
        a.2.1. Sobreposição de argumentos (adjacentes) ≤ 0,482: classe de textos traduzidos
        a.2.2. Sobreposição de argumentos (adjacentes) > 0,482: classe de textos originais

Essa análise classificou as métricas de acordo com duas classes: textos originais e textos traduzidos, independentemente da língua. O índice Flesch aparece como o primeiro discriminador: textos com índices *menores* do que 46 são classificados como sendo traduções. Quando o Flesch é *maior* do que 46, há uma nova discriminação: se a incidência de sintagmas nominais for *menor* do que 284,526, os textos são classificados como originais. Quando a incidência de sintagmas nominais é *menor* do que esse valor, é

a incidência de sobreposição de argumentos a próxima métrica discriminativa: quando *menor* do que 0,482, a classe de textos é de traduções; quando *maior* do que 0,482, a classe de textos é a de originais.

Assim, a classificação aponta que as métricas mais discriminativas são o índice Flesch – como raiz da árvore de decisão –, a incidência de sintagmas nominais – primeiro nó da árvore – e a proporção de sobreposição de argumentos em sentenças adjacentes – segundo nó da árvore. No entanto, os resultados de precisão, cobertura e medida-f são relativamente baixos, o que pode ser reflexo da discrepância entre os dados em função de as duas línguas, inglês e português, terem sido consideradas conjuntamente.

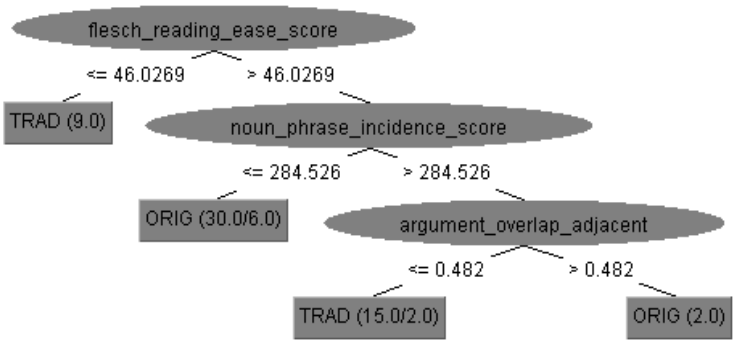

Figura 9. Árvore de decisão da classificação entre textos originais x textos traduzidos.

Precisão	Cobertura	Medida-f	Classe
0,645	0,714	0,678	Originais
0,68	0,607	0,642	Traduções

Tabela 13. Precisão, cobertura e medida-f da classificação entre textos originais x textos traduzidos.

6.2.3 ANÁLISE 3: MÉTRICAS DOS TEXTOS ORIGINAIS EM PORTUGUÊS X MÉTRICAS DOS TEXTOS TRADUZIDOS PARA O PORTUGUÊS

Essa análise classificou as métricas em duas classes: textos originalmente escritos em português e textos traduzidos para o português. Assim, os resultados indicam a discriminação entre as métricas mais características de textos em português, tendo como traço distintivo o fato de serem originais ou fruto de tradução. Os resultados dessa análise, que serão explicados no parágrafo seguinte ao esquema abaixo, são:

a. Índice Flesch ≤ 51,5337: classe de textos em português-tradução
b. Índice Flesch > 51,5337:
    a. Sobreposição de radicais ≤ 0,245979
        (a) Incidência de sintagmas nominais ≤ 248,366: classe de textos em português-tradução
        (b) Incidência de sintagmas nominais > 248,366: classe de textos em português-original

O índice Flesch é o primeiro discriminador: quando menor que 51,5337, trata-se de textos traduzidos para o português; quando maior que 51,5337, é a métrica de sobreposição de radicais que faz a discriminação. Se essa métrica for *menor* do que 0,245979, com incidência de sintagmas nominais *menor* do que 248,366, trata-se de textos traduzidos para o português; se for *maior* do que esse valor, trata-se de textos originalmente escritos em português.

Percebe-se, então, que a métrica mais discriminativa é o índice Flesch, que aparece como raiz da árvore de decisão, indicando que, se o valor do índice for *menor* do 51,53, a classe apontada é a de textos traduzidos para o português. O primeiro nó é a sobreposição de argumentos: se *acima* de 0,24, a classe é a de textos originalmente escritos em português. O segundo nó é a incidência de sintagmas nominais: se *menor* que 248,36, a classe apontada é a de textos traduzidos para o português; se *maior* do que 248,35, a classe apontada é a de textos originalmente escritos em português. Os resultados de precisão, cobertura e medida-f são relativamente bons (quase 80%), indicando validade estatística aceitável para os fins desta pesquisa. A Figura 10 adiante ilustra a árvore de decisão aqui descrita.

Em outras palavras, a classe de textos originalmente escritos em português é caracterizada por incidência *maior* de sintagmas nominais, *maior* proporção de sobreposição de radicais de palavras e índice Flesch *acima* de 51,53. A classe de textos traduzidos para o português é caracterizada por índice Flesch *menor* do 51,53; se *maior* do que esse valor, caracteriza-se por *menor* incidência de sintagmas nominais e *menor* proporção de sobreposição de radicais de palavras.

O atributo principal na discriminação entre textos originalmente escritos em português é o índice Flesch, que indica que textos mais difíceis são mais caracteristicamente fruto de tradução. A incidência de sintagmas nominais teria mais validade como atributo discriminativo se os textos a partir dos quais a análise foi realizada fossem originais e suas respectivas traduções, o que não é o caso dessa análise, pois trata-se da comparação entre classes de texto independentes, cujo único vínculo é o fato de serem em língua portuguesa. Já a proporção de sobreposição de radicais de palavras indica que os textos originalmente escritos em português com índice Flesch superior a 51,13 apresentam maior repetição de vocabulário.

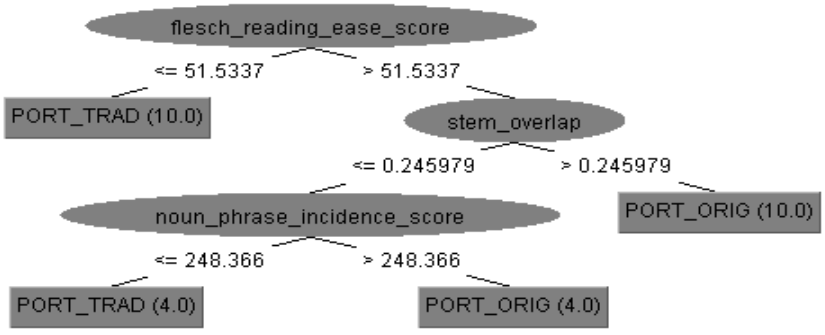

Figura 10. Árvore de decisão da classificação entre textos originais em português x textos traduzidos para o português.

Precisão	Cobertura	Medida-f	Classe
0,786	0,786	0,786	Português-original
0,786	0,786	0,786	Português-tradução

Tabela 14. Precisão, cobertura e medida-f da classificação entre textos originais em português x textos traduzidos para o português.

6.2.4 ANÁLISE 4: MÉTRICAS DOS TEXTOS ORIGINAIS EM INGLÊS X MÉTRICAS DOS TEXTOS TRADUZIDOS PARA O INGLÊS

Os resultados dessa análise foram os seguintes:

a. Incidência de sintagmas nominais ≤ 284,314: classe de textos em inglês-original
b. Incidência de sintagmas nominais > 284,314:
    i. Sobreposição de argumentos (adjacentes) ≤ 0,482: classe de textos em inglês-tradução
    ii. Sobreposição de argumentos (adjacentes) > 0,482: classe de textos em inglês-original

Essa análise classificou as métricas em duas classes: textos originalmente escritos em inglês e textos traduzidos para o inglês. A raiz da árvore é a incidência de sintagmas nominais: se menor do que o valor discriminante 284,31, a classe é a dos textos em inglês-original. O primeiro (e único) nó é a proporção de sobreposição de argumentos em sentenças adjacentes: se menor do que 0,48, a classe indicada é a dos textos em inglês-tradução; se maior do que 0,48, a classe é a dos textos em inglês-original. Os resultados indicam precisão, cobertura e medida-f razoáveis.

A árvore de decisão mostra que valores de sobreposição de argumentos em sentenças adjacentes superiores a 0,48 e incidência de sintagmas nominais superior a 284,31 são discriminadores na classificação de textos originalmente escritos em inglês. Para textos traduzidos para o inglês, ocorre o inverso: valores de sobreposição de argumentos em sentenças adjacentes inferiores a 0,48 e incidência de sintagmas nominais inferior a 284,31 são os discriminadores distintivos. Ou seja, *mais* sintagmas nominais e *menos* sobreposição de argumentos em sentenças adjacentes caracterizam textos originalmente escritos em inglês, e *mais* sintagmas nominais e *mais* sobreposição de argumentos em sentenças adjacentes caracterizam traduções. Isso indica que os textos originalmente escritos em inglês analisados aqui, quando apresentam incidência de sintagmas nominais mais alta, tendem a repetir o vocabulário, ao passo que textos traduzidos tendem a apresentar maior variação lexical.

A Figura 11 a seguir ilustra a árvore de decisão da comparação entre as métricas dos textos originais em inglês x métricas dos textos traduzidos para o inglês.

Figura 11. Árvore de decisão da classificação entre textos originais em inglês x textos traduzidos para o inglês.

Precisão	Cobertura	Medida-f	Classe
0,714	0,714	0,714	Inglês-original
0,714	0,714	0,714	Inglês-tradução

Tabela 15. Precisão, cobertura e medida-f da classificação entre textos originais em inglês x textos traduzidos para o inglês.

7. DISCUSSÃO DOS DADOS OBTIDOS

Após as seções anteriores, em que foram descritos os textos sob estudo, os resultados das métricas das ferramentas Coh-Metrix e Coh-Metrix-Port com diferenças estatisticamente significativas e a análise classificatória por AM, nesta seção serão discutidos alguns pontos de destaque ao longo desta pesquisa. No entanto, em virtude da grande quantidade e variedade de dados levantados, será impossível ponderar detalhadamente sobre todos os tópicos. Em síntese, as discussões apresentadas neste capítulo não esgotarão todas as análises possíveis dos dados obtidos.

7.1 COH-METRIX E COH-METRIX-PORT: PROBLEMAS E ANÁLISES

Em primeiro lugar, é preciso lembrar que as ferramentas Coh-Metrix e Coh-Metrix-Port não foram criadas com a intenção de contrastar traduções. Isso, evidentemente, gerou uma série de dificuldades que precisaram ser contornadas. O primeiro problema foi avaliar quais as métricas, dentre as adaptadas do Coh-Metrix para o Coh-Metrix-Port, seriam equiparáveis. Das 48 métricas disponíveis, somente 31 seriam comparáveis. Num segundo momento, avaliou-se que seria mais produtivo contrastar apenas as métricas estatisticamente relevantes, deixando de lado as métricas sem diferenças significativas. Dentre as 31 métricas, 18 apresentaram diferenças importantes sob o ponto de vista estatístico. Dessas 18 métricas, 6 foram excluídas, por diferentes motivos, conforme explicado nas seções anteriores, restando somente 12 para classificação por métodos de AM. Ainda que as 12 métricas restantes sejam as realmente dignas de mérito, elas não são suficientes para uma descrição mais completa e abrangente das diferenças coesivas entre textos no que diz respeito à questão da complexidade e da inteligibilidade textual. Tampouco se pode categorizar a complexidade textual dos textos analisados a partir de um número tão reduzido de características contrastáveis.

Para resultados mais confiáveis e completos, seria preciso mais do que uma equiparação direta entre as métricas, isto é, seria preciso sincronizar as métricas entre as duas línguas a fim de poder usá-las na comparação dos níveis de inteligibilidade e complexidade textual entre textos originais e traduções, visto que o inglês e o português

são duas línguas com funcionamentos gramaticais bastante distintos. Uma sugestão seria revisar métricas cujos resultados comparativos tenham mostrado as maiores diferenças significativas, como o caso dos pronomes pessoais, pois envolvem condições inerentes da gramática do inglês. Além disso, revisar métricas que repercutem sobre outras, gerando um "efeito cascata" (métricas que contêm outras, por exemplo), pode qualificar o sistema para o português. Seria possível incluir também métricas que ficaram de fora da análise empreendida aqui em função de basearem-se em bancos de dados distintos em português e inglês, como é o caso da métrica frequência de palavras de conteúdo. Indo mais além, poderíamos aventar a criação de métricas específicas para o português, como uma métrica de aferição de pronomes oblíquos para o Coh-Metrix-Port. As métricas que calculam a incidência de conectivos também necessitariam ser reformuladas, como mencionado no Capítulo 6, seção 6.1.2.1.

É importante relembrar também que, no caso do sistema Coh-Metrix, a unidade de processamento é a unidade do texto, destacado em meio a um *corpus*. O movimento analítico *corpus*-texto-*corpus,* embutido no sistema, é realizado a partir de um *corpus* de treinamento que pode ter características bem diferentes dos textos literários aqui em foco, e essa é uma limitação deste trabalho.

Os textos-fonte e as traduções para o inglês, comparados aos textos em português, mostraram algumas métricas com índices de complexidade mais altos. No entanto, não se pode afirmar que a tradução para o português gere um texto mais fácil. Afinal, as métricas que calculam a incidência de sintagmas nominais e a incidência de pronomes pessoais, por exemplo, mostram particularidades específicas do inglês que não correspondem *pari passu* ao português, pois refletem a natureza essencialmente lexicológica do inglês em comparação ao português, uma língua mais organizada pela gramática do que pelo léxico. Além disso, a média do índice Flesch dos textos traduzidos para o português é 48,2, ou seja, textos considerados difíceis, enquanto as médias de todos os outros conjuntos de textos foram superiores a 60, ou seja, textos com nível de dificuldade médio para fácil (ver seção 7.3).

7.2 CLASSIFICAÇÕES POR AM: PROBLEMAS E ANÁLISES

Adotar a metodologia de AM nesta pesquisa foi, sob muitos aspectos, uma aventura. Em primeiro lugar, porque a pesquisadora é *linguista* e nunca havia tido contato com métodos estatísticos que exigissem mais do que o cálculo de uma simples regra de três, e, em segundo lugar, porque a decisão de incluir análises de AM só se deu ao final do curso de mestrado. Contudo, apesar dos percalços inerentes ao encontro de um linguista com o universo assustador dos números, acredito que a decisão de incluir a metodologia de AM na classificação das métricas de coesão textual tenha sido proveitosa, sobretudo porque possibilitou a visualização das relações determinantes entre as métricas na discriminação entre as classes textuais.

Apesar do número reduzido de atributos distintivos (as 12 métricas das ferramentas Coh-Metrix e Coh-Metrix-Port com diferenças estatisticamente significativas), as classificações apresentaram resultados condizentes com as expectativas. Das quatro análises feitas, a que merece maior destaque é a análise comparativa entre textos originais em português e textos traduzidos para o português, pois a métrica mais discriminativa foi o índice Flesch, que, além de ser a única métrica totalmente adaptada ao português, é a única que aponta, em uma escala hierárquica de dificuldade, níveis objetivos de complexidade textual. Além disso, a questão de pesquisa principal deste trabalho refere-se à complexidade e à inteligibilidade de textos em português. Assim, um exame mais detalhado dos resultados envolvendo o índice Flesch se faz necessário.

7.3 ÍNDICE FLESCH

O índice Flesch me fascinou intensamente desde o início desta pesquisa. Não só pela simplicidade e elegância da fórmula, mas pela história de Rudolph Flesch, um advogado exilado num país de língua estranha que veio a se tornar bibliotecário e revolucionou todas as instâncias de comunicação nos Estados Unidos – desde a imprensa

e a comunicação oficial dos órgãos governamentais, até as editoras e os modelos de divulgação do conhecimento científico.

Dentre todas as métricas, o índice Flesch foi a que causou mais impacto em todas as etapas deste trabalho: apontou para uma potencial maior complexidade no nível de inteligibilidade de textos traduzidos para o português e foi o atributo raiz na árvore de decisão classificatória de textos originalmente escritos em português e textos em português fruto de tradução. **Assim, os resultados sugerem não só uma maior complexidade das traduções em relação aos seus textos-fonte em inglês, como também uma maior complexidade dos textos traduzidos para o português em comparação com textos originalmente escritos em língua portuguesa**.

A Tabela 16, a seguir, lista os quatro autores brasileiros e os 14 textos selecionados para este estudo; os índices Flesch dos textos-fonte e das traduções; e os tradutores dos textos para o inglês. Todos os textos-fonte são anteriores a 1950, sendo que nenhum é classificado abaixo da escala de dificuldade de textos razoavelmente difíceis, conforme a classificação proposta pelos desenvolvedores da ferramenta Coh-Metrix-Port (ver Quadro 1, na p. 91). As traduções são relativamente recentes, tendo sido feitas nas últimas duas décadas; somente as traduções de Isaac Goldberg são mais antigas, em torno da década de 1920. Assim, não é surpreendente que os tradutores tenham produzido textos mais fáceis, considerando que atualizaram os textos para um inglês em uso nos dias de hoje.

O único texto cuja tradução tem índice Flesch menor – portanto, com nível de complexidade considerado mais difícil – foi traduzido por Isaac Goldberg, o tradutor mais antigo dentre todos os três. Além disso, o fato de 10 entre os 14 textos terem sido traduzidos pelo mesmo indivíduo (Francis Johnson) também pode revelar uma tendência facilitadora do próprio tradutor, que imprime um estilo próprio às traduções. De qualquer modo, os índices Flesch dos textos-fonte, considerando a época em que foram escritos, revelam que os textos são bastante fáceis.

Autor e texto	Índice Flesch: Texto-fonte em português	Índice Flesch: Tradução para o inglês	Tradutor
001. Lima Barreto - Cazuza	60,56	69,38	Francis Johnson
002. Humberto de Campos - Promessa	54,50	70,16	Francis Johnson
003. Coelho Neto - Firmo	70,07	84,79	Francis Johnson
004. Lima Barreto - Javanês	61,40	71,39	Gregory Rabassa
005. Humberto de Campos - Caveiras	66,76	84,29	Francis Johnson
006. Coelho Neto - Duplo	63,66	80,58	Francis Johnson
007. Lima Barreto - Sepultura	58,21	68,55	Francis Johnson
008. Humberto de Campos - Vingança	52,20	69,22	Francis Johnson
009. Machado de Assis - Cartomante	60,87	69,40	Isaac Goldberg
010. **Machado de Assis - Viver**	71,66	**60,24**	**Isaac Goldberg**
011. Machado de Assis - Cantiga	67,09	78,62	Gregory Rabassa
012. Machado de Assis - Enfermeiro	61,25	79,94	Francis Johnson
013. Machado de Assis - Marcha Fúnebre	65,84	78,62	Francis Johnson
014. Machado de Assis - Eterna	63,35	79,17	Francis Johnson
Média:	62,67	74,60	
DVP:	5,48	7,18	

Tabela 16. Índices Flesch: textos-fonte em português e respectivas traduções para o inglês.

A Tabela 17, abaixo, lista os cinco autores de literatura em língua inglesa e os 14 textos selecionados para este trabalho; os índices Flesch dos textos-fonte e das traduções; e os tradutores dos textos para o português. Nathaniel Hawthorne e Poe são os autores mais antigos, com textos publicados a partir de 1840; os textos restantes foram publicados entre o início do século XX até a década de 1950. Nesta relação contrastiva, o número de tradutores supera o número de autores: temos 5 autores e 9 tradutores. Dos 9 tradutores, todos produziram textos com índice Flesch menor, ou seja, com nível de complexidade maior, do que os textos-fonte em inglês.

Com exceção das traduções de Oscar Mendes, publicadas na década de 1980, todas as outras traduções têm, no máximo, entre 20 e 15 anos, sendo que a tradução do conto de O. Henry, feita por mim, é de 2010. Esses dados são extremamente importantes, pois mostram que, de acordo com os índices Flesch, **100% dos tradutores dos textos aqui analisados produziram traduções contemporâneas com níveis de inteligibilidade não apenas mais complexos do que os textos-fonte, datados do século XIX e do início do século XX, mas mais complexos do que textos originalmente produzidos em português e publicados no século XIX.**

Autor e texto	Índice Flesch: Textos-fonte em inglês	Índice Flesch: Traduções para o português	Tradutor
001. Poe – Oval Portrait	61,06	40,63	Marcelo Bueno
002. Poe – Mesmeric Revelation	55,05	46,03	Oscar Mendes
003. Poe – Black Cat Port	60,75	38,28	Bernardo Carvalho
004. Poe – Imp of Perversity	55,02	42,41	Rodrigo Breunig
005. Poe – Tell-Tale Heart	81,27	67,43	Celina Portocarrero
006. Poe – Berenice	50,79	40,76	Oscar Mendes
007. Poe – Eleonora	53,55	42,04	Oscar Mendes
008. Poe – Red Masque	63,97	44,80	Oscar Mendes
009. Poe – Cask of Amontillado	78,82	62,98	Bernardo Carvalho
010. Poe – Man of Crowd	52,09	39,14	Dorothée de Bruchard
011. James Joyce – Araby	79,92	53,88	Roberto Schmitt-Prym
012. Virginia Woolf – Monday Tuesday	67,93	39,69	Roberto Schmitt-Prym
013. O. Henry – Red Chief	82,26	65,28	Bianca Pasqualini[29]
014. Nathaniel Hawthorne – Wakefield	58,70	51,53	Zaida Maldonado
Média:	64,37	48,20	
DVP:	11,60	10,29	

Tabela 17. Índices Flesch: textos-fonte em inglês e respectivas traduções para o português.

Por fim, como se pode perceber na Tabela 17, acima, a variação entre os índices dos textos-fonte é grande (desvio padrão de 11,6), mas somente 6 ficam abaixo de 60, ao passo que 11 dos textos traduzidos ficam com índices abaixo dessa faixa. Na análise por AM, o valor discriminador do índice Flesch para diferenciar entre textos traduzidos e originais em português é 51,13 (ver Figura 10). Nove dos 14 textos traduzidos para o português têm índices Flesch abaixo desse valor, sendo que **nenhum** dos textos originais em português apresenta índices abaixo disso. A média dos índices Flesch dos textos traduzidos para o português é 48,2, **muito inferior ao recomendado para leitores com proficiência de leitura em nível básico, que é o nível médio de proficiência do leitor brasileiro típico** (ver Tabela 1), pois 48,2 é um valor que indica, de acordo com a classificação de dificuldade textual do índice Flesch mostrado no Quadro 1 (página 91), um nível de dificuldade compatível com a de alunos do Ensino Médio, e a maioria dos

[29] Foi durante a tradução do conto de O. Henry que várias questões sobre o nível de complexidade do texto que eu estava traduzindo surgiram. Posso dizer que esta dissertação é, também, o resultado das inquietações que me acometeram na tradução do conto. Não me surpreendeu perceber que o nível de complexidade da minha tradução foi maior do que a do texto-fonte, e as razões para isso ainda me são desconhecidas. Será um inconsciente coletivo que nos dirige para a maior complexidade? Aqui poderíamos evocar as visões psicanalíticas de tradução.

leitores brasileiros tem letramento básico e proficiência de leitura compatível à de alunos do Ensino Fundamental.

Sabemos, no entanto, que o índice Flesch é considerado um cálculo superficial e temos consciência de que há limitações decorrentes disso. É preciso levar em conta, porém, que o índice Flesch foi mais um entre vários outros elementos nesta pesquisa; mesmo assim, destacou-se em todas as análises e resultados como altamente relevante na comparação dos níveis de inteligibilidade textual do *corpus* estudado, tanto no contraste das métricas das ferramentas Coh-Metrix e Coh-Metrix-Port quanto na classificação das métricas por AM. **Desse modo, o índice Flesch, contextualizado e enriquecido pelo acréscimo de outros elementos textuais e de abordagens estatísticas de análise de complexidade textual, parece ser um indicador bastante confiável de que, no que tange aos textos processados neste trabalho, as traduções para o português são mais complexas do que os seus textos-fonte e são também mais complexas do que os textos dos autores brasileiros selecionados**. É preciso lembrar também da transversalidade semântica proposta por Bouquet (2004), conforme comentamos na sessão de revisão da literatura, no Capítulo 1: **o sentido permeia todas as instâncias da língua, ainda que essas instâncias sejam fragmentos do todo da língua**, ou, o que é o caso agora em questão, sejam constituídas de aspectos lexicais de um texto, em detrimento de outros.

8. RETOMADA DAS HIPÓTESES E DAS QUESTÕES DE PESQUISA

Neste capítulo, voltaremos às hipóteses e questões formuladas no início da dissertação, procurando respondê-las. Em seguida, será feita uma avaliação final da trajetória percorrida, com indicativos e perspectivas futuras de pesquisa. Por fim, são tecidos os comentários finais. Antes de prosseguirmos, é importante reiterar que os comentários tecidos neste capítulo referem-se somente aos textos analisados e investigados *nesta dissertação*, sem fazer generalizações.

8.1 PRIMEIRA HIPÓTESE

— Textos literários em inglês traduzidos para o português brasileiro tendem a ser mais complexos do que os seus textos-fonte.

Essa hipótese se confirma. Evidentemente, é preciso considerar as limitações desta pesquisa, sobretudo no que diz respeito ao tamanho do *corpus* e à variedade de autores e de tradutores, que é bastante restrita. No entanto, no conjunto de textos investigados, a hipótese se confirma.

Na análise dos resultados das métricas obtidos com as ferramentas Coh-Metrix e Coh-Metrix-Port, a confirmação da hipótese de que textos literários em inglês traduzidos para o português brasileiro tendem a ser mais complexos do que os seus textos-fonte ficou, de certa forma, diluída na comparação entre as medidas, pois nem todas indicaram maior complexidade das traduções para o português. Entretanto, na análise por AM, as métricas mais características a cada grupo textual foram determinadas, e aí, sim, pudemos ter mais confiança nessa afirmação, uma vez que o índice Flesch revelou, sem sombra de dúvida, o nível de complexidade maior das traduções aqui investigadas.

8.2 SEGUNDA HIPÓTESE

— O índice Flesch (medida lexical de avaliação do nível de complexidade textual oriunda de trabalhos na área de Processamento de Língua Natural) é um recurso importante para um trabalho linguístico de avaliação de complexidade textual.

Hipótese confirmada. Conforme exposto no Capítulo 7.3, o índice Flesch, contextualizado e enriquecido pelo acréscimo de outros elementos textuais e de abordagens estatísticas de análise de complexidade textual, como a técnica de AM aplicada nesta pesquisa, é um indicador bastante confiável de que as traduções para o português investigadas são mais complexas do que os seus textos-fonte e são também mais complexas do que os textos dos autores brasileiros selecionados, produzidos no século XIX.

Tendo feito essas ponderações sobre as hipóteses das quais partimos, respondamos, agora, às perguntas.

8.3 PRIMEIRA PERGUNTA

— Se textos traduzidos do inglês para o português tenderem a ser mais complexos que os textos-fonte, as traduções do português para o inglês também seriam mais complexas?

Não, não há indicativos de que os textos traduzidos do português para o inglês sejam mais complexos que seus textos-fonte. Na verdade, o que se observou foi o contrário: traduções para o inglês apresentam nível de complexidade inferior ao de seus textos de origem em português, inclusive na análise isolada dos resultados das métricas obtidos com as ferramentas Coh-Metrix e Coh-Metrix-Port. Na análise por AM (ver Capítulo 6, seção 6.2.4), o índice Flesch sequer aparece entre as métricas mais discriminativas entre os textos em inglês. Vê-se, portanto, que as características distintivas entre traduções e textos originais em inglês não envolvem, necessariamente, uma diferença substancial no nível de complexidade entre eles.

8.4 SEGUNDA PERGUNTA

— Qual a contribuição de uma comparação entre a complexidade textual de originais e traduções, nos moldes da pesquisa de PLN, para os estudos linguísticos em geral?

Uma pesquisa de comparação entre a complexidade textual de originais e traduções, nos moldes da pesquisa de PLN, pode contribuir com os estudos linguísticos em geral ao “quantificar” o que o linguista intui. Traz ao linguista uma nova

possibilidade de abordagem textual, em que a individualidade do texto é mantida dentro do conjunto e em que se podem identificar características compartilhadas por textos pertencentes a um determinado grupo. Além disso, torna possível uma avaliação mais objetiva de um traço altamente impalpável e extremamente evasivo como a complexidade de um texto.

Por outro lado, contribui também para uma aproximação entre o que o linguista teoriza e o que o PLN aplica, unindo essas duas perspectivas sobre a língua, uma de cunho mais abstrato, outra de cunho mais prático, apontando para um caminho de conciliação, no campo dos estudos linguísticos, com profissionais de PLN. Ademais, reforça o importante papel desempenhado pelo linguista na produção e na testagem, por meio do uso, de recursos computacionais criados em PLN.

9. PERSPECTIVAS E CONCLUSÕES

A partir do que foi verificado, percebe-se que associar métricas de complexidade e inteligibilidade textual entre originais e traduções pode render importantes *insights* para Estudos de Tradução, de Linguística de Corpus, de PLN e de Linguística em geral. Além disso, ao reiterar-se a percepção de uma tendência para menor inteligibilidade textual quando se traduz textos de literatura para o português do Brasil, conforme aponta o índice Flesch, fica o questionamento se isso ocorreria também na tradução de textos de outra natureza, tais como, por exemplo, o texto científico e o texto jornalístico.

A respeito do referencial teórico de Bouquet, apresentado como sustentação desta pesquisa, cabe dizer que é necessário aprofundar de que forma tais postulados podem ser formalizáveis para aplicação em recursos de PLN, tal como Dias da Silva (2006) propôs para projetos na área. De todo modo, abre-se a perspectiva de trabalho sob essa visão de língua e linguagem, em que um estudo lexical ganha fôlego semântico, o que, acredito, ainda não foi investigado no âmbito dos estudos em PLN.

Ainda no que se refere à contribuição desta pesquisa aos estudos de PLN, acredito que a apresentação de um suporte teórico desvinculado do gerativismo de Chomsky é uma contribuição importante, pois amplia, e muito, as possibilidades "formalizáveis" de língua e linguagem com que os profissionais de PLN vêm trabalhando. Em segundo lugar, nesta investigação ficou clara a utilidade de recursos e ferramentas criados no âmbito do PLN, inclusive para fins completamente diferentes dos pretendidos pelas ferramentas.

Outra perspectiva aberta por este trabalho é a criação de uma ferramenta de análise de métricas de complexidade textual voltada para tradutores. Para que uma ferramenta como essa venha a ser criada, uma diversidade de variáveis precisa ser levada em conta, como, por exemplo, o par de línguas envolvido, a proficiência do tradutor, o público-alvo, os gêneros textuais e aspectos diacrônicos das línguas envolvidas – questões não abordadas por esta investigação.

A existência de uma ferramenta dessa natureza, que auxilie a avaliar os textos em sua complexidade, faz-se ainda mais importante à medida que iniciativas recentes do Ministério da Educação do Brasil (MEC) visam popularizar o acesso a clássicos da literatura nacional e internacional para *neoleitores* por meio de versões mais facilitadas

dos textos para um primeiro contato de leitores iniciantes, em projetos como o "Leitura para Todos" e "É Só o Começo", tal como mencionamos na introdução desta dissertação.

No entanto, as simplificações são feitas sem uma uniformidade conceitual do que vem a ser um texto "facilitado", tampouco se preocupam em definir os aspectos textuais que dificultam a leitura: são feitas a partir da subjetividade do linguista, que altera os textos a seu bel-prazer, de acordo com o que julga ser "mais simples". Não há, no Brasil, uma definição de "português simplificado" – o que se tem, e em abundância, são especulações.

É preciso mencionar também os preconceitos contra qualquer iniciativa de facilitação textual, pois há uma noção, entre o público geral e também entre autoproclamados eruditos e defensores do purismo linguístico, de que simplificar o português é empobrecer a língua. Ora, em primeiro lugar, não se trata de querer empobrecer a língua, pois a língua *não é a escrita*. A escrita é uma *representação* da língua e deve, **sempre**, a ela subjugar-se. E, em função de uma insistência tirânica na superioridade da escrita em detrimento da língua real, há uma parcela da comunidade de falantes que não está tendo acesso à compreensão de textos produzidos em português, o que leva esses leitores a tornarem-se *não leitores,* pois não se sentem bem recebidos pelo texto. Essa é a fonte principal da polêmica em torno da questão da facilitação de textos. E é esse o principal motivo, na minha opinião, que contribui para que o número de leitores no Brasil seja baixo. Esta dissertação, ainda que não tenha se proposto a investigar essas questões, aponta para essa carência e abre a perspectiva de que, em pesquisas futuras, isso seja investigado.

Retomando o comentário sobre as ferramentas aqui utilizadas, fica a perspectiva de avaliar-se o papel das diferentes métricas e contagens já desenvolvidas para textos literários brasileiros, inclusive sem envolver contrapontos com a sua tradução para o inglês, com destaque para aqueles mais demandados como leitura no Ensino Fundamental e Médio e no programa para *neoleitores* do MEC. Assim se poderia também propor a inserção e a testagem de novas medidas, como, por exemplo, a presença de elipses ou omissões de elementos dos textos, ou a incidência de pronomes oblíquos, marca incontestável da língua portuguesa escrita que influencia o nível de complexidade textual.

No que se refere à prática dos profissionais do texto, esta dissertação contribui com a perspectiva de que a avaliação da complexidade textual possa ser um dos critérios que justifiquem escolhas tradutórias e alterações no texto no momento da revisão. Christiane Nord, numa perspectiva funcionalista (1998, p. 98, apud HURTADO ALBIR, 2008, p. 297), classifica os erros de tradução em erros pragmáticos, erros culturais e erros linguísticos. A autora considera os erros pragmáticos os mais graves, pois tais erros são indetectáveis pelo leitor, ou seja, são erros de concepção da tradução, e, retomando a noção de Biber e Conrad, estão ligados sobretudo à noção de gênero. Já os erros culturais estão relacionados ao estilo e às normas da cultura de chegada, como, por exemplo, convenções de medidas, saudações de cortesia, etc. Os erros linguísticos constituem falhas sintáticas e gramaticais e podem ser relacionados à noção de registro de Biber e Conrad. Mas há, também, os erros de *complexidade*, que, como repetido diversas vezes ao longo desta exposição, são difíceis de identificar objetivamente.

Gostaria também de fazer alguns comentários sobre a contribuição desta pesquisa à prática de revisão de textos. A função do revisor costuma restringir-se a identificar erros sintáticos e gramaticais. Contudo, deixar de lado a questão da complexidade e dos erros pragmáticos e culturais põe em risco a qualidade da tradução e sua aceitação por parte da comunidade leitora a que se destina. A relação fundamental do revisor é com o texto de chegada, enquanto o tradutor envolve-se muito mais com o texto de partida. De certa forma, pode-se dizer que o tradutor estabelece uma relação com o autor do texto-fonte, enquanto o revisor estabelece uma relação com o leitor do texto de chegada. Na realidade, ao contrário do que se acredita, a maior parte do trabalho de revisão é identificar erros pragmáticos e culturais. Para tanto, o revisor deve ser capaz de não apenas reconhecer tais erros, mas também de sugerir alterações adequadas e pertinentes à função do texto final. Entretanto, se o nível de complexidade não for avaliado pelo revisor, de nada adiantará ter feito sugestões apenas lexicais ou pragmáticas. Ou seja: o texto pode não ter "erros", mas não será entendido pelo seu público leitor.

Finalizo afirmando que, por trás da proposta desta investigação e que a motivou, está a convicção de que a leitura, seja de textos traduzidos ou não, deve ser *inclusiva*, deve trazer o leitor para o texto, a fim de permitir que a produção artística e intelectual humana de todas as épocas seja compartilhada, e não compartimentalizada por aqueles que dela se apropriam e que lhe atribuem um significado acessível apenas a poucos. E é

por esse motivo que uma abordagem empirista de linguagem e língua serve tão bem a essa proposta, uma vez que é na língua em uso que se desvendam as características e as necessidades dos leitores. Com isso em mente, mesmo que os resultados observados não sejam definitivos, esta pesquisa visa contribuir para a reflexão sobre o papel do tradutor na sociedade como um portal entre línguas e culturas, sobre o texto como registro dos saberes de um povo e sobre o leitor como agente produtor de sentidos.

BIBLIOGRAFIA

Ação Educativa/ Instituto Paulo Montenegro. *Indicador de Alfabetismo Funcional (Inaf) - Principais Resultados.* Relatório, Instituto Paulo Montenegro, 2009.

ALBIR, Hurtado Amparo. *Traducción y traductologia.* Madrid: Cátedra, 2008.

ALPAYDIN, Ethem. *Introduction to machine learning.* 2. Cambridge, Massachusetts: MIT Press, 2010.

ALUISIO, Sandra, Lucia SPECIA, Carolina GASPERIN, e Caroline E. SCARTON. "Readability Assessment for Text Simplification." *The 5th Workshop on Innovative Use of NLP for Building Educational Applications.* Los Angeles, 2010.

AMORIM, G. *Retratos da leitura no Brasil.* São Paulo: Instituto Pró-livro/ Imprensa Oficial do Estado de São Paulo, 2008.

BAKER, Mona. "Corpus Linguistics and Translation Studies – Implications and Applications." In: *Text and Technology: In Honour of John Sinclair*, por M. BAKER, M. G. FRANCIS e E. TOGNINI-BORELLI. Amsterdam & Philadelphia: John Benjamins, 1993.

BAKTHIN, M. *Estética da criação verbal.* São Paulo: Martins-Fontes, 2010.

BASKIN, Wade. "Translator's Introduction." In: *Course in General Linguistics*, por Ferdinand de Saussure. Nova York: The Philosophical Library, 1959.

BASSNETT, Susan. *Estudos de Tradução.* Tradução: Sônia Terezinha Gehring, Letícia Vasconcellos Abreu e Paula Azambuja Antinolfi. Porto Alegre: Editora da UFRGS, 2005.

BEGTHOL, Clare. "The concept of genre and its characteristics." *Bulletin of The American Society for Information, Science and Technology*, 2011: 17-19.

BERBER SARDINHA, Tony. *Linguística de corpus.* São Paulo: Manole, 2004.

BHATIA, Vijay K. *Analysing genre: language use in professional settings.* London: Longman, 1993.

BIBER, Douglas, e Susan CONRAD. *Register, genre and style.* Cambridge: Cambridge, 2009.

BIBER, Douglas, Susan CONRAD, e Randy REPPEN. *Corpus linguistics: investigating language structure and use.* New York: Cambridge University Press, 1998.

BOUQUET, Simon. *Introdução à Leitura de Saussure.* 9. Tradução: Carlos Augusto Lebum Salum e Ana Lucia Franco. São Paulo: Cultrix, 2004.

—. "De um pseudo-Saussure aos textos saussurianos originais." *Letras & Letras*, janeiro - junho de 2009: 161-175.

BOUQUET, Simon. "Saussure's unfinished semantics." In: *The Cambridge Companion to Saussure*, por Carol SANDERS, tradução: Matthew Pires e Carol Sanders. Cambridge: Cambridge University Press, 2006.

BRITTO, Luis P. L. *A sombra do caos: ensino de língua X tradição gramatical.* Campinas: Mercado das Letras, 1997.

CANDIDO JR., Arnaldo, Erick MAZIERO, Caroline GASPERIN, Thiago A. S. PARDO, Lucia SPECIA, e Sandra ALUISIO. "Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese." *NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications.* 2009. 34-42.

CHRISTIE, Agatha. *Dumb witness.* Londres: Collins Crime Club, 1937.

COMPAGNON, Antoine. *O demônio da teoria: literatura e senso comum.* Tradução: Cleonice P. B. Mourão. Belo Horizonte: Ed. UFMG, 1999.

COSERIU, Eugenio. *Lições de linguística geral.* Rio de Janeiro: Ao Livro Técnico, 1980.

DAELEMANS, Walter, e Véronique HOSTE. "Evaluation of Machine Learning Methods for Natural Language Processing Tasks." *Proceedings of the Third International Conference on Language Resources and Evaluation.* 2002. 755-760.

DAVISON, Alice, e Georgia GREEN. *Linguistic complexity and text comprehension - readability issues reconsired.* New Jersey: Lawrenc e Erlbaum Associates, 1988.

DELISLE, Jean, e Judith WOODSWORTH. *Os tradutores na história.* Tradução: Sérgio BATH. São Palo: Ática, 1998.

DEPECKER, L. “Un autre Saussure.” In: *Comprendre Saussure: d'après les manuscrits*, por L. Depecker. Paris: Armand Colin, 2009.

DIAS DA SILVA, Bento. “A face tecnológica dos estudos da linguagem: o processamento automático de línguas naturais.” Tese de doutorado, Universidade Estadual Paulista, Araraquara, 1996.

—. “O estudo linguístico-computacional da linguagem.” *Letras de Hoje*, n. 2, v. 41, junho de 2006: p. 103-138.

DUBAY, William. *The principles of readability.* Costa Mesa, California: Impact Information, 2004.

DURAN, Magali Sanches, e Claudia XATARA. *Critérios para categorização de dicionários bilíngues.* Vol. III, em *As ciências do léxico: lexicologia, lexicografia, terminologia*, por Aparecida Negri ISQUERDO e Ieda Maria ALVES, 311-320. Campo Grande: Ed. UFMS, 2007.

ECO, Umberto. *Quase a mesma coisa.* Tradução: Eliana Aguiar. Rio de Janeiro: Record, 2007.

FARIAS, Virginia Sita. “Dicionários escolares de língua portuguesa: uma breve análise de aspectos macroestruturais.” *Lusorama*, 71-72 de 2007: 160-206.

FINATTO, Maria José B. “Complexidade textual em artigos científicos: contribuições para o estudo do texto científico em português.” *Organon (UFRGS)*, 2011: 30-45.

FISH, Stanley. “Interpreting the Variorum.” *Critical Inquiry*, 1976.

Folha de São Paulo. *Classe C é a única que continua a crescer, aponta FGV.* 7 de Julho de 2011. http://www.folha.uol.com.br/ (acesso em 07 de 07 de 2011).

FOUCAULT, Michel. “What is an author.” In: *Contemporary Literary Criticism*, por Robert DAVIS e Ronald SCHLEIFER, 342-53. New York: Longman, 1994.

FULGÊNCIO, Lúcia, e Yara LIBERATO. *Como facilitar a leitura: como se processa a leitura.* São Paulo: Contexto, 1992.

GADET, Françoise. *Saussure: Une Science de la Langue.* Paris: Presses Universitaires de France, 1996.

GRAESSER, Art, Moongee JEON, Cai ZHIQIANG, e Danielle McNAMARA. *AUTOMATIC ANALYSES OF LANGUAGE, DISCOURSE, AND SITUATION MODELS.* Projeto de pesquisa, Memphis: http://cohmetrix.memphis.edu/CohMetrixWeb2/HelpFile2.htm#CONCSpi, 2001.

GRAESSER, Arthur C., M. A. GERNSBACHER, e S. R. GOLDMAN. “Cognition.” In: *Discourse studies: A multidisciplinary introduction. Vol 1: Discourse as structure and process*, por T. A. (ed.) Van DIJK, 292-319. London: Sage, 1997.

HAENSCH, G., L. WOLF, e R. WERNER. *La lexicografía: de la linguística teórica a la lexicografia prática.* Madrid: Editorial Gredos, 1982.

HALLIDAY, M. A. K., Wolfgang TEUBERG, Collin YALLOP, e Anna CERMAKOVA. *Lexicology and Corpus Linguistics: an introduction.* London/New York: Continuum, 2004.

HARRIS, Roy. “Translator's Introduction.” In: *Course in General Linguistics*, por Ferdinand de Saussure. Londres: Open Court, 1986.

HOEY, Michael. *Patterns of lexis in text.* London: Oxford University Press, 1991.

“Oxford Advanced Learner's Dictionary of Current English.” De A. S. Hornby. Oxford: Oxford University Press, 2003.

ISER, Wolfgang. *Teoría de la recepción.* Madri: Cátedra, 1995.

JACKSON, K. David (ed.). *Oxford Anthology of the Brazilian Short Story.* New York: Oxford, 2006.

KATO, Mary. *O aprendizado de leitura.* São Paulo: Martins Fontes, 1995.

KLEIMAN, Angela. *Leitura: ensino e pesquisa.* Campinas: Pontes, 1987.

—. *Oficina de leitura: teoria e prática.* Campinas: Pontos, 1997.

—. *Os significados do letramento.* Campinas: Mercado das Letras, 1995.

KOCH, Ingedore Villaça, e Maria Elias VANDA. *Ler e compreender: os sentidos do texto.* São Paulo: Contexto, 2007.

LANDAU, Sidney. *Dictionaries: the art and craft of lexicography*. Cambridge: Cambridge University Press, 2001.

LAVIOSA, Sara. "Corpus-based translation studies 15 years on: theory, findings applications." *SYNAPS*, 2010, 24 ed.

LEFFA, Vilson J. "Fatores da compreensão na leitura." *Cadernos do IL* 15, n. 15 (1996): 143-159.

LOPES, Lucelene, Renata VIEIRA, Maria José B. FINATTO, Daniel MARTINS, Adriano ZANETTE, e Luiz Carlos RIBEIRO JR. "Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde." *Revista eletrônica de comunicação, informação & inovação em saúde* 3 (2009): 76-88.

LOPES, Lucelene, Renata VIEIRA, Maria José B. FINATTO, e Daniel MARTINS. "Extracting Compound Terms from Domain Corpora: Combining linguistic and statistical approaches." *Journal of the Brazilian Computer Society*, 2010: 1-13.

MARCUSCHI, Luiz Antônio. *Produção textual, análise de gêneros e compreensão.* São Paulo: Parábola, 2008.

MARTINS, T. B. F., C. M. GHIRALDELO, M. G. NUNES, e O. N. OLIVERIA JÚNIOR. *Readability formulas applied to textbooks in Brazilian-Portuguese.* Technical Report, São Carlos: ICMC/USP, 1996.

Michaelis. *Moderno dicionário inglês-português, português-inglês.* São Paulo: Melhoramentos, 2000.

MILTON, John. "Translation and adaptation studies." In: *Translation Research Project 2*, por Anthony PYM e Alexander PEREKRESTENKO, 51-58. Tarragona: Intercultural Studies, 2009.

MOURA NEVES, Maria Helena de. *Gramática de Usos do Português.* São Paulo: UNESP, 2000.

NEIS, Ignacio Antonio. "A competência de leitura." *Letras de Hoje*, 1982: p.43-57.

NILSSON, Nils J. *Introduction to Machine Learning.* http://robotics.stanford.edu/~nilsson/MLBOOK.pdf. Stanford, California, 1998.

NORD, Christiane. "Translating as a purposeful activity: a prospective approach." *TEFLIN Journal* 17 (Agosto 2006): 131-143.

NORMAND, Claudine. *Saussure.* Tradução: Ana de Alencar e Marcelo Diniz. São Paulo: Estação Liberdade, 2009.

NORMAND, Claudine. "System, arbitrariness, value." Cáp. 6 em *The Cambridge Companion to Saussure*, por Carol SANDERS. Cambridge: Cambridge University Press, 2006.

PARDO, Thiago A. S., e Maria da Graça Volpe NUNES. "Relações Retóricas e seus Marcadores Superficiais Análise de um Corpus de Textos Científicos em Português do Brasil." Relatório Técnico NILC, São Carlos, 2004.

PASQUALINI, Bianca F. *Análise de traduções do conto "O retrato oval", de Edgar Allan Poe, e sua adequação para leitores de Ensino Médio.* Porto Alegre: EdiPUCRS, 2009.

PASQUALINI, Bianca F., Aline EVERS, e Maria José B. FINATTO. *MEDIDAS DE COMPLEXIDADE TEXTUAL ENTRE TRADUÇÕES BRASILEIRAS E ORIGINAIS DE LITERATURA INGLESA: UM ESTUDO-PILOTO BASEADO EM CORPUS.* Porto Alegre: IX Encontro de Linguística de Corpus e IV Escola Brasileira de Computação, 2010.

PASQUALINI, Bianca F., Carolina SCARTON, e Maria José B. FINATTO. "Comparando Avaliações de Inteligibilidade Textual entre Originais e Traduções de Textos Literários." *Anais do 8th Brazilian Symposium in Information and Human Language Technology*, 2010: 30-39.

PAULINO, Graça, e Ivete WALTY. "Leitura literária - enunciação e encenação." In: *Ensaios sobre leitura*, por Hugo MARI, Ivete WALTY e Zélia VERSIANI. Belo Horizonte: Editora Pucminas, 2005.

RIBEIRO, Vera M. "Matriz de referência para a medição do alfabetismo nos domínios do letramento e do numeramento." *Est. Aval. Educ.* 21, n. 45 (jan./abril 2010): 147-168.

RODRIGUES, Cristina Carneiro. *Tradução e diferença.* São Paulo: UNESP, 2000.

ROGER FISHER, Steven. *Uma breve história da linguagem: introdução à origem das línguas.* São Paulo: Novo Século, 2009.

SALUM, Isaac Nicolau. "Prefácio à edição brasileira." In: *Curso de Linguística Geral*, por Ferdinand de Saussure. São Paulo: Cultrix, 2006.

SARTRE, Jean Paul. *O que é literatura?* 3a ed. Tradução: Carlos Felipe Moisés. São Paulo: Ática, 2004.

SAUSSURE, Ferdinand de. *Cours de Linguistique Générale.* Edição: Charles Bally, Albert Sechehaye e Albert Riedlinger. Paris: Payot, 1995.

—. *Course in General Linguistics.* Edição: Charles Bally, Albert Sechehaye e Albert Riedlinger. Tradução: Roy Harris. Londres: Open Court, 1986.

—. *Course in General Linguistics.* Edição: Charles Bally, Albert Sechehaye e Albert Riedlinger. Tradução: Wade Baskin. Nova York: The Philosophical Library, 1959.

—. *Curso de linguística geral.* 27a. ed. Tradução: Antônio CHELINI, José Paulo PAES e Izidoro BLIKSTEIN. São Paulo: Cultrix, 2006.

—. *Curso de Linguística Geral.* 27. Edição: Charles Bally, Albert Sechehaye e Albert Riedlinger. Tradução: Antônio Chelini, José Paulo Paes e Izidoro Blikstein. São Paulo: Cultrix, 2006.

—. *Escritos de Línguística Geral.* 10. Edição: Simon Bouquet e Rudolf Engler. Tradução: Carlos Augusto Leuba Salum e Ana Lucia Franco. São Paulo: Cultrix, 2004.

SCARTON, Carolina, Daniel Machado ALMEIDA, e Sandra ALUISIO. *Coh-Metrix-Port.* 2009. http://caravelas.icmc.usp.br:3000/ (acesso em agosto de 2010).

SCARTON, Carolina, e Sandra Maria ALUÍSIO. "Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o português." *LinguaMática* 2 (2010): 45-62.

SCHMITZ, John Robert. *A problemática dos dicionários bilíngues.* Vol. I, em *As ciências do léxico: lexicologia, lexicografia, terminologia*, por Ana Maria P. P. OLIVEIRA e Aparecida Negri ISQUERDO, 161-170. Campo Grande: Ed. UFMS, 2001.

SOUZA, Jacqueline A. "TIPOLOGIA DE TRAÇOS LINGUÍSTICOS DE TEXTOS DO PORTUGUÊS DO BRASIL DOS SÉCULOS XVI, XVII, XVIII E XIX: UMA PROPOSTA PARA A CLASSIFICAÇÃO AUTOMÁTICA DE GÊNEROS TEXTUAIS." Dissertação de mestrado, UFSCar, São Carlos, 2010.

SPECIA, Lucia. "Translating from Complex to Simplified Sentences." *PROPOR.* Porto Alegre, 2010.

STANLEY, Fish. *Is there a text in this class? The authority of interpretative communities.* Cambridge: Harvard, 1980.

STUBBS, M. *Text and corpus analysis: computer assisted studiesof language and culture.* Oxford: Blackwell Publishers, 1996.

SWALES, John M. *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press, 1990.

TAGNIN, Stella. *O jeito que a gente diz: expressões convencionais e idiomáticas.* São Paulo: Disau, 2005.

TEIXEIRA, Elisa Duarte. *A Linguística de Corpus a serviço do tradutor: proposta de um dicionário de culinária voltado para a produção textual.* Tese, Departamento de Letras Modernas, USP, São Paulo: USP, 2008.

TIEPOLO, Eliane V. "Neoleitores no Brasil alfabetizado." 2008. http://www.ipm.org.br (acesso em 01 de julho de 2011).

TOGNINI BONELLI, E. "Theoretical overview of the evolution of corpus linguistics." In: *The Routledge Handbook of Corpus Linguistics*, por Anne O'KEEFFE e Michael MCCARTHY, 14-27. New York: Routledge, 2010.

TRABANT, Jürgen. "Faut-il défendre Saussure contre sés amateurs?" *Langages*, 2005.

VIEIRA, Renata. "Linguística Computacional: uma entrevista com Renata Vieira." *Revista Virtual de Estudos da Linguagem. ReVEL.* , Vol. 2, n. 3, agosto de 2004.

VOLPE NUNES, Maria da Graça. "O Processamento de Línguas Naturais: para quê e para quem?" Notas Didáticas , ICMC-USP, São Carlos, 2008.

WELKER, Herbert A. *Dicionários: uma pequena introdução à lexicografia.* Brasília: Thesaurus, 2004.

WELKER, Herbert Andreas. "Sobre lexicografia e tradução." *Horizontes de Linguística Aplicada*, n. 6 de 2007: 132-148.

WITTEN, Ian, e Eibe FRANK. *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Elsevier, 2005.

XATARA, Claudia Maria. *Os dicionários bilíngues e o problema da tradução.* Vol. I, em *As ciências do léxico: lexicologia, lexicografia, terminologia*, por Ana Maria Pires de OLIVEIRA e Aparecida Negri ISQUERDO, 181-187. Campo Grande: Ed. UFMS, 2001.

ZGUSTA, Ladislav. *Manual of lexicography.* Paris: Mouton, The Hague, 1971.

ANEXO A: LISTA DE CONECTIVOS

Coh-Metrix e Coh-Metrix-Port

Conectivos Aditivos	Conectivos Aditivos	Conectivos Causais	Conectivos Causais
a propósito	after all	a fim de	a consequence of
adicionalmente	again	a partir de	after all
afinal de contas	all in all	afinal de contas	arise from
além de	also	agora que	arise out of
além disso	and	assim	as a consequence
além disto	as a final point	através	as a result
ao invés de	as well	aí	as soon as
ao mesmo tempo	at least	cada vez que	because
apesar de	besides	com	by
assim	by the way	como	Cause
bem como	correspondingly	como consequência	conditional upon
como	finally	como resultado	consequently
como exemplo	first (next/second)	condicional a	due to
como também	for example	condicional à	Enable
de fato	for instance	consequentemente	even then
de novo	fortunately	dado que	follow that
dessa forma	further	daí	For
desse modo	furthermore	desde	for (the/these/that)
e	in actual fact	dessa forma	purpose
em adição	in addition	dessa maneira	hence
em geral	in fact	desse modo	if
em paralelo	in other words	devido a	in case
em resumo	in sum	devido à	in order that
em segundo lugar	incidentally	então	it follow that
em vez de	instead	**habilita**	it follows
felizmente	it follows	logo	make
finalmente	moreover	mesmo assim	now that
incidentalmente	next	nesse caso	on (the)* condition that
inclusive	on (the)* one hand	nesse contexto	on condition that
isto é	once again	nesse sentido	only if
na verdade	secondly	neste caso	provided that
novamente	similarly	para	purpose (of/for) which
ou seja	summarizing	para esse fim	pursuant to
pelo menos	summing up	para isso	since
por exemplo	that is (to say)*	para tanto	so
por fim	thereupon	pois	the consequence of
por outro lado	to (these/this) ends	por causa	then again
resumindo	to conclude	por conseguinte	therefore
segue	to return to	por essa razão	thus
segue que	to sum up	por esse motivo	to (these/this) ends
similarmente	to summarize	por esta razão	to that end
também	to take an example	por este motivo	to those ends
ainda	too	por fim	Whenever
alternativamente	at any rate	por isso	Although
antes	alternatively	porque	even though
contudo	and conversely	portanto	nevertheless
em contraste	anyhow	propósito da	nonetheless
entretanto	but	propósito de	though
mas	by contrast	propósito do	unless
pelo contrário	contrasted with	quer dizer que	
porém	except that	relativo a	
todavia	however	relativo à	
	in contrast	se	
FALTAM:	notwithstanding that	sempre que	
Ademais	on the (one/other) hand	sendo assim	
Além do mais	on the contrary	somente se	
Ao contrário (de)	or (else)*	uma vez que	
Ao menos	otherwise	visto isso	
Ao passo que	rather	é claro	
Por conseguinte	whereas	a menos que	
Paralelamente a	yet	apesar de	

A despeito de Em primeiro lugar Em outras palavras Exceto (que) Em comparação No mínimo Uma vez mais Próximo		apesar disso apesar disto contudo embora mesmo embora no entanto FALTAM: Apenas se Contanto que Com a finalidade de	
Conectivos Temporais	**Conectivos Temporais**	**Conectivos Lógicos**	**Conectivos Lógicos**
à medida que	(an/one/two etc.) hour	a seguir	a consequence of
a partir de	later	além da	actually
a seguir	A consequence of	além de	all in all
agora que	after (a/some) time	além disso	also
além de	after (this/that/all)*	além disto	anyway
anteriormente	again	além do	arise from
antes	all this time	ao invés de	arise out of
ao passo que	as	após	as a consequence
apenas	as (long/soon) as	assim	as a final point
assim	as a consequence	bem como	as a result
até agora	at first .. in the end	caso	as if
até aqui	at first… finally	como	as well
até que	at last	como consequência	at least
atualmente	at once	como exemplo	at this point
bem	at the same time	como resultado	because
bem como	at this (moment/point)	como também	besides
daqui a pouco	before	consequentemente	cause
de novo	by this time	dado que	conditional upon
de repente	earlier	de novo	consequently
depois	even then	de qualquer forma	correspondingly
durante o tempo	finally	de qualquer maneira	due to
em breve	first (next/second)	de qualquer modo	enable
em condições de	first then	depois	essentially then
em outra hora	follow that	desde que	even then
em outra ocasião	from now on	devido a	finally
em outra vez	further	devido à	first ¡¬ (next/second)
em outro momento	immediately	em conclusão	first ¡¬ then
em seguida	in the meantime	em geral	follow that
enquanto	instantly	em outras palavras	For
finalmente	It follows (that)*	em paralelo	for (the/these/that)
imediatamente	just before	em resumo	purpose
já	later	em vez de	for example
logo	meanwhile	enquanto	for instance
logo após	next	então	fortunately
mais ao longo	now that	felizmente	further
mais distante	on another occasion	finalmente	furthermore
mais tarde	once again	habilita	hence
mais tarde	once more	isto é	if
mais uma vez	only when	logo	in (short/brief)
nesse momento	presently	mesmo que	in actual fact
nesse ponto	previously	na realidade	in any (case/event)
neste momento	secondly ¡¬	na verdade	in case
neste ponto	simultaneously	ou seja	in conclusion
novamente	since	para	in fact
pelo menos	so far	para dar um exemplo	in order that
primeira coisa que	soon	para este fim	in other words
primeiramente	suddenly	pelo menos	in sum
quando	the consequence of	por exemplo	incidentally
simultaneamente	the last time	por isso	instead
subitamente	the previous moment	por outro lado	it follow that
tardio	then (again/at last)*	porque	likewise
todas as vezes que	this time	portanto	moreover
última vez	throughout	posteriormente	Next
um momento antes de	to that end	primeiramente	on (the)* condition that
vezes	up till that time	resumindo	on (the)* one hand
	up to now	se	on condition that

FALTAM: Ao longo de Brevemente Até o momento Previamente Nesse meio-tempo Ínterim	when whenever while until (then)*	segue que sendo assim similarmente somente se a menos que ainda alternativamente ao acaso ao menos que apesar de apesar disso apesar disto contudo de outra forma de outra maneira de outro modo em todo caso embora entretanto exceto que mesmo embora nem no entanto ou então pelo contrário por outro lado FALTAM: A despeito do fato A despeito de Ademais	once again Only if provided that purpose (of/for) which pursuant to secondly similarly since so summarizing summing up That is (to say)* the consequence of Then Then again therefore thereupon Thus to (these/this) ends to conclude to return to to sum up to summarize to take an example to that end to those ends Well, at any rate while admittedly x, but y alternatively although and conversely anyhow But by contrast contrasted with despite the fact even though except that however in contrast nevertheless nonetheless Nor notwithstanding that on the (one/other) hand on the contrary or (else)* otherwise rather though unless whereas yet

ANEXO B: ESTATÍSTICAS COH-METRIX E COH-METRIX-PORT

PORTUGUÊS - INGLÊS				**INGLÊS - PORTUGUÊS**			
PORTUGUÊS		**INGLÊS**		**INGLÊS**		**PORTUGUÊS**	
MÉDIA	**DSVP**	**MÉDIA**	**DSVP**	**MÉDIA**	**DSVP**	**MÉDIA**	**DSVP**

NÚMERO PALAVRAS	1613,21	323,12	1758,29	379,63	1855,04	626,40	1823,04	555,66
NÚMERO SENTENÇAS	121,46	40,00	131,96	33,82	104,25	58,76	118,75	60,91
NÚMERO PARÁGRAFOS	42,89	22,44	42,68	22,41	24,18	23,25	27,50	22,72
PALAVRAS POR SENTENÇAS	14,07	2,90	13,85	2,96	20,08	6,28	17,24	5,30
SENTENÇAS POR PARÁGRAFOS	3,24	1,22	3,52	1,19	5,57	2,31	5,21	2,25
SÍLABAS POR PALAVRAS	**2,64**	**0,11**	**1,40**	**0,08**	**1,44**	**0,09**	**2,85**	**0,14**
FLESCH	**62,67**	**5,48**	**74,60**	**7,19**	**64,37**	**11,61**	**48,20**	**10,29**
TODOS CONECTIVOS	87,11	8,77	73,24	13,16	84,62	13,34	89,11	11,70
CONECTIVOS ADTIVOS POSITIVOS	34,81	4,15	31,35	6,22	38,55	10,69	35,34	4,26
CONECTIVOS ADTIVOS NEGATIVOS	9,10	2,28	9,68	2,42	13,76	3,05	9,54	3,87
CONECTIVOS TEMPORAIS POSITIVOS	**14,88**	**2,35**	**11,03**	**2,65**	**10,29**	**2,57**	**14,06**	**3,86**
CONECTIVOS TEMPORAIS NEGATIVOS	0,31	0,44	0,39	0,29	0,50	0,39	0,38	0,46
CONECTIVOS CAUSAIS POSITIVOS	**34,85**	**5,85**	**21,10**	**5,04**	**22,27**	**5,86**	**38,42**	**5,05**
CONECTIVOS CAUSAIS NEGATIVOS	0,49	0,50	0,56	0,57	0,72	0,72	1,78	0,96
CONECTIVOS LÓGICOS POSITIVOS	**28,39**	**5,22**	**18,63**	**4,39**	**20,93**	**6,92**	**31,54**	**4,44**
CONECTIVOS LÓGICOS NEGATIVOS	**4,63**	**1,41**	**11,06**	**3,16**	**14,97**	**3,53**	**4,24**	**2,07**
NEGAÇÕES	**5,29**	**3,55**	**7,78**	**2,78**	**11,34**	**3,94**	**3,65**	**1,57**
OPERADORES LÓGICOS	45,10	6,36	41,99	6,10	54,10	9,64	47,48	8,63
HIPERÔNIMOS DE VERBOS	**0,37**	**0,04**	**1,54**	**0,22**	**1,63**	**0,09**	**0,41**	**0,09**
MÍNIMO DE FREQUÊNCIA DE PALAVRAS DE CONTEÚDO	**67126,39**	**38884,93**	**292,18**	**222,78**	**164,48**	**176,05**	**46531,41**	**62012,25**
FREQUÊNCIA DE PALAVRAS DE CONTEÚDO	**229164,25**	**53426,18**	**2378,02**	**379,96**	**2261,99**	**778,54**	**201565,26**	**53501,01**
SINTAGMAS NOMINAIS	**248,88**	**14,39**	**298,74**	**12,44**	**278,53**	**19,36**	**234,35**	**23,83**
MODIFICADORES POR SINTAGMAS	**0,53**	**0,05**	**0,74**	**0,16**	**0,81**	**0,11**	**0,59**	**0,06**
PRONOMES POR SINTAGMAS	**0,23**	**0,07**	**0,37**	**0,08**	**0,35**	**0,11**	**0,21**	**0,05**
PRONOMES PESSOAIS	**14,75**	**7,36**	**111,74**	**26,74**	**97,99**	**33,73**	**13,23**	**8,59**
TIPO TOKEN	0,68	0,05	0,65	0,05	0,69	0,08	0,70	0,08
PALAVRAS ANTES DE VERBOS	2,74	0,92	3,02	0,99	4,19	1,56	3,76	1,24
REFERÊNCIA ANAFÓRICA	**0,44**	**0,22**	**0,24**	**0,08**	**0,27**	**0,11**	**0,49**	**0,22**
REFERÊNCIA ANAFÓRICA ADJACENTE	**0,31**	**0,16**	**0,40**	**0,12**	**0,49**	**0,18**	**0,34**	**0,17**
SOBREPOSIÇÃO DE PALAVRAS DE CONTEÚDO (ADJACENTE)	**0,18**	**0,07**	**0,08**	**0,02**	**0,08**	**0,04**	**0,21**	**0,08**
SOBREPOSIÇÃO DE RADICAL DE PALAVRAS (ADJACENTE)	**0,33**	**0,11**	**0,13**	**0,05**	**0,14**	**0,08**	**0,40**	**0,13**
SOBREPOSIÇÃO DE ARGUMENTOS (ADJACENTE)	**0,20**	**0,09**	**0,36**	**0,10**	**0,44**	**0,14**	**0,25**	**0,13**
SOBREPOSIÇÃO DE RADICAL DE PALAVRAS	**0,26**	**0,08**	**0,10**	**0,03**	**0,12**	**0,05**	**0,30**	**0,14**
SOBREPOSIÇÃO DE ARGUMENTOS	**0,15**	**0,05**	**0,29**	**0,07**	**0,37**	**0,12**	**0,19**	**0,11**

ANEXO C: ARQUIVOS ARFF

Análise 1: Textos em inglês e textos em português

@relation 'textos em inglês e textos em português'

@attribute title string

@attribute argument_overlap_adjacent numeric
@attribute stem_overlap_adjacent numeric
@attribute anaphor_reference_adjacent numeric
@attribute argument_overlap numeric
@attribute stem_overlap numeric
@attribute anaphor_reference numeric
@attribute noun_phrase_incidence_score numeric
@attribute ratio_of_pronouns_to_noun_phrases numeric
@attribute personal_pronoun_incidence_score numeric
@attribute number_of_paragraphs numeric
@attribute number_of_sentences numeric
@attribute number_of_words numeric
@attribute average_sentences_per_paragraph numeric
@attribute average_words_per_sentence numeric
@attribute average_syllables_per_word numeric
@attribute flesch_reading_ease_score numeric
@attribute mean_number_of_modifiers_per_noun_phrase numeric
@attribute mean_number_of_words_before_the_main_verb numeric
@attribute type_token_ratio numeric
@attribute proportion_of_content_words_that_overlap_between_adjacent_sentences numeric
@attribute class {EN,PORT}

@data
text1-
poe_en,0.551,0.163,0.633,0.418,0.178,0.298,243.808,0.352,85.913,7,50,1292,7.143,25.84,1.413,61.068,0.867,6.84,0.711,0.07,EN
text2-
poe_en,0.43,0.365,0.28,0.3325,0.245,0.1565,281.952,0.272,76.6135,38,108,1872.5,2.8435,17.3415,1.586,55.0575,0.8025,4.0995,0.563,0.1015,EN
text3-
poe_en,0.5415,0.105,0.6355,0.5195,0.0945,0.332,278.5845,0.4015,111.879,13.5,91.5,1934.5,6.7855,21.1435,1.473,60.7585,0.7705,4.29,0.732,0.0825,EN
text4-
poe_en,0.562,0.14,0.678,0.477,0.085,0.368,276.19,0.403,111.387,16,122,2415,7.625,19.795,1.557,55.021,0.727,4.066,0.707,0.108,EN
text5-
poe_en,0.51,0.134,0.618,0.446,0.069,0.453,292.556,0.468,137.032,16,158,2109,9.875,13.348,1.324,81.276,0.645,2.532,0.574,0.136,EN
text6-
poe_en,0.498,0.14,0.5355,0.413,0.1035,0.276,267.831,0.353,95.2565,12,66.5,1606,5.7785,24.4195,1.5515,50.7925,0.8755,3.764,0.7415,0.0975,EN
text7-
poe_en,0.581,0.149,0.635,0.446,0.184,0.317,277.089,0.31,86.022,19,75,2418,3.947,32.24,1.425,53.556,0.843,7.373,0.662,0.066,EN
text8-
poe_en,0.311,0.204,0.301,0.201,0.142,0.107,256.188,0.156,40.017,15,104,2424,6.933,23.308,1.409,63.976,0.969,4.481,0.654,0.066,EN
text9-
poe_en,0.321,0.04,0.438,0.261,0.034,0.263,320.924,0.479,153.616,90,250,2337,2.778,9.348,1.401,78.822,0.635,1.856,0.649,0.155,EN
text10-
poe_en,0.4335,0.064,0.5635,0.379,0.0805,0.304,267.8525,0.279,74.7405,11,66.5,1749,6.5,27.266,1.502,52.0905,0.898,4.085,0.757,0.0465,EN
text1-
litingl_en,0.58,0.12,0.633,0.493,0.089,0.354,299.363,0.465,139.278,38,151,2355,3.974,15.596,1.313,79.925,0.784,3.344,0.666,0.101,EN
text2-litingl_en,0.1,0.1,0.05,0.129,0.123,0.044,284.314,0.149,42.484,7,21,306,3,14.571,1.467,67.937,0.92,4.143,0.834,0.02,EN
text3-
litingl_en,0.421,0.114,0.459,0.3595,0.097,0.324,291.3995,0.401,116.667,50.5,145,2145.5,2.8765,14.8755,1.294,82.2635,0.668,2.3505,0.635,0.0725,EN
text4-
litingl_en,0.3175,0.083,0.453,0.275,0.1205,0.1885,261.353,0.386,101.003,5.5,51,1007,7.9375,22.0025,1.487,58.702,0.879,5.4905,0.813,0.0525,EN
001.lima.barreto.cazuza.port.TXTparte1.txt,0.217391,0.4,0.181034,0.146177,0.250225,0.318966,243.161,0.246201,12.1581,45,116,1645,2.57778,14.181,2.67417,60.5612,0.4525,2.68966,0.696677,0.26087,PORT
002.humberto.campos.promessa.port.txt,0.1922725,0.281818,0.36174,0.150967,0.249163,0.509459,284.526,0.184442,5.89292,22,78.5,1253,3.55357,17.21145,2.79626,54.50415,0.515039,4.5122,0.752203,0.096818,PORT
003.coelho.neto.firmo.port.TXTparte1.txt,0.100775,0.170543,0.330769,0.0911151,0.141205,0.369231,255.276,0.183345,13.6147,65,130,1469,2,11.3,2.5345,70.066,0.554667,2.54615,0.702924,0.0930233,PORT
004.lima.barreto.javanês.port.txt,0.188734,0.3001165,0.222231,0.1475665,0.261363,0.327906,241.711,0.277004,15.2226,44,115,1539.5,2.706975,13.42255,2.672775,61.40485,0.515081,2.089775,0.715465,0.131707,PORT
005.humberto.campos.caveiras.port.TXTparte1.txt,0.135135,0.234234,0.0803571,0.0883205,0.157658,0.116071,258.871,0.268817,16.129,39,112,1240,2.87179,11.0714,2.62158,66.7601,0.445483,2.89286,0.709352,0.135135,PORT
006.coelho.neto.duplo.port.TXTparte1.txt,0.45679,0.555556,0.0853659,0.288768,0.370069,0.195122,238.594,0.379833,11.9671,34,82,1337,2.41176,16.3049,2.71773,63.655,0.53605,2.57317,0.774468,0.222222,PORT

007.lima.barreto.número.sepultura.port.txt,0.2109245,0.3957985,0.5545545,0.137802,0.31453,0.69845,241.0095,0.2272185,14.5815,52,103,1540.5,1.96927,15.6873,2.718755,58.2148,0.542219,3.0812,0.67902,0.247059,PORT
008.humberto.campos.vinganÇõa.port.txt,0.151786,0.3125,0.424779,0.159608,0.243837,0.486726,268.085,0.110907,4.25532,18,113,2115,6.27778,18.7168,2.83962,52.1974,0.552028,4.99115,0.718696,0.160714,PORT
001.machado.cartomante.port.txt,0.1540405,0.2590185,0.5305905,0.1497375,0.2637215,0.896063,241.6365,0.1765985,18.4421,30,113.5,1543.5,3.772525,13.5482,2.63982,60.8679,0.54723,2.17744,0.6718225,0.1349205,PORT
002.machado.viver.port.txt,0.08299425,0.1438525,0.10221695,0.0664613,0.1282905,0.169167,255.231,0.3394305,34.08935,63.5,236.5,1685.5,3.671685,6.932825,2.51247,71.66345,0.598677,2.07288,0.589336,0.0718646,PORT
003.machado.cantiga.port.TXTparte1.txt,0.273684,0.389474,0.40625,0.175,0.32193,0.625,234.838,0.243544,11.9887,29,96,1418,3.31034,14.7708,2.45863,67.0887,0.597598,2.05208,0.613475,0.273684,PORT
004.machado.enfermeiro.port.txt,0.1858245,0.3945855,0.3502645,0.142883,0.331777,0.484025,248.366,0.2019005,13.0719,23,106.5,1530,4.645835,14.38175,2.6293,61.2499,0.504197,2.18707,0.685674,0.2528785,PORT
005.machado.marcha.fúnebre.port.txt,0.2167115,0.4183085,0.3943955,0.171942,0.3445855,0.5491575,242.2145,0.1793665,11.05845,33.5,126.5,1880.5,3.875545,15.06985,2.54492,65.83525,0.5410775,2.447435,0.6432325,0.213838,PORT
006.machado.vida.eterna.port.txt,0.236583,0.3457725,0.339491,0.1749515,0.286537,0.4768925,230.8105,0.1907825,23.97465,102.5,172,2388.5,1.678065,14.3395,2.627375,63.34945,0.581404,2.06455,0.6018745,0.1917555,PORT
text1-
litbras_en,0.464,0.161,0.446,0.372,0.1,0.327,301.282,0.381,114.85,41,113,1872,2.756,16.566,1.426,69.381,0.617,3.77,0.658,0.098,EN
text2-
litbras_en,0.3435,0.181,0.415,0.2655,0.152,0.201,286.535,0.327,93.6705,21,73,1374.5,3.5715,19.4645,1.382,70.1615,0.844,4.9335,0.735,0.062,EN
text3-
litbras_en,0.2,0.069,0.325,0.183,0.084,0.189,302.57,0.369,111.565,69,161,1712,2.333,10.634,1.315,84.792,0.687,2.317,0.648,0.07,EN
text4-
litbras_en,0.3195,0.049,0.433,0.2755,0.0605,0.2815,310.2145,0.416,129.2115,51.5,146,1673,2.8405,11.461,1.4635,71.39,0.6805,2.1505,0.6945,0.0675,EN
text5-
litbras_en,0.267,0.164,0.302,0.213,0.115,0.168,287.729,0.36,103.667,44,117,1418,2.659,12.12,1.303,84.299,0.725,3.034,0.643,0.079,EN
text6-
litbras_en,0.482,0.105,0.5,0.351,0.071,0.291,319.322,0.454,144.918,42,115,1594,2.738,13.861,1.326,80.586,0.611,2.426,0.668,0.116,EN
text7-
litbras_en,0.4685,0.1835,0.5105,0.327,0.116,0.2815,295.9155,0.377,111.742,47.5,110,1702,2.357,16.4165,1.4375,68.5595,0.7415,3.246,0.6595,0.1,EN
text8-
litbras_en,0.26,0.165,0.331,0.254,0.154,0.144,276.351,0.247,68.202,21,128,2258,6.095,17.641,1.415,69.22,0.965,5.055,0.658,0.055,EN
text1-
machado_en,0.3555,0.096,0.461,0.307,0.1155,0.242,286.281,0.4095,117.4525,31,128.5,1949.5,4.2265,15.125,1.443,69.4055,0.649,3.053,0.6355,0.0625,EN
text2-
machado_en,0.1685,0.125,0.101,0.1475,0.1085,0.0735,291.36,0.192,61.834,41.5,158.5,1340.5,3.8485,8.393,1.632,60.249,1.1735,1.684,0.586,0.0475,EN
text3-
machado_en,0.44,0.233,0.405,0.312,0.16,0.219,300.06,0.317,94.979,28,117,1653,4.179,14.128,1.346,78.623,0.754,3.171,0.576,0.122,EN
text4-
machado_en,0.479,0.086,0.581,0.3995,0.0595,0.382,318.1105,0.4795,152.5515,30,150.5,1913,5.0165,12.7395,1.347,79.948,0.563,2.3985,0.646,0.097,EN
text5-
machado_en,0.375,0.1245,0.425,0.285,0.084,0.2625,299.716,0.382,114.6305,24.5,112.5,1423.5,4.6365,12.6655,1.3635,78.627,0.6495,2.5905,0.6995,0.081,EN
text6-
machado_en,0.374,0.091,0.424,0.305,0.0735,0.2795,306.9805,0.4725,145.0735,105.5,217.5,2733,2.074,12.75,1.356,79.176,0.67,2.479,0.5665,0.0845,EN
001.poe.oval.portrait.port.txtparte1.txt,0.32,0.56,0.627451,0.305098,0.505098,0.666667,229.35,0.333313,7.90861,9,51,1138,5.66667,22.3137,2.91821,40.6316,0.655172,4.96078,0.756173,0.16,PORT
002.poe.mesmeric.revelation.port.txt,0.252364,0.356721,0.215038,0.1110633,0.2297005,0.297078,242.3935,0.179732,7.93387,38.5,134.5,1765,3.491215,13.15235,2.91515,46.0269,0.6340425,3.0717,0.581353,0.2922155,PORT
003.poe.black.cat.port.txt,0.2458415,0.336944,0.405026,0.169322,0.2871545,0.571882,235.3185,0.2237775,12.6017,17,103,1704.5,6.10764,16.6729,3.03418,38.2795,0.598444,3.50607,0.742789,0.1260156,PORT
004.poe.imp.of.perversity.port.txtparte1.txt,0.364964,0.510949,0.181159,0.159314,0.299164,0.297101,220.889,0.186899,5.33333,19,138,2250,7.26316,16.3043,2.97991,42.4061,0.591549,3.47826,0.673107,0.277372,PORT
005.poe.tell-
tale.heart.port.txtparte1.txt,0.238636,0.392045,0.310734,0.116911,0.232987,0.446328,196.866,0.259476,32.3213,22,177,2042,8.04545,11.5367,2.56961,67.4281,0.584577,2.44068,0.556027,0.215909,PORT
006.poe.berenice.port.txt,0.1944445,0.378788,0.3064385,0.1587455,0.29549,0.5760275,231.851,0.222835,10.99825,19.5,86.5,1620.5,4.421055,19.077,2.978345,40.75865,0.626173,4.966025,0.751642,0.2392675,PORT
007.poe.eleonora.port.txt,0.5625,0.6375,0.518519,0.441975,0.574383,0.679012,220.423,0.20656,9.89654,26,81,2223,3.11538,27.4444,2.84341,42.0364,0.577551,5.24691,0.680199,0.225,PORT

008.poe.red.masque.port.txt,0.339623,0.5,0.196262,0.23964,0.383883,0.327103,233.808,0.145295,3.67478,22,107,2177,4.86364,20.3458,2.91905,44.7998,0.587426,4.71963,0.684383,0.396226,PORT
009.poe.cask.of.amontillado.port.txt,0.0557621,0.133829,0.148148,0.032521,0.0846482,0.240741,233.01,0.18305,17.4757,94,270,2060,2.87234,7.62963,2.71548,62.9792,0.514583,1.85926,0.678661,0.0892193,PORT
010.poe.man.of.crowd.port.txt,0.3041355,0.444314,0.426018,0.2665825,0.4268845,0.635296,214.314,0.1822725,20.22775,16,107.5,2543.5,7.413045,23.24665,2.91838,39.1402,0.6969545,3.44251,0.691412,0.199013,PORT
001.james.joyce.araby.port.txt,0.0927152,0.225166,0.282895,0.0942837,0.174974,0.427632,244.18,0.171796,14.3635,36,152,2019,4.22222,13.2829,2.81865,53.8756,0.555781,2.76316,0.729706,0.10596,PORT
002.virginia.woolf.monday.tuesday.port.txt,0.0588235,0.411765,0.166667,0.0784314,0.150327,0.222222,304.918,0.282038,3.27869,7,18,305,2.57143,16.9444,2.87701,39.6915,0.494624,5.66667,0.86631,0.294118,PORT
003.O.Henry.Red.Chief.port.txt,0.2485955,0.3635005,0.3370575,0.167234,0.245979,0.46674,239.614,0.18562,27.3452,50.5,154,1999,3.054705,13.0114,2.589745,65.2826,0.5042335,2.11719,0.640197,0.2094015,PORT
004.nathaniel.hawthorne.wakefield.port.txt,0.2540675,0.4184205,0.658071,0.2827545,0.3662415,1.0150045,234.03,0.2407515,11.80988,8.5,83,1676,9.833335,20.34205,2.799155,51.5337,0.6495435,4.42089,0.7442545,0.146141,PORT

Análise 2: Textos originais e textos traduzidos

@relation 'textos originais e textos traduzidos'

@attribute title string
@attribute argument_overlap_adjacent numeric
@attribute stem_overlap_adjacent numeric
@attribute anaphor_reference_adjacent numeric
@attribute argument_overlap numeric
@attribute stem_overlap numeric
@attribute anaphor_reference numeric
@attribute noun_phrase_incidence_score numeric
@attribute ratio_of_pronouns_to_noun_phrases numeric
@attribute personal_pronoun_incidence_score numeric
@attribute number_of_paragraphs numeric
@attribute number_of_sentences numeric
@attribute number_of_words numeric
@attribute average_sentences_per_paragraph numeric
@attribute average_words_per_sentence numeric
@attribute average_syllables_per_word numeric
@attribute flesch_reading_ease_score numeric
@attribute mean_number_of_modifiers_per_noun_phrase numeric
@attribute mean_number_of_words_before_the_main_verb numeric
@attribute type_token_ratio numeric
@attribute proportion_of_content_words_that_overlap_between_adjacent_sentences numeric
@attribute class {ORIG,TRAD}

@data
text1-
poe_en,0.551,0.163,0.633,0.418,0.178,0.298,243.808,0.352,85.913,7,50,1292,7.143,25.84,1.413,61.068,0.867,6.84,0.711,0.07,ORIG
text2-
poe_en,0.43,0.365,0.28,0.3325,0.245,0.1565,281.952,0.272,76.6135,38,108,1872.5,2.8435,17.3415,1.586,55.0575,0.8025,4.0995,0.563,0.1015,ORIG
text3-
poe_en,0.5415,0.105,0.6355,0.5195,0.0945,0.332,278.5845,0.4015,111.879,13.5,91.5,1934.5,6.7855,21.1435,1.473,60.7585,0.7705,4.29,0.732,0.0825,ORIG
text4-
poe_en,0.562,0.14,0.678,0.477,0.085,0.368,276.19,0.403,111.387,16,122,2415,7.625,19.795,1.557,55.021,0.727,4.066,0.707,0.108,ORIG
text5-
poe_en,0.51,0.134,0.618,0.446,0.069,0.453,292.556,0.468,137.032,16,158,2109,9.875,13.348,1.324,81.276,0.645,2.532,0.574,0.136,ORIG
text6-
poe_en,0.498,0.14,0.5355,0.413,0.1035,0.276,267.831,0.353,95.2565,12,66.5,1606,5.7785,24.4195,1.5515,50.7925,0.8755,3.764,0.7415,0.0975,ORIG
text7-
poe_en,0.581,0.149,0.635,0.446,0.184,0.317,277.089,0.31,86.022,19,75,2418,3.947,32.24,1.425,53.556,0.843,7.373,0.662,0.066,ORIG
text8-
poe_en,0.311,0.204,0.301,0.201,0.142,0.107,256.188,0.156,40.017,15,104,2424,6.933,23.308,1.409,63.976,0.969,4.481,0.654,0.066,ORIG
text9-
poe_en,0.321,0.04,0.438,0.261,0.034,0.263,320.924,0.479,153.616,90,250,2337,2.778,9.348,1.401,78.822,0.635,1.856,0.649,0.155,ORIG

text10-
poe_en,0.4335,0.064,0.5635,0.379,0.0805,0.304,267.8525,0.279,74.7405,11,66.5,1749,6.5,27.266,1.502,52.0905,0.898,4.085,0.757,0
.0465,ORIG
text1-
litingl_en,0.58,0.12,0.633,0.493,0.089,0.354,299.363,0.465,139.278,38,151,2355,3.974,15.596,1.313,79.925,0.784,3.344,0.666,0.101,
ORIG
text2-litingl_en,0.1,0.1,0.05,0.129,0.123,0.044,284.314,0.149,42.484,7,21,306,3,14.571,1.467,67.937,0.92,4.143,0.834,0.02,ORIG
text3-
litingl_en,0.421,0.114,0.459,0.3595,0.097,0.324,291.3995,0.401,116.667,50.5,145,2145.5,2.8765,14.8755,1.294,82.2635,0.668,2.350
5,0.635,0.0725,ORIG
text4-
litingl_en,0.3175,0.083,0.453,0.275,0.1205,0.1885,261.353,0.386,101.003,5.5,51,1007,7.9375,22.0025,1.487,58.702,0.879,5.4905,0.8
13,0.0525,ORIG
001.lima.barreto.cazuza.port.TXTparte1.txt,0.217391,0.4,0.181034,0.146177,0.250225,0.318966,243.161,0.246201,12.1581,45,116,1
645,2.57778,14.181,2.67417,60.5612,0.4525,2.68966,0.696677,0.26087,ORIG
002.humberto.campos.promessa.port.txt,0.1922725,0.281818,0.36174,0.150967,0.249163,0.509459,284.526,0.184442,5.89292,22,78.
5,1253,3.55357,17.21145,2.79626,54.50415,0.515039,4.5122,0.752203,0.096818,ORIG
003.coelho.neto.firmo.port.TXTparte1.txt,0.100775,0.170543,0.330769,0.0911151,0.141205,0.369231,255.276,0.183345,13.6147,65,
130,1469,2,11.3,2.5345,70.066,0.554667,2.54615,0.702924,0.0930233,ORIG
004.lima.barreto.javanês.port.txt,0.188734,0.3001165,0.222231,0.1475665,0.261363,0.327906,241.711,0.277004,15.2226,44,115,153
9.5,2.706975,13.42255,2.672775,61.40485,0.515081,2.089775,0.715465,0.131707,ORIG
005.humberto.campos.caveiras.port.TXTparte1.txt,0.135135,0.234234,0.0803571,0.0883205,0.157658,0.116071,258.871,0.268817,1
6.129,39,112,1240,2.87179,11.0714,2.62158,66.7601,0.445483,2.89286,0.709352,0.135135,ORIG
006.coelho.neto.duplo.port.TXTparte1.txt,0.45679,0.555556,0.0853659,0.288768,0.370069,0.195122,238.594,0.379833,11.9671,34,8
2,1337,2.41176,16.3049,2.71773,63.655,0.53605,2.57317,0.774468,0.222222,ORIG
007.lima.barreto.número.sepultura.port.txt,0.2109245,0.3957985,0.5545545,0.137802,0.31453,0.69845,241.0095,0.2272185,14.5815
5,52,103,1540.5,1.96927,15.6873,2.718755,58.2148,0.542219,3.0812,0.67902,0.247059,ORIG
008.humberto.campos.vinganÇõa.port.txt,0.151786,0.3125,0.424779,0.159608,0.243837,0.486726,268.085,0.110907,4.25532,18,113
,2115,6.27778,18.7168,2.83962,52.1974,0.552028,4.99115,0.718696,0.160714,ORIG
001.machado.cartomante.port.txt,0.1540405,0.2590185,0.5305905,0.1497375,0.2637215,0.896063,241.6365,0.1765985,18.4421,30,1
13.5,1543.5,3.772525,13.5482,2.63982,60.8679,0.54723,2.17744,0.6718225,0.1349205,ORIG
002.machado.viver.port.txt,0.08299425,0.1438525,0.10221695,0.0664613,0.1282905,0.169167,255.231,0.3394305,34.08935,63.5,23
6.5,1685.5,3.671685,6.932825,2.51247,71.66345,0.598677,2.07288,0.589336,0.0718646,ORIG
003.machado.cantiga.port.TXTparte1.txt,0.273684,0.389474,0.40625,0.175,0.32193,0.625,234.838,0.243544,11.9887,29,96,1418,3.3
1034,14.7708,2.45863,67.0887,0.597598,2.05208,0.613475,0.273684,ORIG
004.machado.enfermeiro.port.txt,0.1858245,0.3945855,0.3502645,0.142883,0.331777,0.484025,248.366,0.2019005,13.0719,23,106.
5,1530,4.645835,14.38175,2.6293,61.2499,0.504197,2.18707,0.685674,0.2528785,ORIG
005.machado.marcha.fúnebre.port.txt,0.2167115,0.4183085,0.3943955,0.171942,0.3445855,0.5491575,242.2145,0.1793665,11.0584
5,33.5,126.5,1880.5,3.875545,15.06985,2.54492,65.83525,0.5410775,2.447435,0.6432325,0.213838,ORIG
006.machado.vida.eterna.port.txt,0.236583,0.3457725,0.339491,0.1749515,0.286537,0.4768925,230.8105,0.1907825,23.97465,102.5
,172,2388.5,1.678065,14.3395,2.627375,63.34945,0.581404,2.06455,0.6018745,0.1917555,ORIG
text1-
litbras_en,0.464,0.161,0.446,0.372,0.1,0.327,301.282,0.381,114.85,41,113,1872,2.756,16.566,1.426,69.381,0.617,3.77,0.658,0.098,T
RAD
text2-
litbras_en,0.3435,0.181,0.415,0.2655,0.152,0.201,286.535,0.327,93.6705,21,73,1374.5,3.5715,19.4645,1.382,70.1615,0.844,4.9335,0
.735,0.062,TRAD
text3-
litbras_en,0.2,0.069,0.325,0.183,0.084,0.189,302.57,0.369,111.565,69,161,1712,2.333,10.634,1.315,84.792,0.687,2.317,0.648,0.07,T
RAD
text4-
litbras_en,0.3195,0.049,0.433,0.2755,0.0605,0.2815,310.2145,0.416,129.2115,51.5,146,1673,2.8405,11.461,1.4635,71.39,0.6805,2.1
505,0.6945,0.0675,TRAD
text5-
litbras_en,0.267,0.164,0.302,0.213,0.115,0.168,287.729,0.36,103.667,44,117,1418,2.659,12.12,1.303,84.299,0.725,3.034,0.643,0.079
,TRAD
text6-
litbras_en,0.482,0.105,0.5,0.351,0.071,0.291,319.322,0.454,144.918,42,115,1594,2.738,13.861,1.326,80.586,0.611,2.426,0.668,0.116
,TRAD
text7-
litbras_en,0.4685,0.1835,0.5105,0.327,0.116,0.2815,295.9155,0.377,111.742,47.5,110,1702,2.357,16.4165,1.4375,68.5595,0.7415,3.
246,0.6595,0.1,TRAD
text8-
litbras_en,0.26,0.165,0.331,0.254,0.154,0.144,276.351,0.247,68.202,21,128,2258,6.095,17.641,1.415,69.22,0.965,5.055,0.658,0.055,
TRAD
text1-
machado_en,0.3555,0.096,0.461,0.307,0.1155,0.242,286.281,0.4095,117.4525,31,128.5,1949.5,4.2265,15.125,1.443,69.4055,0.649,3.
053,0.6355,0.0625,TRAD
text2-
machado_en,0.1685,0.125,0.101,0.1475,0.1085,0.0735,291.36,0.192,61.834,41.5,158.5,1340.5,3.8485,8.393,1.632,60.249,1.1735,1.6
84,0.586,0.0475,TRAD

text3-
machado_en,0.44,0.233,0.405,0.312,0.16,0.219,300.06,0.317,94.979,28,117,1653,4.179,14.128,1.346,78.623,0.754,3.171,0.576,0.122
,TRAD
text4-
machado_en,0.479,0.086,0.581,0.3995,0.0595,0.382,318.1105,0.4795,152.5515,30,150.5,1913,5.0165,12.7395,1.347,79.948,0.563,2.
3985,0.646,0.097,TRAD
text5-
machado_en,0.375,0.1245,0.425,0.285,0.084,0.2625,299.716,0.382,114.6305,24.5,112.5,1423.5,4.6365,12.6655,1.3635,78.627,0.649
5,2.5905,0.6995,0.081,TRAD
text6-
machado_en,0.374,0.091,0.424,0.305,0.0735,0.2795,306.9805,0.4725,145.0735,105.5,217.5,2733,2.074,12.75,1.356,79.176,0.67,2.47
9,0.5665,0.0845,TRAD
001.poe.oval.portrait.port.txtparte1.txt,0.32,0.56,0.627451,0.305098,0.505098,0.666667,229.35,0.333313,7.90861,9,51,1138,5.66667,
22.3137,2.91821,40.6316,0.655172,4.96078,0.756173,0.16,TRAD
002.poe.mesmeric.revelation.port.txt,0.252364,0.356721,0.215038,0.1110633,0.2297005,0.297078,242.3935,0.179732,7.93387,38.5,
134.5,1765,3.491215,13.15235,2.91515,46.0269,0.6340425,3.0717,0.581353,0.2922155,TRAD
003.poe.black.cat.port.txt,0.2458415,0.336944,0.405026,0.169322,0.2871545,0.571882,235.3185,0.2237775,12.6017,17,103,1704.5,
6.10764,16.6729,3.03418,38.2795,0.598444,3.50607,0.742789,0.1260156,TRAD
004.poe.imp.of.perversity.port.txtparte1.txt,0.364964,0.510949,0.181159,0.159314,0.299164,0.297101,220.889,0.186899,5.33333,19,
138,2250,7.26316,16.3043,2.97991,42.4061,0.591549,3.47826,0.673107,0.277372,TRAD
005.poe.tell-
tale.heart.port.txtparte1.txt,0.238636,0.392045,0.310734,0.116911,0.232987,0.446328,196.866,0.259476,32.3213,22,177,2042,8.045
45,11.5367,2.56961,67.4281,0.584577,2.44068,0.556027,0.215909,TRAD
006.poe.berenice.port.txt,0.1944445,0.378788,0.3064385,0.1587455,0.29549,0.5760275,231.851,0.222835,10.99825,19.5,86.5,1620.
5,4.421055,19.077,2.978345,40.75865,0.626173,4.966025,0.751642,0.2392675,TRAD
007.poe.eleonora.port.txt,0.5625,0.6375,0.518519,0.441975,0.574383,0.679012,220.423,0.20656,9.89654,26,81,2223,3.11538,27.444
4,2.84341,42.0364,0.577551,5.24691,0.680199,0.225,TRAD
008.poe.red.masque.port.txt,0.339623,0.5,0.196262,0.23964,0.383883,0.327103,233.808,0.145295,3.67478,22,107,2177,4.86364,20.
3458,2.91905,44.7998,0.587426,4.71963,0.684383,0.396226,TRAD
009.poe.cask.of.amontillado.port.txt,0.0557621,0.133829,0.148148,0.032521,0.0846482,0.240741,233.01,0.18305,17.4757,94,270,20
60,2.87234,7.62963,2.71548,62.9792,0.514583,1.85926,0.678661,0.0892193,TRAD
010.poe.man.of.crowd.port.txt,0.3041355,0.444314,0.426018,0.2665825,0.4268845,0.635296,214.314,0.1822725,20.22775,16,107.5,
2543.5,7.413045,23.24665,2.91838,39.1402,0.6969545,3.44251,0.691412,0.199013,TRAD
001.james.joyce.araby.port.txt,0.0927152,0.225166,0.282895,0.0942837,0.174974,0.427632,244.18,0.171796,14.3635,36,152,2019,4
.22222,13.2829,2.81865,53.8756,0.555781,2.76316,0.729706,0.10596,TRAD
002.virginia.woolf.monday.tuesday.port.txt,0.0588235,0.411765,0.166667,0.0784314,0.150327,0.222222,304.918,0.282038,3.27869,
7,18,305,2.57143,16.9444,2.87701,39.6915,0.494624,5.66667,0.86631,0.294118,TRAD
003.O.Henry.Red.Chief.port.txt,0.2485955,0.3635005,0.3370575,0.167234,0.245979,0.46674,239.614,0.18562,27.3452,50.5,154,199
9,3.054705,13.0114,2.589745,65.2826,0.5042335,2.11719,0.640197,0.2094015,TRAD
004.nathaniel.hawthorne.wakefield.port.txt,0.2540675,0.4184205,0.658071,0.2827545,0.3662415,1.0150045,234.03,0.2407515,11.80
988,8.5,83,1676,9.833335,20.34205,2.799155,51.5337,0.6495435,4.42089,0.7442545,0.146141,TRAD

Análise 3: Textos originais em português e textos traduzidos para o português

@relation 'textos originais em português e textos traduzidos para o português'

@attribute title string
@attribute argument_overlap_adjacent numeric
@attribute stem_overlap_adjacent numeric
@attribute anaphor_reference_adjacent numeric
@attribute argument_overlap numeric
@attribute stem_overlap numeric
@attribute anaphor_reference numeric
@attribute noun_phrase_incidence_score numeric
@attribute ratio_of_pronouns_to_noun_phrases numeric
@attribute personal_pronoun_incidence_score numeric
@attribute number_of_paragraphs numeric
@attribute number_of_sentences numeric
@attribute number_of_words numeric
@attribute average_sentences_per_paragraph numeric
@attribute average_words_per_sentence numeric
@attribute average_syllables_per_word numeric
@attribute flesch_reading_ease_score numeric
@attribute mean_number_of_modifiers_per_noun_phrase numeric
@attribute mean_number_of_words_before_the_main_verb numeric
@attribute type_token_ratio numeric
@attribute proportion_of_content_words_that_overlap_between_adjacent_sentences numeric
@attribute class {PORT_ORIG,PORT_TRAD}

@data
001.lima.barreto.cazuza.port.TXTparte1.txt,0.217391,0.4,0.181034,0.146177,0.250225,0.318966,243.161,0.246201,12.1581,45,116,1645,2.57778,14.181,2.67417,60.5612,0.4525,2.68966,0.696677,0.26087,PORT_ORIG
002.humberto.campos.promessa.port.txt,0.1922725,0.281818,0.36174,0.150967,0.249163,0.509459,284.526,0.184442,5.89292,22,78.5,1253,3.55357,17.21145,2.79626,54.50415,0.515039,4.5122,0.752203,0.096818,PORT_ORIG
003.coelho.neto.firmo.port.TXTparte1.txt,0.100775,0.170543,0.330769,0.0911151,0.141205,0.369231,255.276,0.183345,13.6147,65,130,1469,2,11.3,2.5345,70.066,0.554667,2.54615,0.702924,0.0930233,PORT_ORIG
004.lima.barreto.javanês.port.txt,0.188734,0.3001165,0.222231,0.1475665,0.261363,0.327906,241.711,0.277004,15.2226,44,115,1539.5,2.706975,13.42255,2.672775,61.40485,0.515081,2.089775,0.715465,0.131707,PORT_ORIG
005.humberto.campos.caveiras.port.TXTparte1.txt,0.135135,0.234234,0.0803571,0.0883205,0.157658,0.116071,258.871,0.268817,16.129,39,112,1240,2.87179,11.0714,2.62158,66.7601,0.445483,2.89286,0.709352,0.135135,PORT_ORIG
006.coelho.neto.duplo.port.TXTparte1.txt,0.45679,0.555556,0.0853659,0.288768,0.370069,0.195122,238.594,0.379833,11.9671,34,82,1337,2.41176,16.3049,2.71773,63.655,0.53605,2.57317,0.774468,0.222222,PORT_ORIG
007.lima.barreto.número.sepultura.port.txt,0.2109245,0.3957985,0.5545545,0.137802,0.31453,0.69845,241.0095,0.2272185,14.58155,52,103,1540.5,1.96927,15.6873,2.718755,58.2148,0.542219,3.0812,0.67902,0.247059,PORT_ORIG
008.humberto.campos.vinganÇõa.port.txt,0.151786,0.3125,0.424779,0.159608,0.243837,0.486726,268.085,0.110907,4.25532,18,113,2115,6.27778,18.7168,2.83962,52.1974,0.552028,4.99115,0.718696,0.160714,PORT_ORIG
001.machado.cartomante.port.txt,0.1540405,0.2590185,0.5305905,0.1497375,0.2637215,0.896063,241.6365,0.1765985,18.4421,30,113.5,1543.5,3.772525,13.5482,2.63982,60.8679,0.54723,2.17744,0.6718225,0.1349205,PORT_ORIG
002.machado.viver.port.txt,0.08299425,0.1438525,0.10221695,0.0664613,0.1282905,0.169167,255.231,0.3394305,34.08935,63.5,236.5,1685.5,3.671685,6.932825,2.51247,71.66345,0.598677,2.07288,0.589336,0.0718646,PORT_ORIG
003.machado.cantiga.port.TXTparte1.txt,0.273684,0.389474,0.40625,0.175,0.32193,0.625,234.838,0.243544,11.9887,29,96,1418,3.31034,14.7708,2.45863,67.0887,0.597598,2.05208,0.613475,0.273684,PORT_ORIG
004.machado.enfermeiro.port.txt,0.1858245,0.3945855,0.3502645,0.142883,0.331777,0.484025,248.366,0.2019005,13.0719,23,106.5,1530,4.645835,14.38175,2.6293,61.2499,0.504197,2.18707,0.685674,0.2528785,PORT_ORIG
005.machado.marcha.fúnebre.port.txt,0.2167115,0.4183085,0.3943955,0.171942,0.3445855,0.5491575,242.2145,0.1793665,11.05845,33.5,126.5,1880.5,3.875545,15.06985,2.54492,65.83525,0.5410775,2.447435,0.6432325,0.213838,PORT_ORIG
006.machado.vida.eterna.port.txt,0.236583,0.3457725,0.339491,0.1749515,0.286537,0.4768925,230.8105,0.1907825,23.97465,102.5,172,2388.5,1.678065,14.3395,2.627375,63.34945,0.581404,2.06455,0.6018745,0.1917555,PORT_ORIG
001.poe.oval.portrait.port.txtparte1.txt,0.32,0.56,0.627451,0.305098,0.505098,0.666667,229.35,0.333313,7.90861,9,51,1138,5.66667,22.3137,2.91821,40.6316,0.655172,4.96078,0.756173,0.16,PORT_TRAD
002.poe.mesmeric.revelation.port.txt,0.252364,0.356721,0.215038,0.1110633,0.2297005,0.297078,242.3935,0.179732,7.93387,38.5,134.5,1765,3.491215,13.15235,2.91515,46.0269,0.6340425,3.0717,0.581353,0.2922155,PORT_TRAD
003.poe.black.cat.port.txt,0.2458415,0.336944,0.405026,0.169322,0.2871545,0.571882,235.3185,0.2237775,12.6017,17,103,1704.5,6.10764,16.6729,3.03418,38.2795,0.598444,3.50607,0.742789,0.1260156,PORT_TRAD
004.poe.imp.of.perversity.port.txtparte1.txt,0.364964,0.510949,0.181159,0.159314,0.299164,0.297101,220.889,0.186899,5.33333,19,138,2250,7.26316,16.3043,2.97991,42.4061,0.591549,3.47826,0.673107,0.277372,PORT_TRAD
005.poe.tell-tale.heart.port.txtparte1.txt,0.238636,0.392045,0.310734,0.116911,0.232987,0.446328,196.866,0.259476,32.3213,22,177,2042,8.04545,11.5367,2.56961,67.4281,0.584577,2.44068,0.556027,0.215909,PORT_TRAD
006.poe.berenice.port.txt,0.1944445,0.378788,0.3064385,0.1587455,0.29549,0.5760275,231.851,0.222835,10.99825,19.5,86.5,1620.5,4.421055,19.077,2.978345,40.75865,0.626173,4.966025,0.751642,0.2392675,PORT_TRAD
007.poe.eleonora.port.txt,0.5625,0.6375,0.518519,0.441975,0.574383,0.679012,220.423,0.20656,9.89654,26,81,2223,3.11538,27.4444,2.84341,42.0364,0.577551,5.24691,0.680199,0.225,PORT_TRAD
008.poe.red.masque.port.txt,0.339623,0.5,0.196262,0.23964,0.383883,0.327103,233.808,0.145295,3.67478,22,107,2177,4.86364,20.3458,2.91905,44.7998,0.587426,4.71963,0.684383,0.396226,PORT_TRAD
009.poe.cask.of.amontillado.port.txt,0.0557621,0.133829,0.148148,0.032521,0.0846482,0.240741,233.01,0.18305,17.4757,94,270,2060,2.87234,7.62963,2.71548,62.9792,0.514583,1.85926,0.678661,0.0892193,PORT_TRAD
010.poe.man.of.crowd.port.txt,0.3041355,0.444314,0.426018,0.2665825,0.4268845,0.635296,214.314,0.1822725,20.22775,16,107.5,2543.5,7.413045,23.24665,2.91838,39.1402,0.6969545,3.44251,0.691412,0.199013,PORT_TRAD
001.james.joyce.araby.port.txt,0.0927152,0.225166,0.282895,0.0942837,0.174974,0.427632,244.18,0.171796,14.3635,36,152,2019,4.22222,13.2829,2.81865,53.8756,0.555781,2.76316,0.729706,0.10596,PORT_TRAD
002.virginia.woolf.monday.tuesday.port.txt,0.0588235,0.411765,0.166667,0.0784314,0.150327,0.222222,304.918,0.282038,3.27869,7,18,305,2.57143,16.9444,2.87701,39.6915,0.494624,5.66667,0.86631,0.294118,PORT_TRAD
003.O.Henry.Red.Chief.port.txt,0.2485955,0.3635005,0.3370575,0.167234,0.245979,0.46674,239.614,0.18562,27.3452,50.5,154,1999,3.054705,13.0114,2.589745,65.2826,0.5042335,2.11719,0.640197,0.2094015,PORT_TRAD
004.nathaniel.hawthorne.wakefield.port.txt,0.2540675,0.4184205,0.658071,0.2827545,0.3662415,1.0150045,234.03,0.2407515,11.80988,8.5,83,1676,9.833335,20.34205,2.799155,51.5337,0.6495435,4.42089,0.7442545,0.146141,PORT_TRAD

Análise 4: Textos originais em inglês e textos traduzidos para o inglês

@relation 'textos originais em inglês e textos traduzidos para o inglês '

@attribute title string
@attribute argument_overlap_adjacent numeric
@attribute stem_overlap_adjacent numeric
@attribute anaphor_reference_adjacent numeric
@attribute argument_overlap numeric
@attribute stem_overlap numeric
@attribute anaphor_reference numeric
@attribute noun_phrase_incidence_score numeric
@attribute ratio_of_pronouns_to_noun_phrases numeric

@attribute personal_pronoun_incidence_score numeric
@attribute number_of_paragraphs numeric
@attribute number_of_sentences numeric
@attribute number_of_words numeric
@attribute average_sentences_per_paragraph numeric
@attribute average_words_per_sentence numeric
@attribute average_syllables_per_word numeric
@attribute flesch_reading_ease_score numeric
@attribute mean_number_of_modifiers_per_noun_phrase numeric
@attribute mean_number_of_words_before_the_main_verb numeric
@attribute type_token_ratio numeric
@attribute proportion_of_content_words_that_overlap_between_adjacent_sentences numeric
@attribute class {EN_ORIG,EN_TRAD}

@data
text1-
poe_en,0.551,0.163,0.633,0.418,0.178,0.298,243.808,0.352,85.913,7,50,1292,7.143,25.84,1.413,61.068,0.867,6.84,0.711,0.07,EN_ORIG
text2-
poe_en,0.43,0.365,0.28,0.3325,0.245,0.1565,281.952,0.272,76.6135,38,108,1872.5,2.8435,17.3415,1.586,55.0575,0.8025,4.0995,0.563,0.1015,EN_ORIG
text3-
poe_en,0.5415,0.105,0.6355,0.5195,0.0945,0.332,278.5845,0.4015,111.879,13.5,91.5,1934.5,6.7855,21.1435,1.473,60.7585,0.7705,4.29,0.732,0.0825,EN_ORIG
text4-
poe_en,0.562,0.14,0.678,0.477,0.085,0.368,276.19,0.403,111.387,16,122,2415,7.625,19.795,1.557,55.021,0.727,4.066,0.707,0.108,EN_ORIG
text5-
poe_en,0.51,0.134,0.618,0.446,0.069,0.453,292.556,0.468,137.032,16,158,2109,9.875,13.348,1.324,81.276,0.645,2.532,0.574,0.136,EN_ORIG
text6-
poe_en,0.498,0.14,0.5355,0.413,0.1035,0.276,267.831,0.353,95.2565,12,66.5,1606,5.7785,24.4195,1.5515,50.7925,0.8755,3.764,0.7415,0.0975,EN_ORIG
text7-
poe_en,0.581,0.149,0.635,0.446,0.184,0.317,277.089,0.31,86.022,19,75,2418,3.947,32.24,1.425,53.556,0.843,7.373,0.662,0.066,EN_ORIG
text8-
poe_en,0.311,0.204,0.301,0.201,0.142,0.107,256.188,0.156,40.017,15,104,2424,6.933,23.308,1.409,63.976,0.969,4.481,0.654,0.066,EN_ORIG
text9-
poe_en,0.321,0.04,0.438,0.261,0.034,0.263,320.924,0.479,153.616,90,250,2337,2.778,9.348,1.401,78.822,0.635,1.856,0.649,0.155,EN_ORIG
text10-
poe_en,0.4335,0.064,0.5635,0.379,0.0805,0.304,267.8525,0.279,74.7405,11,66.5,1749,6.5,27.266,1.502,52.0905,0.898,4.085,0.757,0.0465,EN_ORIG
text1-
litingl_en,0.58,0.12,0.633,0.493,0.089,0.354,299.363,0.465,139.278,38,151,2355,3.974,15.596,1.313,79.925,0.784,3.344,0.666,0.101,EN_ORIG
text2-
litingl_en,0.1,0.1,0.05,0.129,0.123,0.044,284.314,0.149,42.484,7,21,306,3,14.571,1.467,67.937,0.92,4.143,0.834,0.02,EN_ORIG
text3-
litingl_en,0.421,0.114,0.459,0.3595,0.097,0.324,291.3995,0.401,116.667,50.5,145,2145.5,2.8765,14.8755,1.294,82.2635,0.668,2.3505,0.635,0.0725,EN_ORIG
text4-
litingl_en,0.3175,0.083,0.453,0.275,0.1205,0.1885,261.353,0.386,101.003,5.5,51,1007,7.9375,22.0025,1.487,58.702,0.879,5.4905,0.813,0.0525,EN_ORIG
text1-
litbras_en,0.464,0.161,0.446,0.372,0.1,0.327,301.282,0.381,114.85,41,113,1872,2.756,16.566,1.426,69.381,0.617,3.77,0.658,0.098,EN_TRAD
text2-
litbras_en,0.3435,0.181,0.415,0.2655,0.152,0.201,286.535,0.327,93.6705,21,73,1374.5,3.5715,19.4645,1.382,70.1615,0.844,4.9335,0.735,0.062,EN_TRAD
text3-
litbras_en,0.2,0.069,0.325,0.183,0.084,0.189,302.57,0.369,111.565,69,161,1712,2.333,10.634,1.315,84.792,0.687,2.317,0.648,0.07,EN_TRAD
text4-
litbras_en,0.3195,0.049,0.433,0.2755,0.0605,0.2815,310.2145,0.416,129.2115,51.5,146,1673,2.8405,11.461,1.4635,71.39,0.6805,2.1505,0.6945,0.0675,EN_TRAD
text5-
litbras_en,0.267,0.164,0.302,0.213,0.115,0.168,287.729,0.36,103.667,44,117,1418,2.659,12.12,1.303,84.299,0.725,3.034,0.643,0.079,EN_TRAD

text6-
litbras_en,0.482,0.105,0.5,0.351,0.071,0.291,319.322,0.454,144.918,42,115,1594,2.738,13.861,1.326,80.586,0.611,2.426,0.668,0.116
,EN_TRAD
text7-
litbras_en,0.4685,0.1835,0.5105,0.327,0.116,0.2815,295.9155,0.377,111.742,47.5,110,1702,2.357,16.4165,1.4375,68.5595,0.7415,3.
246,0.6595,0.1,EN_TRAD
text8-
litbras_en,0.26,0.165,0.331,0.254,0.154,0.144,276.351,0.247,68.202,21,128,2258,6.095,17.641,1.415,69.22,0.965,5.055,0.658,0.055,
EN_TRAD
text1-
machado_en,0.3555,0.096,0.461,0.307,0.1155,0.242,286.281,0.4095,117.4525,31,128.5,1949.5,4.2265,15.125,1.443,69.4055,0.649,3.
053,0.6355,0.0625,EN_TRAD
text2-
machado_en,0.1685,0.125,0.101,0.1475,0.1085,0.0735,291.36,0.192,61.834,41.5,158.5,1340.5,3.8485,8.393,1.632,60.249,1.1735,1.6
84,0.586,0.0475,EN_TRAD
text3-
machado_en,0.44,0.233,0.405,0.312,0.16,0.219,300.06,0.317,94.979,28,117,1653,4.179,14.128,1.346,78.623,0.754,3.171,0.576,0.122
,EN_TRAD
text4-
machado_en,0.479,0.086,0.581,0.3995,0.0595,0.382,318.1105,0.4795,152.5515,30,150.5,1913,5.0165,12.7395,1.347,79.948,0.563,2.
3985,0.646,0.097,EN_TRAD
text5-
machado_en,0.375,0.1245,0.425,0.285,0.084,0.2625,299.716,0.382,114.6305,24.5,112.5,1423.5,4.6365,12.6655,1.3635,78.627,0.649
5,2.5905,0.6995,0.081,EN_TRAD
text6-
machado_en,0.374,0.091,0.424,0.305,0.0735,0.2795,306.9805,0.4725,145.0735,105.5,217.5,2733,2.074,12.75,1.356,79.176,0.67,2.47
9,0.5665,0.0845,EN_TRAD

ANEXO D: RESULTADOS DAS FERRAMENTAS COH-METRIX E COH-METRIX-PORT COM DIFERENÇAS SIGNIFICATIVAS

PRONOMES POR SINTAGMA

PORTUGUÊS - INGLÊS					
titulo	pronomessintagmas	DENSPR2 'Ratio of pronouns to noun phrases'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	0,246201	0,381		0,25	0,381
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	0,168451	0,296		0,18	0,327
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	0,200433	0,358		0,18	0,369
003.coelho.neto.firmo.port.TXTparte1.txt	0,183345	0,369		0,28	0,416
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,257778	0,391		0,27	0,36
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	0,29623	0,441		0,38	0,454
005.humberto.campos.caveiras.port.TXTparte1.txt	0,268817	0,36		0,23	0,377
006.coelho.neto.duplo.port.TXTparte1.txt	0,379833	0,454		0,11	0,247
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	0,184861	0,354		0,18	0,4095
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	0,269576	0,4		0,34	0,192
008.humberto.campos.vinganÃ§a.port.txt	0,110907	0,247		0,24	0,317
001.machado.cartomante.port.TXTparte1.txt	0,180854	0,43		0,20	0,4795
001.machado.cartomante.port.TXTparte2.txt	0,172343	0,389		0,18	0,382
002.machado.viver.port.TXTparte1.txt	0,187809	0,37		0,19	0,4725
002.machado.viver.portpart2.txt	0,491052	0,014	**Média:**	0,229242143	0,37025
003.machado.cantiga.port.TXTparte1.txt	0,243544	0,317	**DVP:**	0,070582813	0,080914901
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	0,226966	0,461			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	0,176835	0,498			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	0,12009	0,359	**p-value (t-test)**	**0,00026433**	
005.machado.marcha.fÃºnebre.portparte2.txt	0,238643	0,405			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	0,178102	0,471			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	0,203463	0,474			
INGLÊS - PORTUGUÊS					
titulo	pronomessintagmas	DENSPR2 'Ratio of pronouns to noun phrases'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	0,333313	0,352		0,33	0,352
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	0,188049	0,331		0,18	0,272
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	0,171415	0,213		0,22	0,4015
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	0,225841	0,412		0,19	0,403
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	0,221714	0,391		0,26	0,468
004.poe.imp.of.perversity.port.txtparte1.txt	0,186899	0,403		0,22	0,353
005.poe.tell-tale.heart.port.txtparte1.txt	0,259476	0,468		0,21	0,31
006.poe.berenice.port.DUAS.PARTES.txt	0,216991	0,273		0,15	0,156
006.poe.berenice.port.DUAS.PARTES2.txt	0,228679	0,433		0,18	0,479
007.poe.eleonora.port.txt	0,20656	0,31		0,18	0,279
008.poe.red.masque.port.txt	0,145295	0,156		0,17	0,465
009.poe.cask.of.amontillado.port.txt	0,18305	0,479		0,28	0,149
010.poe.man.of.crowd.port.DUAS.PARTES.txt	0,102458	0,195		0,19	0,401
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	0,262087	0,363		0,24	0,386
001.james.joyce.araby.port.txt	0,171796	0,465	**Média:**	0,214529679	0,348178571
002.virginia.woolf.monday.tuesday.port.txt	0,282038	0,149	**DVP:**	0,050424223	0,105201932
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	0,163993	0,385			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	0,207247	0,417			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	0,228057	0,357	**p-value (t-test)**	**0,000874981**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	0,253446	0,415			

TYPES/TOKEN

PORTUGUÊS - INGLÊS					
titulo	tipotoken	TYPTOKc 'Type-token ratio for all content words'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	0,696677	0,658		0,70	0,658
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	0,782414	0,762		0,75	0,735
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	0,721992	0,708		0,70	0,648
003.coelho.neto.firmo.port.TXTparte1.txt	0,702924	0,648		0,72	0,6945
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,711882	0,681		0,71	0,643
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	0,719048	0,708		0,77	0,668
005.humberto.campos.caveiras.port.TXTparte1.txt	0,709352	0,643		0,68	0,6595
006.coelho.neto.duplo.port.TXTparte1.txt	0,774468	0,668		0,72	0,658
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	0,715921	0,707		0,67	0,6355
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	0,642119	0,612		0,59	0,586
008.humberto.campos.vinganÃ§a.port.txt	0,718696	0,658		0,61	0,576
001.machado.cartomante.port.TXTparte1.txt	0,635148	0,615		0,69	0,646
001.machado.cartomante.port.TXTparte2.txt	0,708497	0,656		0,64	0,6995
002.machado.viver.port.TXTparte1.txt	0,555556	0,628		0,60	0,5665
002.machado.viver.portpart2.txt	0,623116	0,544	**Média:**	0,68244425	0,648107143
003.machado.cantiga.port.TXTparte1.txt	0,613475	0,576	**DVP:**	0,054543673	0,047207471
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	0,681767	0,641			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	0,689581	0,651			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	0,604195	0,692	**p-value (t-test)**	**0,003731827**	
005.machado.marcha.fÃºnebre.portparte2.txt	0,68227	0,707			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	0,610953	0,574			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	0,592796	0,559			
INGLÊS - PORTUGUÊS					
titulo	tipotoken	TYPTOKc 'Type-token ratio for all content words'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	0,756173	0,711		0,76	0,711
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	0,642857	0,62		0,58	0,563
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	0,519849	0,506		0,74	0,732
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	0,720197	0,715		0,67	0,707
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	0,765381	0,749		0,56	0,574
004.poe.imp.of.perversity.port.txtparte1.txt	0,673107	0,707		0,75	0,7415
005.poe.tell-tale.heart.port.txtparte1.txt	0,556027	0,574		0,68	0,662
006.poe.berenice.port.DUAS.PARTES.txt	0,766667	0,738		0,68	0,654
006.poe.berenice.port.DUAS.PARTES2.txt	0,736617	0,745		0,68	0,649
007.poe.eleonora.port.txt	0,680199	0,662		0,69	0,757
008.poe.red.masque.port.txt	0,684383	0,654		0,73	0,666
009.poe.cask.of.amontillado.port.txt	0,678661	0,649		0,87	0,834
010.poe.man.of.crowd.port.DUAS.PARTES.txt	0,672243	0,817		0,64	0,635
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	0,710581	0,697		0,74	0,813
001.james.joyce.araby.port.txt	0,729706	0,666	**Média:**	0,698300964	0,69275
002.virginia.woolf.monday.tuesday.port.txt	0,86631	0,834	**DVP:**	0,077649486	0,079586129
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	0,636716	0,627			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	0,643678	0,643			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	0,734164	0,881	**p-value (t-test)**	**0,604723162**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	0,754345	0,745			

PRONOMES PESSOAIS

PORTUGUÊS - INGLÊS					
titulo	pronomespessoais	DENPRPi 'Personal pronoun incidence score'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	12,1581	114,85		12,16	114,85
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	3,31126	84,531		5,89	93,6705
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	8,47458	102,81		13,61	111,565
003.coelho.neto.firmo.port.TXTparte1.txt	13,6147	111,565		15,22	129,2115
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	16,1692	118,224		16,13	103,667
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	14,276	140,199		11,97	144,918
005.humberto.campos.caveiras.port.TXTparte1.txt	16,129	103,667		14,58	111,742
006.coelho.neto.duplo.port.TXTparte1.txt	11,9671	144,918		4,26	68,202
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	12,1387	101,075		18,44	117,4525
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	17,0244	122,409		34,09	61,834
008.humberto.campos.vinganÃ§a.port.txt	4,25532	68,202		11,99	94,979
001.machado.cartomante.port.TXTparte1.txt	18,5915	127,149		13,07	152,5515
001.machado.cartomante.port.TXTparte2.txt	18,2927	107,756		11,06	114,6305
002.machado.viver.port.TXTparte1.txt	28,4638	120,025		23,97	145,0735
002.machado.viver.portpart2.txt	39,7149	3,643	**Média:**	14,74617429	111,7390714
003.machado.cantiga.port.TXTparte1.txt	11,9887	94,979	**DVP:**	7,35524558	26,73545856
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	14,3791	147,262			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	11,7647	157,841			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	10,6132	107,072	**p-value (t-test)**	**8,82787E-09**	
005.machado.marcha.fÃºnebre.portparte2.txt	11,5037	122,189			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	22,1256	138,324			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	25,8237	151,823			
INGLÊS - PORTUGUÊS					
titulo	pronomespessoais	DENPRPi 'Personal pronoun incidence score'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	7,90861	85,913		7,91	85,913
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	8,05524	92,733		7,93	76,6135
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	7,8125	60,494		12,60	111,879
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	13,2488	117,043		5,33	111,387
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	11,9546	106,715		32,32	137,032
004.poe.imp.of.perversity.port.txtparte1.txt	5,33333	111,387		11,00	95,2565
005.poe.tell-tale.heart.port.txtparte1.txt	32,3213	137,032		9,90	86,022
006.poe.berenice.port.DUAS.PARTES.txt	8,3709	71,09		3,67	40,017
006.poe.berenice.port.DUAS.PARTES2.txt	13,6256	119,423		17,48	153,616
007.poe.eleonora.port.txt	9,89654	86,022		20,23	74,7405
008.poe.red.masque.port.txt	3,67478	40,017		14,36	139,278
009.poe.cask.of.amontillado.port.txt	17,4757	153,616		3,28	42,484
010.poe.man.of.crowd.port.DUAS.PARTES.txt	15,2695	52,184		27,35	116,667
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	25,186	97,297		11,81	101,003
001.james.joyce.araby.port.txt	14,3635	139,278	**Média:**	13,22636429	97,99346429
002.virginia.woolf.monday.tuesday.port.txt	3,27869	42,484	**DVP:**	8,59138327	33,72896334
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	33,5512	113,851			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	21,1392	119,483			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	8,98876	92,593	**p-value (t-test)**	**6,78035E-08**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	14,631	109,413			

OPERADORES LÓGICOS

PORTUGUÊS - INGLÊS					
titulo	numerooplog	DENLOGi 'Logical operator incidence score (and + if + or + cond + neg)'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	43,769	44,338		43,77	44,338
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	37,2517	49,028		33,26	43,352
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	29,2758	37,676		39,48	42,056
003.coelho.neto.firmo.port.TXTparte1.txt	39,4826	42,056		43,69	38,8165
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	39,801	39,792		50,81	43,724
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	47,5867	37,841		41,14	40,151
005.humberto.campos.caveiras.port.TXTparte1.txt	50,8064	43,724		48,92	50,088
006.coelho.neto.duplo.port.TXTparte1.txt	41,1369	40,151		43,97	41,63
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	54,9133	54,839		53,80	44,2875
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	42,9312	45,337		38,00	26,003
008.humberto.campos.vinganÃ§a.port.txt	43,9716	41,63		55,71	37,508
001.machado.cartomante.port.TXTparte1.txt	55,7746	42,986		46,08	42,6985
001.machado.cartomante.port.TXTparte2.txt	51,8293	45,589		51,06	52,497
002.machado.viver.port.TXTparte1.txt	39,347	49,274		41,72	40,6765
002.machado.viver.portpart2.txt	36,6599	2,732	**Média:**	45,10120357	41,98757143
003.machado.cantiga.port.TXTparte1.txt	55,7123	37,508	**DVP:**	6,361957412	6,096119241
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	50,9804	48,379			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	41,1765	37,018			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	49,5283	55,519	**p-value (t-test)**	**0,11949541**	
005.machado.marcha.fÃºnebre.portparte2.txt	52,5883	49,475			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	44,2513	45,276			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	39,1808	36,077			
INGLÊS - PORTUGUÊS					
titulo	numerooplog	DENLOGi 'Logical operator incidence score (and + if + or + cond + neg)'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	48,3304	61,92		48,33	61,92
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	37,9747	47,18		33,78	37,793
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	29,5759	28,406		41,94	50,922
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	42,6267	50,308		43,11	49,275
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	41,2433	51,536		43,58	52,632
004.poe.imp.of.perversity.port.txtparte1.txt	43,1111	49,275		51,42	67,191
005.poe.tell-tale.heart.port.txtparte1.txt	43,5847	52,632		53,53	57,072
006.poe.berenice.port.DUAS.PARTES.txt	48,9375	62,204		61,09	66,007
006.poe.berenice.port.DUAS.PARTES2.txt	53,91	72,178		37,38	43,646
007.poe.eleonora.port.txt	53,5313	57,072		50,35	52,27
008.poe.red.masque.port.txt	61,0932	66,007		44,08	46,709
009.poe.cask.of.amontillado.port.txt	37,3786	43,646		65,57	71,895
010.poe.man.of.crowd.port.DUAS.PARTES.txt	50,8982	56,432		48,28	51,87
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	49,7997	48,108		42,32	48,255
001.james.joyce.araby.port.txt	44,0812	46,709	**Média:**	47,48345714	54,10407143
002.virginia.woolf.monday.tuesday.port.txt	65,5738	71,895	**DVP:**	8,626717992	9,638266326
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	44,8802	52,61			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	51,6735	51,13			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	47,7528	55,556	**p-value (t-test)**	**3,49294E-05**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	36,8957	40,954			

NEGAÇÕES

PORTUGUÊS - INGLÊS

titulo	numeroneg	DENNEGi 'Number of negations, incidence score'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	5,47112	8,013		5,47	8,013
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	2,48344	5,072		2,01	3,813
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	1,54083	2,554		2,04	4,089
003.coelho.neto.firmo.port.TXTparte1.txt	2,04221	4,089		3,71	8,1125
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,621891	6,92		3,23	7,052
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	6,7981	9,305		2,24	6,274
005.humberto.campos.caveiras.port.TXTparte1.txt	3,22581	7,052		5,08	8,9935
006.coelho.neto.duplo.port.TXTparte1.txt	2,24383	6,274		1,89	5,314
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	8,67052	10,215		6,99	8,0945
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	1,48038	7,772		8,47	11,1575
008.humberto.campos.vinganÃ§a.port.txt	1,89125	5,314		14,81	13,309
001.machado.cartomante.port.TXTparte1.txt	7,88732	10,86		6,86	6,815
001.machado.cartomante.port.TXTparte2.txt	6,09756	5,329		7,65	11,635
002.machado.viver.port.TXTparte1.txt	8,79029	19,583		3,56	6,186
002.machado.viver.portpart2.txt	8,14664	2,732	**Média:**	5,286438964	7,775571429
003.machado.cantiga.port.TXTparte1.txt	14,8096	13,309	**DVP:**	3,547531839	2,777467613
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	11,7647	7,974			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	1,96078	5,656			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	7,07547	16,523	**p-value (t-test)**	**0,000111637**	
005.machado.marcha.fÃºnebre.portparte2.txt	8,21693	6,747			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	3,55591	7,487			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	3,56189	4,885			

INGLÊS - PORTUGUÊS

titulo	numeroneg	DENNEGi 'Number of negations, incidence score'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	1,75747	10,836		1,76	10,836
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	4,60299	18,438		5,37	14,7425
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	6,13839	11,047		3,81	13,6865
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	4,03226	15,4		4,44	14,907
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	3,58637	11,973		5,39	17,07
004.poe.imp.of.perversity.port.txtparte1.txt	4,44444	14,907		3,66	15,48
005.poe.tell-tale.heart.port.txtparte1.txt	5,38688	17,07		4,50	9,098
006.poe.berenice.port.DUAS.PARTES.txt	2,57566	17,18		2,76	6,188
006.poe.berenice.port.DUAS.PARTES2.txt	4,73934	13,78		3,40	16,26
007.poe.eleonora.port.txt	4,49843	9,098		3,94	7,3585
008.poe.red.masque.port.txt	2,75609	6,188		1,98	11,04
009.poe.cask.of.amontillado.port.txt	3,39806	16,26		0,00	6,536
010.poe.man.of.crowd.port.DUAS.PARTES.txt	3,29341	6,068		4,81	8,5725
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	4,57928	8,649		5,24	6,924
001.james.joyce.araby.port.txt	1,98118	11,04	**Média:**	3,646437143	11,33564286
002.virginia.woolf.monday.tuesday.port.txt	0	6,536	**DVP:**	1,565775722	3,941785482
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	6,10022	10,686			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	3,52319	6,459			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	7,30337	5,291	**p-value (t-test)**	**3,01744E-06**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	3,18066	8,557			

SOBREPOSIÇÃO DE PALAVRAS DE CONTEÚDO

PORTUGUÊS - INGLÊS					
titulo	cwovl	CREFC1u 'Proportion of content words that overlap between adjacent sentences'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	0,26087	0,098		0,26	0,098
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	0,163636	0,043		0,10	0,062
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	0,03	0,081		0,09	0,07
003.coelho.neto.firmo.port.TXTparte1.txt	0,093023	0,07		0,13	0,0675
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,130081	0,05		0,14	0,079
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	0,133333	0,085		0,22	0,116
005.humberto.campos.caveiras.port.TXTparte1.txt	0,135135	0,079		0,25	0,1
006.coelho.neto.duplo.port.TXTparte1.txt	0,222222	0,116		0,16	0,055
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	0,317647	0,082		0,13	0,0625
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	0,176471	0,118		0,07	0,0475
008.humberto.campos.vinganÃ§a.port.txt	0,160714	0,055		0,27	0,122
001.machado.cartomante.port.TXTparte1.txt	0,15873	0,072		0,25	0,097
001.machado.cartomante.port.TXTparte2.txt	0,111111	0,053		0,21	0,081
002.machado.viver.port.TXTparte1.txt	0,084906	0,051		0,19	0,0845
002.machado.viver.portpart2.txt	0,058824	0,044	**Média:**	0,177606386	0,081571429
003.machado.cantiga.port.TXTparte1.txt	0,273684	0,122	**DVP:**	0,068521289	0,022580503
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	0,303922	0,095			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	0,201835	0,099			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	0,206897	0,085	**p-value (t-test)**	**7,44221E-06**	
005.machado.marcha.fÃºnebre.portparte2.txt	0,220779	0,077			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	0,241379	0,079			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	0,142132	0,09			
INGLÊS - PORTUGUÊS					
titulo	cwovl	CREFC1u 'Proportion of content words that overlap between adjacent sentences'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	0,16	0,07		0,16	0,07
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	0,293651	0,094		0,29	0,10
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	0,29078	0,109		0,13	0,08
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	0,170213	0,087		0,28	0,11
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	0,081818	0,078		0,22	0,14
004.poe.imp.of.perversity.port.txtparte1.txt	0,277372	0,108		0,24	0,10
005.poe.tell-tale.heart.port.txtparte1.txt	0,215909	0,136		0,23	0,07
006.poe.berenice.port.DUAS.PARTES.txt	0,347222	0,07		0,40	0,07
006.poe.berenice.port.DUAS.PARTES2.txt	0,131313	0,125		0,09	0,16
007.poe.eleonora.port.txt	0,225	0,066		0,20	0,05
008.poe.red.masque.port.txt	0,396226	0,066		0,11	0,10
009.poe.cask.of.amontillado.port.txt	0,089219	0,155		0,29	0,02
010.poe.man.of.crowd.port.DUAS.PARTES.txt	0,210526	0,035		0,21	0,07
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	0,1875	0,058		0,15	0,05
001.james.joyce.araby.port.txt	0,10596	0,101	**Média:**	0,212561314	0,083928571
002.virginia.woolf.monday.tuesday.port.txt	0,294118	0,02	**DVP:**	0,084691166	0,035703295
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	0,207865	0,062			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	0,210938	0,083			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	0,142857	0,059	**p-value (t-test)**	**0,000391832**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	0,149425	0,046			

SOBREPOSIÇÃO DE RADICAL DE PALAVRAS

PORTUGUÊS - INGLÊS					
titulo	stmovl	CREFSau 'Stem Overlap, all distances, unweighted'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	0,250225	0,1		0,25	0,1
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	0,345455	0,196		0,25	0,152
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	0,152871	0,108		0,14	0,084
003.coelho.neto.firmo.port.TXTparte1.txt	0,141205	0,084		0,26	0,0605
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,254262	0,063		0,16	0,115
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	0,268464	0,058		0,37	0,071
005.humberto.campos.caveiras.port.TXTparte1.txt	0,157658	0,115		0,31	0,116
006.coelho.neto.duplo.port.TXTparte1.txt	0,370069	0,071		0,24	0,154
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	0,414774	0,136		0,26	0,1155
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	0,214286	0,096		0,13	0,1085
008.humberto.campos.vinganÃ§a.port.txt	0,243837	0,154		0,32	0,16
001.machado.cartomante.port.TXTparte1.txt	0,306837	0,141		0,33	0,0595
001.machado.cartomante.port.TXTparte2.txt	0,220606	0,09		0,34	0,084
002.machado.viver.port.TXTparte1.txt	0,137576	0,043		0,29	0,0735
002.machado.viver.portpart2.txt	0,119005	0,174	**Média:**	0,261777964	0,103821429
003.machado.cantiga.port.TXTparte1.txt	0,32193	0,16	**DVP:**	0,075607337	0,034041608
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	0,338283	0,069			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	0,325271	0,05			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	0,338522	0,079	**p-value (t-test)**	**1,46461E-05**	
005.machado.marcha.fÃºnebre.portparte2.txt	0,350649	0,089			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	0,361927	0,087			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	0,211147	0,06			
INGLÊS - PORTUGUÊS					
titulo	stmovl	CREFSau 'Stem Overlap, all distances, unweighted'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	0,505098	0,178		0,51	0,18
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	0,212473	0,201		0,23	0,25
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	0,246928	0,289		0,29	0,09
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	0,359239	0,094		0,30	0,09
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	0,21507	0,095		0,23	0,07
004.poe.imp.of.perversity.port.txtparte1.txt	0,299164	0,085		0,30	0,10
005.poe.tell-tale.heart.port.txtparte1.txt	0,232987	0,069		0,57	0,18
006.poe.berenice.port.DUAS.PARTES.txt	0,332192	0,132		0,38	0,14
006.poe.berenice.port.DUAS.PARTES2.txt	0,258788	0,075		0,08	0,03
007.poe.eleonora.port.txt	0,574383	0,184		0,43	0,08
008.poe.red.masque.port.txt	0,383883	0,142		0,17	0,09
009.poe.cask.of.amontillado.port.txt	0,084648	0,034		0,15	0,12
010.poe.man.of.crowd.port.DUAS.PARTES.txt	0,418584	0,088		0,25	0,10
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	0,435185	0,073		0,37	0,12
001.james.joyce.araby.port.txt	0,174974	0,089	**Média:**	0,3040653	0,1175
002.virginia.woolf.monday.tuesday.port.txt	0,150327	0,123	**DVP:**	0,136322807	0,054550964
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	0,251648	0,075			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	0,24031	0,119			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	0,430236	0,126	**p-value (t-test)**	**6,58384E-05**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	0,302247	0,115			

SOBREPOSIÇÃO DE RADICAL DE PALAVRAS (ADJACENTES)

PORTUGUÊS - INGLÊS					
titulo	stmadjovl	CREFS1u 'Stem Overlap, adjacent, unweighted'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	0,4	0,161		0,40	0,161
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	0,363636	0,208		0,28	0,181
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	0,2	0,154		0,17	0,069
003.coelho.neto.firmo.port.TXTparte1.txt	0,170543	0,069		0,30	0,049
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,276423	0,033		0,23	0,164
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	0,32381	0,065		0,56	0,105
005.humberto.campos.caveiras.port.TXTparte1.txt	0,234234	0,164		0,40	0,1835
006.coelho.neto.duplo.port.TXTparte1.txt	0,555556	0,105		0,31	0,165
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	0,447059	0,184		0,26	0,096
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	0,344538	0,183		0,14	0,125
008.humberto.campos.vinganÃ§a.port.txt	0,3125	0,165		0,39	0,233
001.machado.cartomante.port.TXTparte1.txt	0,285714	0,113		0,39	0,086
001.machado.cartomante.port.TXTparte2.txt	0,232323	0,079		0,42	0,1245
002.machado.viver.port.TXTparte1.txt	0,163522	0,06		0,35	0,091
002.machado.viver.portpart2.txt	0,124183	0,19	**Média:**	0,328684107	0,130928571
003.machado.cantiga.port.TXTparte1.txt	0,389474	0,233	**DVP:**	0,108714171	0,051870718
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	0,431373	0,128			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	0,357798	0,044			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	0,408046	0,142	**p-value (t-test)**	**1,84714E-05**	
005.machado.marcha.fÃºnebre.portparte2.txt	0,428571	0,107			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	0,427586	0,119			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	0,263959	0,063			
INGLÊS - PORTUGUÊS					
titulo	stmadjovl	CREFS1u 'Stem Overlap, adjacent, unweighted'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	0,56	0,163		0,56	0,163
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	0,373016	0,296		0,36	0,365
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	0,340426	0,434		0,34	0,105
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	0,382979	0,133		0,51	0,14
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	0,290909	0,077		0,39	0,134
004.poe.imp.of.perversity.port.txtparte1.txt	0,510949	0,14		0,38	0,14
005.poe.tell-tale.heart.port.txtparte1.txt	0,392045	0,134		0,64	0,149
006.poe.berenice.port.DUAS.PARTES.txt	0,555556	0,167		0,50	0,204
006.poe.berenice.port.DUAS.PARTES2.txt	0,20202	0,113		0,13	0,04
007.poe.eleonora.port.txt	0,6375	0,149		0,44	0,064
008.poe.red.masque.port.txt	0,5	0,204		0,23	0,12
009.poe.cask.of.amontillado.port.txt	0,133829	0,04		0,41	0,1
010.poe.man.of.crowd.port.DUAS.PARTES.txt	0,451128	0,078		0,36	0,114
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	0,4375	0,05		0,42	0,083
001.james.joyce.araby.port.txt	0,225166	0,12	**Média:**	0,404995857	0,137214286
002.virginia.woolf.monday.tuesday.port.txt	0,411765	0,1	**DVP:**	0,128756322	0,077551153
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	0,398876	0,095			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	0,328125	0,133			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	0,480519	0,143	**p-value (t-test)**	**4,4462E-06**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	0,356322	0,023			

SOBREPOSIÇÃO DE ARGUMENTOS

PORTUGUÊS - INGLÊS					
titulo	argovl	CREFAau 'Argument Overlap, all distances, unweighted'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	0,146177	0,372		0,15	0,372
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	0,194805	0,309		0,15	0,2655
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	0,107129	0,222		0,09	0,183
003.coelho.neto.firmo.port.TXTparte1.txt	0,091115	0,183		0,15	0,2755
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,142932	0,253		0,09	0,213
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	0,152201	0,298		0,29	0,351
005.humberto.campos.caveiras.port.TXTparte1.txt	0,088321	0,213		0,14	0,327
006.coelho.neto.duplo.port.TXTparte1.txt	0,288768	0,351		0,16	0,254
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	0,1658	0,395		0,15	0,307
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	0,109804	0,259		0,07	0,1475
008.humberto.campos.vinganÃ§a.port.txt	0,159608	0,254		0,18	0,312
001.machado.cartomante.port.TXTparte1.txt	0,167354	0,325		0,14	0,3995
001.machado.cartomante.port.TXTparte2.txt	0,132121	0,289		0,17	0,285
002.machado.viver.port.TXTparte1.txt	0,071213	0,15		0,17	0,305
002.machado.viver.portpart2.txt	0,06171	0,145	**Média:**	0,149378529	0,2855
003.machado.cantiga.port.TXTparte1.txt	0,175	0,312	**DVP:**	0,052289125	0,070347763
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	0,166	0,358			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	0,119766	0,441			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	0,169392	0,258	**p-value (t-test)**	**4,31117E-07**	
005.machado.marcha.fÃºnebre.portparte2.txt	0,174492	0,312			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	0,23146	0,338			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	0,118443	0,272			
INGLÊS - PORTUGUÊS					
titulo	argovl	CREFAau 'Argument Overlap, all distances, unweighted'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	0,305098	0,418		0,31	0,42
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	0,098863	0,352		0,11	0,33
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	0,123264	0,313		0,17	0,52
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	0,209406	0,552		0,16	0,48
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	0,129238	0,487		0,12	0,45
004.poe.imp.of.perversity.port.txtparte1.txt	0,159314	0,477		0,16	0,41
005.poe.tell-tale.heart.port.txtparte1.txt	0,116911	0,446		0,44	0,45
006.poe.berenice.port.DUAS.PARTES.txt	0,164764	0,319		0,24	0,20
006.poe.berenice.port.DUAS.PARTES2.txt	0,152727	0,507		0,03	0,26
007.poe.eleonora.port.txt	0,441975	0,446		0,27	0,38
008.poe.red.masque.port.txt	0,23964	0,201		0,09	0,49
009.poe.cask.of.amontillado.port.txt	0,032521	0,261		0,08	0,13
010.poe.man.of.crowd.port.DUAS.PARTES.txt	0,251375	0,346		0,17	0,36
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	0,28179	0,412		0,28	0,28
001.james.joyce.araby.port.txt	0,094284	0,493	**Média:**	0,187419707	0,367821429
002.virginia.woolf.monday.tuesday.port.txt	0,078431	0,129	**DVP:**	0,10898987	0,115596215
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	0,173373	0,323			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	0,161095	0,396			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	0,310023	0,221	**p-value (t-test)**	**0,00041141**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	0,255486	0,329			

SOBREPOSIÇÃO DE ARGUMENTOS (ADJACENTES)

PORTUGUÊS - INGLÊS					
titulo	argadjovl	CREFA1u 'Argument Overlap, adjacent, unweighted'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	0,217391	0,464		0,22	0,464
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	0,254545	0,302		0,19	0,3435
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	0,13	0,385		0,10	0,2
003.coelho.neto.firmo.port.TXTparte1.txt	0,100775	0,2		0,19	0,3195
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,186992	0,272		0,14	0,267
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	0,190476	0,367		0,46	0,482
005.humberto.campos.caveiras.port.TXTparte1.txt	0,135135	0,267		0,21	0,4685
006.coelho.neto.duplo.port.TXTparte1.txt	0,45679	0,482		0,15	0,26
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	0,211765	0,517		0,15	0,3555
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	0,210084	0,42		0,08	0,1685
008.humberto.campos.vinganÃ§a.port.txt	0,151786	0,26		0,27	0,44
001.machado.cartomante.port.TXTparte1.txt	0,166667	0,404		0,19	0,479
001.machado.cartomante.port.TXTparte2.txt	0,141414	0,307		0,22	0,375
002.machado.viver.port.TXTparte1.txt	0,100629	0,167		0,24	0,374
002.machado.viver.portpart2.txt	0,06536	0,17	**Média:**	0,200260411	0,356892857
003.machado.cantiga.port.TXTparte1.txt	0,273684	0,44	**DVP:**	0,090288738	0,104208773
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	0,215686	0,496			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	0,155963	0,462			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	0,212644	0,4	**p-value (t-test)**	**2,05278E-06**	
005.machado.marcha.fÃºnebre.portparte2.txt	0,220779	0,35			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	0,331034	0,394			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	0,142132	0,354			
INGLÊS - PORTUGUÊS					
titulo	argadjovl	CREFA1u 'Argument Overlap, adjacent, unweighted'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	0,32	0,551		0,32	0,551
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	0,277778	0,426		0,25	0,43
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	0,22695	0,434		0,25	0,5415
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	0,255319	0,578		0,36	0,562
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	0,236364	0,505		0,24	0,51
004.poe.imp.of.perversity.port.txtparte1.txt	0,364964	0,562		0,19	0,498
005.poe.tell-tale.heart.port.txtparte1.txt	0,238636	0,51		0,56	0,581
006.poe.berenice.port.DUAS.PARTES.txt	0,277778	0,433		0,34	0,311
006.poe.berenice.port.DUAS.PARTES2.txt	0,111111	0,563		0,06	0,321
007.poe.eleonora.port.txt	0,5625	0,581		0,30	0,4335
008.poe.red.masque.port.txt	0,339623	0,311		0,09	0,58
009.poe.cask.of.amontillado.port.txt	0,055762	0,321		0,06	0,1
010.poe.man.of.crowd.port.DUAS.PARTES.txt	0,308271	0,392		0,25	0,421
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	0,3	0,475		0,25	0,3175
001.james.joyce.araby.port.txt	0,092715	0,58	**Média:**	0,25231945	0,439821429
002.virginia.woolf.monday.tuesday.port.txt	0,058824	0,1	**DVP:**	0,132703506	0,137886046
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	0,247191	0,375			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	0,25	0,467			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	0,220779	0,286	**p-value (t-test)**	**0,000203268**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	0,287356	0,349			

MODIFICADORES POR SINTAGMA

PORTUGUÊS - INGLÊS

titulo	modsintnominais	SYNNP 'Mean number of modifiers per noun-phrase'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	0,4525	0,617		0,45	0,617
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	0,563953	0,861		0,52	0,844
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	0,466125	0,827		0,55	0,687
003.coelho.neto.firmo.port.TXTparte1.txt	0,554667	0,687		0,52	0,6805
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,471503	0,685		0,45	0,725
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	0,558659	0,676		0,54	0,611
005.humberto.campos.caveiras.port.TXTparte1.txt	0,445483	0,725		0,54	0,7415
006.coelho.neto.duplo.port.TXTparte1.txt	0,53605	0,611		0,55	0,965
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	0,575179	0,817		0,55	0,649
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	0,509259	0,666		0,60	1,1735
008.humberto.campos.vinganÃ§a.port.txt	0,552028	0,965		0,60	0,754
001.machado.cartomante.port.TXTparte1.txt	0,473441	0,599		0,50	0,563
001.machado.cartomante.port.TXTparte2.txt	0,621019	0,699		0,54	0,6495
002.machado.viver.port.TXTparte1.txt	0,572848	0,544		0,58	0,67
002.machado.viver.portpart2.txt	0,624506	1,803	**Média:**	0,534517893	0,737857143
003.machado.cantiga.port.TXTparte1.txt	0,597598	0,754	**DVP:**	0,04607969	0,16185029
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	0,472585	0,542			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	0,535809	0,584			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	0,540901	0,632	**p-value (t-test)**	**0,000186311**	
005.machado.marcha.fÃºnebre.portparte2.txt	0,541254	0,667			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	0,604491	0,695			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	0,558317	0,645			

INGLÊS - PORTUGUÊS

titulo	modsintnominais	SYNNP 'Mean number of modifiers per noun-phrase'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	0,655172	0,867		0,66	0,867
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	0,626829	0,735		0,63	0,8025
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	0,641256	0,87		0,60	0,7705
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	0,610422	0,747		0,59	0,727
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	0,586466	0,794		0,58	0,645
004.poe.imp.of.perversity.port.txtparte1.txt	0,591549	0,727		0,63	0,8755
005.poe.tell-tale.heart.port.txtparte1.txt	0,584577	0,645		0,58	0,843
006.poe.berenice.port.DUAS.PARTES.txt	0,638356	0,911		0,59	0,969
006.poe.berenice.port.DUAS.PARTES2.txt	0,61399	0,84		0,51	0,635
007.poe.eleonora.port.txt	0,577551	0,843		0,70	0,898
008.poe.red.masque.port.txt	0,587426	0,969		0,56	0,784
009.poe.cask.of.amontillado.port.txt	0,514583	0,635		0,49	0,92
010.poe.man.of.crowd.port.DUAS.PARTES.txt	0,680426	0,957		0,50	0,668
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	0,713483	0,839		0,65	0,879
001.james.joyce.araby.port.txt	0,555781	0,784	**Média:**	0,590761	0,805964286
002.virginia.woolf.monday.tuesday.port.txt	0,494624	0,92	**DVP:**	0,059516031	0,105787062
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	0,454545	0,664			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	0,553922	0,672			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	0,665842	0,867	**p-value (t-test)**	**1,47533E-06**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	0,633245	0,891			

INCIDÊNCIA DE SINTAGMAS NOMINAIS

PORTUGUÊS - INGLÊS					
titulo	incidsintnominais	DENSNP 'Noun Phrase Incidence Score (per thousand words)'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	243,161	301,282		243,2	301,282
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	284,768	285,714		284,5	286,535
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	284,284	287,356		255,3	302,57
003.coelho.neto.firmo.port.TXTparte1.txt	255,276	302,57		241,7	310,2145
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	240,05	302,191		258,9	287,729
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	243,372	318,238		238,6	319,322
005.humberto.campos.caveiras.port.TXTparte1.txt	258,871	287,729		241,0	295,9155
006.coelho.neto.duplo.port.TXTparte1.txt	238,594	319,322		268,1	276,351
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	242,197	285,484		241,6	286,281
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	239,822	306,347		255,2	291,36
008.humberto.campos.vinganÃ§a.port.txt	268,085	276,351		234,8	300,06
001.machado.cartomante.port.TXTparte1.txt	243,944	295,475		248,4	318,1105
001.machado.cartomante.port.TXTparte2.txt	239,329	277,087		242,2	299,716
002.machado.viver.port.TXTparte1.txt	252,825	324,068		230,8	306,9805
002.machado.viver.portpart2.txt	257,637	258,652	**Média:**	248,8807143	298,7447857
003.machado.cantiga.port.TXTparte1.txt	234,838	300,06	**DVP:**	14,39344801	12,43868238
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	250,327	319,511			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	246,405	316,71			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	235,456	298,083	**p-value (t-test)**	**2,99323E-06**	
005.machado.marcha.fÃºnebre.portparte2.txt	248,973	301,349			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	228,763	293,405			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	232,858	320,556			
INGLÊS - PORTUGUÊS					
titulo	incidsintnominais	DENSNP 'Noun Phrase Incidence Score (per thousand words)'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	229,35	243,808		229,4	243,808
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	235,903	280,369		242,4	281,952
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	248,884	283,535		235,3	278,5845
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	232,143	284,394		220,9	276,19
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	238,494	272,775		196,9	292,556
004.poe.imp.of.perversity.port.txtparte1.txt	220,889	276,19		231,9	267,831
005.poe.tell-tale.heart.port.txtparte1.txt	196,866	292,556		220,4	277,089
006.poe.berenice.port.DUAS.PARTES.txt	235,029	260,071		233,8	256,188
006.poe.berenice.port.DUAS.PARTES2.txt	228,673	275,591		233,0	320,924
007.poe.eleonora.port.txt	220,423	277,089		214,3	267,8525
008.poe.red.masque.port.txt	233,808	256,188		244,2	299,363
009.poe.cask.of.amontillado.port.txt	233,01	320,924		304,9	284,314
010.poe.man.of.crowd.port.DUAS.PARTES.txt	224,85	267,597		239,6	291,3995
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	203,778	268,108		234,0	261,353
001.james.joyce.araby.port.txt	244,18	299,363	**Média:**	234,3546429	278,5288929
002.virginia.woolf.monday.tuesday.port.txt	304,918	284,314	**DVP:**	23,83421098	19,36421519
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	239,651	295,931			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	239,577	286,868			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	226,966	259,259	**p-value (t-test)**	**7,64928E-05**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	241,094	263,447			

REFERÊNCIAS ANAFÓRICAS

PORTUGUÊS - INGLÊS

titulo	refana	CREFPau 'Anaphor reference, all distances, unweighted'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	0,318966	0,327		0,32	0,327
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	0,553571	0,224		0,51	0,201
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	0,465347	0,178		0,37	0,189
003.coelho.neto.firmo.port.TXTparte1.txt	0,369231	0,189		0,33	0,2815
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,193548	0,262		0,12	0,168
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	0,462264	0,301		0,20	0,291
005.humberto.campos.caveiras.port.TXTparte1.txt	0,116071	0,168		0,70	0,2815
006.coelho.neto.duplo.port.TXTparte1.txt	0,195122	0,291		0,49	0,144
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	0,930233	0,334		0,90	0,242
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	0,466667	0,229		0,17	0,0735
008.humberto.campos.vinganÃ§a.port.txt	0,486726	0,144		0,63	0,219
001.machado.cartomante.port.TXTparte1.txt	0,992126	0,232		0,48	0,382
001.machado.cartomante.port.TXTparte2.txt	0,8	0,252		0,55	0,2625
002.machado.viver.port.TXTparte1.txt	0,163009	0,147		0,48	0,2795
002.machado.viver.portpart2.txt	0,175325	0	**Média:**	0,444445429	0,238678571
003.machado.cantiga.port.TXTparte1.txt	0,625	0,219	**DVP:**	0,215090138	0,079667484
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	0,640777	0,354			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	0,327273	0,41			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	0,508571	0,264	**p-value (t-test)**	**0,002853217**	
005.machado.marcha.fÃºnebre.portparte2.txt	0,589744	0,261			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	0,554795	0,313			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	0,39899	0,246			

INGLÊS - PORTUGUÊS

titulo	refana	CREFPau 'Anaphor reference, all distances, unweighted'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	0,666667	0,298		0,67	0,298
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	0,291339	0,23		0,30	0,1565
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	0,302817	0,083		0,57	0,332
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	0,747368	0,309		0,30	0,368
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	0,396396	0,355		0,45	0,453
004.poe.imp.of.perversity.port.txtparte1.txt	0,297101	0,368		0,58	0,276
005.poe.tell-tale.heart.port.txtparte1.txt	0,446328	0,453		0,68	0,317
006.poe.berenice.port.DUAS.PARTES.txt	0,452055	0,169		0,33	0,107
006.poe.berenice.port.DUAS.PARTES2.txt	0,7	0,383		0,24	0,263
007.poe.eleonora.port.txt	0,679012	0,317		0,64	0,304
008.poe.red.masque.port.txt	0,327103	0,107		0,43	0,354
009.poe.cask.of.amontillado.port.txt	0,240741	0,263		0,22	0,044
010.poe.man.of.crowd.port.DUAS.PARTES.txt	0,529851	0,249		0,47	0,324
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	0,740741	0,359		1,02	0,1885
001.james.joyce.araby.port.txt	0,427632	0,354	**Média:**	0,490631	0,270357143
002.virginia.woolf.monday.tuesday.port.txt	0,222222	0,044	**DVP:**	0,218194468	0,110352831
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	0,49162	0,294			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	0,44186	0,354			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	1,0641	0,167	**p-value (t-test)**	**0,002885032**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	0,965909	0,21			

REFERÊNCIAS ANAFÓRICAS (ADJACENTES)

PORTUGUÊS - INGLÊS					
titulo	refanaadj	CREFP1u 'Anaphor reference, adjacent, unweighted'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	0,181034	0,446		0,18	0,446
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte1.txt	0,357143	0,434		0,36	0,415
002.humberto.campos.promessa.port.DUAS.PARTES.TXTparte2.txt	0,366337	0,396		0,33	0,325
003.coelho.neto.firmo.port.TXTparte1.txt	0,330769	0,325		0,22	0,433
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte1.txt	0,104839	0,391		0,08	0,302
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTparte2.txt	0,339623	0,475		0,09	0,5
005.humberto.campos.caveiras.port.TXTparte1.txt	0,080357	0,302		0,55	0,5105
006.coelho.neto.duplo.port.TXTparte1.txt	0,085366	0,5		0,42	0,331
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES.txt	0,767442	0,609		0,53	0,461
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PARTES2.txt	0,341667	0,412		0,10	0,101
008.humberto.campos.vinganÃ§a.port.txt	0,424779	0,331		0,41	0,405
001.machado.cartomante.port.TXTparte1.txt	0,551181	0,475		0,35	0,581
001.machado.cartomante.port.TXTparte2.txt	0,51	0,447		0,39	0,425
002.machado.viver.port.TXTparte1.txt	0,094044	0,202		0,34	0,424
002.machado.viver.portpart2.txt	0,11039	0	**Média:**	0,311717068	0,40425
003.machado.cantiga.port.TXTparte1.txt	0,40625	0,405	**DVP:**	0,155487266	0,115187598
004.machado.enfermeiro.port.DUAS PARTES.TXTparte1.txt	0,436893	0,567			
004.machado.enfermeiro.port.DUAS PARTES.TXTparte2.txt	0,263636	0,595			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	0,365714	0,433	**p-value (t-test)**	**0,038936208**	
005.machado.marcha.fÃºnebre.portparte2.txt	0,423077	0,417			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte1.txt	0,431507	0,461			
006.machado.vida.eterna.port.DUAS.PARTES.TXTparte2.txt	0,247475	0,387			
INGLÊS - PORTUGUÊS					
titulo	refanaadj	CREFP1u 'Anaphor reference, adjacent, unweighted'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	0,627451	0,633		0,63	0,633
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte1.txt	0,204724	0,352		0,22	0,28
002.poe.mesmeric.revelation.port.DUAS.PARTES.txtparte2.txt	0,225352	0,208		0,41	0,6355
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	0,494737	0,667		0,18	0,678
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	0,315315	0,604		0,31	0,618
004.poe.imp.of.perversity.port.txtparte1.txt	0,181159	0,678		0,31	0,5355
005.poe.tell-tale.heart.port.txtparte1.txt	0,310734	0,618		0,52	0,635
006.poe.berenice.port.DUAS.PARTES.txt	0,232877	0,367		0,20	0,301
006.poe.berenice.port.DUAS.PARTES2.txt	0,38	0,704		0,15	0,438
007.poe.eleonora.port.txt	0,518519	0,635		0,43	0,5635
008.poe.red.masque.port.txt	0,196262	0,301		0,28	0,633
009.poe.cask.of.amontillado.port.txt	0,148148	0,438		0,17	0,05
010.poe.man.of.crowd.port.DUAS.PARTES.txt	0,358209	0,49		0,34	0,459
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	0,493827	0,637		0,66	0,453
001.james.joyce.araby.port.txt	0,282895	0,633	**Média:**	0,341391714	0,49375
002.virginia.woolf.monday.tuesday.port.txt	0,166667	0,05	**DVP:**	0,166612688	0,179861458
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	0,340782	0,435			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	0,333333	0,483			
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES.txt	0,782051	0,429	**p-value (t-test)**	**0,008856169**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PARTES2.txt	0,534091	0,477			

ÍNDICE FLESCH

PORTUGUÊS - INGLÊS					
titulo	flesch	READFRE 'Flesch Reading Ease Score (0-100)'		port	ingl
001.lima.barreto.cazuza.port.TXTparte1.txt	60,5612	69,381		60,56	69,381
002.humberto.campos.promessa.port.DUAS.PARTE	49,196	65,06		54,50	70,1615
002.humberto.campos.promessa.port.DUAS.PARTE	59,8123	75,263		70,07	84,792
003.coelho.neto.firmo.port.TXTparte1.txt	70,066	84,792		61,40	71,39
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTp	62,2638	72,501		66,76	84,299
004.lima.barreto.javanÃªs.port.DUAS.PARTES.TXTp	60,5459	70,279		63,66	80,586
005.humberto.campos.caveiras.port.TXTparte1.txt	66,7601	84,299		58,21	68,5595
006.coelho.neto.duplo.port.TXTparte1.txt	63,655	80,586		52,20	69,22
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PA	48,4097	57,128		60,87	69,4055
007.lima.barreto.nÃºmero.sepultura.port.DUAS.PA	68,0199	79,991		71,66	60,249
008.humberto.campos.vinganÃ§a.port.txt	52,1974	69,22		67,09	78,623
001.machado.cartomante.port.TXTparte1.txt	58,0614	64,562		61,25	79,948
001.machado.cartomante.port.TXTparte2.txt	63,6744	74,249		65,84	78,627
002.machado.viver.port.TXTparte1.txt	71,7149	71,866		63,35	79,176
002.machado.viver.portpart2.txt	71,612	48,632	**Média:**	62,672725	74,60125
003.machado.cantiga.port.TXTparte1.txt	67,0887	78,623	**DVP:**	5,482872085	7,186152
004.machado.enfermeiro.port.DUAS PARTES.TXTpa	63,4519	79,857			
004.machado.enfermeiro.port.DUAS PARTES.TXTpa	59,0479	80,039			
005.machado.marcha.fÃºnebre.port.TXTparte1.txt	65,5781	76,549	**p-value (t-test)**	**5,05101E-05**	
005.machado.marcha.fÃºnebre.portparte2.txt	66,0924	80,705			
006.machado.vida.eterna.port.DUAS.PARTES.TXTp	59,9335	76,003			
006.machado.vida.eterna.port.DUAS.PARTES.TXTp	66,7654	82,349			
INGLÊS - PORTUGUÊS					
titulo	flesch	READFRE 'Flesch Reading Ease Score (0-100)'		port	ingl
001.poe.oval.portrait.port.txtparte1.txt	40,6316	61,068		40,63	61,068
002.poe.mesmeric.revelation.port.DUAS.PARTES.t	43,4508	56,588		46,03	55,0575
002.poe.mesmeric.revelation.port.DUAS.PARTES.t	48,603	53,527		38,28	60,7585
003.poe.black.cat.port.DUAS.PARTES.txtparte1.txt	37,3544	58,884		42,41	55,021
003.poe.black.cat.port.DUAS.PARTES.txtparte2.txt	39,2046	62,633		67,43	81,276
004.poe.imp.of.perversity.port.txtparte1.txt	42,4061	55,021		40,76	50,7925
005.poe.tell-tale.heart.port.txtparte1.txt	67,4281	81,276		42,04	53,556
006.poe.berenice.port.DUAS.PARTES.txt	32,5475	40,596		44,80	63,976
006.poe.berenice.port.DUAS.PARTES2.txt	48,9698	60,989		62,98	78,822
007.poe.eleonora.port.txt	42,0364	53,556		39,14	52,0905
008.poe.red.masque.port.txt	44,7998	63,976		53,88	79,925
009.poe.cask.of.amontillado.port.txt	62,9792	78,822		39,69	67,937
010.poe.man.of.crowd.port.DUAS.PARTES.txt	35,1614	43,791		65,28	82,2635
010.poe.man.of.crowd.port.DUAS.PARTES2.txt	43,119	60,39		51,53	58,702
001.james.joyce.araby.port.txt	53,8756	79,925	**Média:**	48,20498929	64,37468
002.virginia.woolf.monday.tuesday.port.txt	39,6915	67,937	**DVP:**	10,29067408	11,60889
003.O.Henry.Red.Chief.port.DUAS.PARTES.txt	66,0685	82,581			
003.O.Henry.Red.Chief.port.DUAS.PARTES2.txt	64,4967	81,946			
004.nathaniel.hawthorne.wakefield.port.DUAS.PAI	48,3449	52,58	**p-value (t-test)**	**3,38255E-07**	
004.nathaniel.hawthorne.wakefield.port.DUAS.PAI	54,7225	64,824			