

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS

AMANDA MICHEL RAUPP

**CARACTERÍSTICAS LEXICAIS DAS PRODUÇÕES ESCRITAS DO EXAME
CELPE-BRAS NA TAREFA 3 DE 2016-2: UMA PESQUISA GUIADA POR CORPUS**

Porto Alegre
2024

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS

AMANDA MICHEL RAUPP

**CARACTERÍSTICAS LEXICAIS DAS PRODUÇÕES ESCRITAS DO EXAME
CELPE-BRAS NA TAREFA 3 DE 2016-2: UMA PESQUISA GUIADA POR CORPUS**

Trabalho de conclusão de curso apresentado como
requisito parcial para o grau de Licenciada em
Letras pela Universidade Federal do Rio Grande
do Sul.

Orientadora: Prof^ª. Dr^ª. Juliana Roquele Schoffen

Porto Alegre
2024

AGRADECIMENTOS

Em primeiro lugar, quero agradecer a Deus por tantas graças concedidas ao longo da trajetória acadêmica e por me dar forças quando eu mais precisava.

À Universidade Federal do Rio Grande do Sul (UFRGS), por todas as oportunidades que tive na graduação e por formar a profissional que sou hoje, me concedendo um ensino gratuito e de qualidade.

À minha orientadora, Juliana, por me apresentar ao mundo da avaliação de língua, por todas as dicas e conselhos valiosos, e por acreditar em mim. Dedico a ela todos os meus enunciados mais bem escritos.

À minha família, meus pais e minhas irmãs, por todo o suporte concedido desde o início da graduação. Eu não teria conseguido realizar o sonho de me formar em Letras sem os sacrifícios e o incentivo constante da minha família. Ao meu noivo, Davi, por ser o meu maior incentivador de sonhos, e por sempre acreditar no meu potencial.

Aos amigos que a faculdade me deu, Aryelle, Gabriela, Keverton, Lucas, Veridiana, Felipe, Julia, Renata, Vitória e Carol, por terem feito o caminho da graduação ser muito mais leve e por me incentivarem a sonhar e almejar mais no mundo acadêmico. Às minhas amigas de vida, Bruna e Carol, por aguentarem meus surtos mesmo não entendendo nada do assunto do meu TCC e por estarem sempre do meu lado.

À banca do meu TCC, Elisa Stumpf e Luiza Sarmiento Divino, por terem aceitado a missão de ler esta pesquisa e por todas as contribuições que me ajudarão a crescer como pesquisadora e profissional na área de Letras.

Ao grupo de pesquisa Avalia, por todas as discussões enriquecedoras e todas as parcerias que foram criadas neste grupo. De forma mais especial ainda, à minha amiga de pesquisa e de vida, Brenda, por estar sempre disposta a me ajudar na minha pesquisa.

A todos aqueles que, de alguma forma, contribuíram para que eu pudesse realizar esta pesquisa.

“Estuda - Estuda com empenho - Se tens de ser sal e luz, necessitas de ciência, de idoneidade.

*Ou julgas que, por seres preguiçoso e comodista, hás de receber ciência infusa?”
(Caminho, 340, São Josemaria Escrivá)*

RESUMO

Este trabalho tem por objetivo analisar quantitativamente as características lexicais dos textos avaliados com nota 2 e nota 5 na Tarefa 3 da edição 2016-2 do Celpe-Bras utilizando ferramentas de Linguística de Corpus, a fim de contribuir para uma descrição mais detalhada dos níveis de proficiência avaliados na Parte Escrita do exame. O Celpe-Bras busca avaliar, em práticas de usos linguísticos reais, o nível de proficiência em português como língua adicional (BRASIL, 2020). Apoiado na noção bakhtiniana de gêneros do discurso (BAKHTIN, 2003), o Celpe-Bras propõe, na Parte Escrita, tarefas nas quais são apresentados ao examinando situações comunicativas específicas, um papel de enunciador específico que deve endereçar o texto a interlocutores determinados com propósitos específicos (SCHLATTER ET AL., 2009). Todos esses aspectos devem ser contemplados nos textos para cumprir os propósitos da tarefa. Os examinandos precisam lançar mão de elementos lexicais para realizar a tarefa com sucesso, selecionados a partir do enunciado da tarefa, do texto de insumo e de seu repertório pessoal (SCHOFFEN, 2009; MENDEL, 2019). A Linguística de Corpus é muito útil para evidenciar esses elementos em avaliações de língua, contribuindo, por exemplo, para o refinamento dos descritores de proficiência, diferenciando suas notas (GABLASOVA, 2020; CALLIES; GÖTZ, 2015). Para este trabalho, foram utilizados dois subcorpora do CorCel (SCHOFFEN, et al., em preparação), nomeados corcel20162t3n2 (referente aos textos que receberam nota 2) e corcel20162t3n5 (referente aos textos que receberam nota 5), com o intuito de realizar análises quantitativas e automáticas, e, em conjunto, análises qualitativas (KENNEDY, 2001). O *software* escolhido para o estudo foi o *Sketch Engine* (KILGARRIFF et al., 2004), que contém várias ferramentas disponíveis para realizar as análises, como *keywords*, *wordlist*, *n-grams* e *concordance*, utilizadas neste trabalho. Com base em estudos anteriores (DIVINO, 2021; 2024; HANAUER, 2023; SOSTRUZNIK, 2023), esta pesquisa, guiada por corpus, identificou características lexicais importantes para a descrição de padrões nos textos que receberam as notas 2 e 5 na Tarefa 3 de 2016-2. Identificou-se que os textos de nota 5 são, em média, mais extensos do que os textos de nota 2. Ademais, em ambos os subcorpora há a presença de inadequações ortográficas, mas essas inadequações são mais recorrentes nos textos de nota 2 e podem ter influenciado alguns resultados das análises realizadas no corpus desta nota, como o cálculo *Type-Token-Ratio* (TTR), que mostra a riqueza lexical de um corpus. Ao recuperar informações do material de insumo, ambos os subcorpora recuperam informações importantes para o cumprimento da tarefa, em alguns casos, utilizando paráfrases e sinônimos. No entanto, os textos de nota 5 fazem isso de forma mais articulada, enquanto os textos de nota 2 realizam essa recuperação de forma fragmentada, não organizando o texto de forma clara. Com base nas análises realizadas, este estudo busca contribuir para a descrição mais robusta dos níveis de proficiência do Celpe-Bras, como também para o refinamento dos parâmetros de avaliação da parte escrita, dando evidência para a importância das análises de corpus para a área de avaliação de proficiência em línguas.

Palavras-chave: Exame Celpe-Bras; Níveis de Proficiência; Português como Língua Adicional; Linguística de Corpus;

ABSTRACT

This study analyzes the characteristics of texts rated grades 2 and 5 in Task 3 of the 2016-2 edition of Celpe-Bras using Corpus Linguistics tools, aiming to contribute to a more detailed description of the proficiency levels assessed in the written section of the exam. Celpe-Bras assesses Portuguese proficiency as an additional language through real linguistic practices (BRASIL, 2020). Grounded in Bakhtin's notion of discourse genres (BAKHTIN, 2003), the exam presents examinees with tasks requiring them to address texts to specific interlocutors with defined purposes (SCHLATTER ET AL., 2009). All these aspects must be incorporated into the texts to meet the task's objectives. Examinees must use lexical elements drawn from the task's prompt, the input text, and their personal repertoire to successfully complete the task (SCHOFFEN, 2009; MENDEL, 2019). Corpus Linguistics is a valuable methodology in language assessment, helping to refine proficiency descriptors by differentiating grades (GABLASOVA, 2020; CALLIES; GÖTZ, 2015). This study utilized two subcorpora from CorCel (SCHOFFEN et al., in preparation), named corcel20162t3n2 (for texts rated grade 2) and corcel20162t3n5 (for texts rated grade 5), to conduct both quantitative and qualitative analyses (KENNEDY, 2001). The study employed Sketch Engine software (KILGARRIFF et al., 2004), chosen for its tools, including keywords, wordlists, n-grams, and concordance. Previous corpus-driven studies (DIVINO, 2021; 2024; HANAUER, 2023; SOSTRUZNIK, 2023) identified key characteristics that describe patterns in texts rated grades 2 and 5 in Task 3 of 2016-2 edition. The analysis revealed that texts rated grade 5 are generally longer than those rated grade 2. Both subcorpora exhibited spelling inadequacies, but these were more frequent in grade 2 texts and may have influenced some results, such as the Type-Token Ratio (TTR), which measures lexical richness. When retrieving information from the input material, both subcorpora extracted essential information, sometimes using paraphrases and synonyms. However, grade 5 texts did so more cohesively, while grade 2 texts retrieved information in a fragmented and less organized manner. This study not only contributes to a more robust description of Celpe-Bras proficiency levels but also offers valuable insights for refining the assessment parameters of the written section, emphasizing the significance of corpus analysis in language proficiency assessment.

Keywords: Celpe-Bras Exam; Proficiency levels; Portuguese as an Additional Language; Corpus Linguistics.

LISTA DE FIGURAS

| | |
|--|----|
| Figura 1 - Número de examinandos anuais do Exame Celpe-Bras..... | 16 |
| Figura 2 - Total de produções sobre o Celpe-Bras por tipo de publicação..... | 17 |
| Figura 3 - Enunciado da Tarefa 3 da Edição de 2016-2..... | 22 |
| Figura 4 - Texto de Insumo da Tarefa 3 da Edição de 2016-2..... | 22 |
| Figura 5 - Resposta Esperada na Tarefa 3 da Edição de 2016-2..... | 24 |
| Figura 6 - <i>Dashboard</i> do <i>Sketch Engine</i> | 39 |
| Figura 7 - Critérios selecionados na ferramenta <i>Keywords</i> | 40 |
| Figura 8 - Resultados da ferramenta <i>Keywords</i> | 41 |
| Figura 9 - Critérios selecionados na ferramenta <i>Wordlist</i> | 42 |
| Figura 10 - Resultados da ferramenta <i>Wordlist</i> | 43 |
| Figura 11 - Critérios selecionados na ferramenta <i>N-grams</i> | 44 |
| Figura 12 - Resultados da ferramenta <i>N-grams</i> | 45 |
| Figura 13 - Critérios selecionados na ferramenta <i>Concordance</i> | 46 |
| Figura 14 - Resultados da ferramenta <i>Concordance</i> | 46 |
| Figura 15 - Página do <i>Corpus Info</i> | 47 |

LISTA DE QUADROS

| | |
|---|----|
| Quadro 1 - Etapa 1 e 2 da Parte Oral do Celpe-Bras..... | 19 |
| Quadro 2 - Parâmetros de Avaliação da Parte Escrita..... | 28 |
| Quadro 3 - Extensão dos subcorpora..... | 49 |
| Quadro 4 - Média de <i>types</i> , <i>tokens</i> e sentenças por texto em cada subcorpora..... | 50 |
| Quadro 5 - Resultado do <i>Type-Token-Ratio</i> | 50 |
| Quadro 6 - 15 primeiras <i>keywords</i> e Relative DOCF (%) de cada subcorpora..... | 51 |
| Quadro 7 - <i>Keywords</i> que não estão presentes no material de insumo ordenados pelo Relative DOCF (%) de cada subcorpora..... | 54 |
| Quadro 8 - Extensão do corpus corcel20162t3mi..... | 57 |
| Quadro 9 - 15 primeiras <i>keywords</i> do corpus corcel20162t3mi..... | 57 |
| Quadro 10 - Quantidade de palavras que aparecem no mínimo em 5% dos textos em cada subcorpora..... | 59 |
| Quadro 11 - Quantidade de palavras que aparecem no mínimo em 10% dos textos em cada subcorpora..... | 60 |
| Quadro 12 - Quantidade de palavras que aparecem no mínimo em 20% dos textos em cada subcorpora..... | 61 |
| Quadro 13 - Quantidade de palavras que aparecem no mínimo em 50% dos textos em cada subcorpora..... | 62 |
| Quadro 14 - Quantidade de itens em sequência que aparecem no mínimo em 10% dos textos em cada subcorpora..... | 65 |
| Quadro 15 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n5 com 4 elementos em sequência..... | 66 |
| Quadro 16 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n5 com 5 elementos em sequência..... | 68 |
| Quadro 17 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n5 com 6 elementos em sequência..... | 69 |
| Quadro 18 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n2 com 4 elementos em sequência..... | 70 |
| Quadro 19 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n2 com 5 elementos em sequência..... | 72 |
| Quadro 20 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n5 com 6 elementos em sequência..... | 73 |

| | |
|--|----|
| Quadro 21 - Resultados encontrados no <i>left context</i> do corcel20162t3n5..... | 75 |
| Quadro 22 - Resultados encontrados no <i>right context</i> do corcel20162t3n5..... | 75 |
| Quadro 23 - Resultados encontrados no <i>left context</i> do corcel20162t3n2..... | 77 |
| Quadro 24 - Resultados encontrados no <i>right context</i> do corcel20162t3n2..... | 78 |

LISTA DE ANEXOS

| | |
|--|-----|
| Anexo 1 – Tarefa 3 de 2016-2..... | 98 |
| Anexo 2 - Lista de <i>keywords</i> do corcel20162t3mi..... | 99 |
| Anexo 3 - Resultado na íntegra da ferramenta <i>n-grams</i> com 4 elementos em sequência que aparecem no mínimo em 10% no corcel20162t3n5..... | 103 |
| Anexo 4 - Resultado na íntegra da ferramenta <i>n-grams</i> com 5 elementos em sequência que aparecem no mínimo em 10% no corcel20162t3n5..... | 106 |
| Anexo 5 - Resultado na íntegra da ferramenta <i>n-grams</i> com 6 elementos em sequência que aparecem no mínimo em 10% no corcel20162t3n5..... | 108 |
| Anexo 6 - Resultado na íntegra da ferramenta <i>n-grams</i> com 4 elementos em sequência que aparecem no mínimo em 10% no corcel20162t3n2..... | 110 |
| Anexo 7 - Resultado na íntegra da ferramenta <i>n-grams</i> com 5 elementos em sequência que aparecem no mínimo em 10% no corcel20162t3n2..... | 112 |
| Anexo 8 - Resultado na íntegra da ferramenta <i>n-grams</i> com 6 elementos em sequência que aparecem no mínimo em 10% no corcel20162t3n2..... | 113 |
| Anexo 9 - Resultado completo da ferramenta <i>concordance</i> no corcel20162t3n5..... | 113 |
| Anexo 10 - Resultado completo da ferramenta <i>concordance</i> no corcel20162t3n2..... | 122 |

SUMÁRIO

| | |
|--|-----------|
| INTRODUÇÃO..... | 12 |
| 1 O EXAME CELPE-BRAS..... | 15 |
| 1.1 A noção de proficiência do exame..... | 17 |
| 1.2 A Parte Escrita..... | 20 |
| 1.2.1 Exemplo de tarefa..... | 21 |
| 1.2.2 Avaliação da Parte Escrita..... | 27 |
| 1.2.3 Parâmetros de avaliação da Parte Escrita..... | 28 |
| 2 PRESSUPOSTOS TEÓRICOS..... | 31 |
| 2.1 Linguística de Corpus..... | 31 |
| 2.1.1 Linguística de Corpus e avaliação de língua..... | 32 |
| 2.2 Estudos sobre o Celpe-Bras à luz da Linguística de Corpus..... | 33 |
| 3 METODOLOGIA..... | 35 |
| 3.1 Objetivos de pesquisa..... | 35 |
| 3.2 CorCel: o corpus de pesquisa..... | 35 |
| 3.3 Procedimentos para as análises..... | 38 |
| 3.3.1 O Sketch Engine..... | 38 |
| 3.3.1.1 Keywords..... | 39 |
| 3.3.1.2 Wordlist..... | 41 |
| 3.3.1.3 N-grams..... | 43 |
| 3.3.1.4 Concordance..... | 45 |
| 3.3.2 Type-token-ratio (TTR)..... | 47 |
| 4 ANÁLISE DOS DADOS..... | 49 |
| 4.1 Extensão dos subcorpora e type-token-ratio (TTR)..... | 49 |
| 4.2 Keywords..... | 51 |
| 4.3 Wordlist..... | 58 |
| 4.4 N-grams..... | 64 |
| 4.5 Concordance..... | 74 |

| | |
|--|-----------|
| 5 DISCUSSÃO DOS DADOS..... | 81 |
| 6 CONSIDERAÇÕES FINAIS..... | 89 |
| REFERÊNCIAS BIBLIOGRÁFICAS..... | 92 |
| ANEXOS..... | 98 |

INTRODUÇÃO

Ingressar na Universidade Federal do Rio Grande do Sul (UFRGS) em 2020 foi a realização de um sonho, mas que logo se tornou em uma época desafiadora, talvez a mais desafiadora dos últimos anos que qualquer calouro já passou na universidade: a pandemia. Apesar de a realidade naquele momento ser o ensino remoto e com certeza temos tido mais motivos para desistir do que para seguir nos estudos, insisti muito para que meu tempo como graduanda não fosse perdido. Por isso, logo na primeira oportunidade que tive, me inscrevi para ser monitora da disciplina *Leitura e Produção Textual*, ministrada pela Prof.^a Dr.^a Juliana Roquele Schoffen como disciplina obrigatória no curso de Letras. Nesta disciplina, tive a oportunidade de ler os textos produzidos pelos alunos e ajudá-los na reescrita com os bilhetes orientadores. Já nessa época, eu entendia, provavelmente de forma inconsciente, que todo enunciado produzido pelos alunos era único e irrepetível (BAKHTIN, 2003), e nisso estava a beleza de ler cada produção escrita, de perceber cada mecanismo escolhido pelos alunos, a partir do repertório pessoal de cada um, para deixar a sua voz autoral no texto (SIMÕES et al., 2012).

Ao longo das disciplinas que ia cursando na faculdade, começou a ficar claro para mim que escrever é, sem dúvidas, “produzir textos com determinados propósitos para determinados interlocutores e, assim, poder inserir-se de modo mais participativo na sociedade” (RIO GRANDE DO SUL, 2009, p. 39). Isso mudou meu olhar sobre o texto, e me instigou a querer mais. Foi então que decidi, em setembro de 2022, ingressar no grupo de pesquisa Avalia - Avaliação de Uso da Linguagem, coordenado pela Prof.^a Dr.^a Juliana Roquele Schoffen, e me encantei com o mundo da avaliação de proficiência de língua, ainda mais da língua em produções escritas. Foi neste grupo de pesquisa que conheci o Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras), entendi sua relevância na vida dos examinandos, como também o seu construto, que consiste em avaliar o desempenho do examinando em diversas práticas sociais de uso da língua portuguesa, buscando nivelar a sua proficiência (BRASIL, 2020).

Após isso, depois de muitos semestres sem oferta, cursei a disciplina de *Avaliação e Educação Linguística*, ministrada pela Prof.^a Dr.^a Juliana Roquele Schoffen como eletiva no curso de Letras da UFRGS. Foi então que compreendi os diversos desafios que circulam o ato de avaliar, entendendo os diferentes tipos de avaliação, e como produzir propostas de avaliação coerentes com o ensino comprometido da educação linguística. Nessa disciplina, tive a oportunidade de me aprofundar ainda mais no exame Celpe-Bras e entender que, para se escrever um texto proficiente em uma língua, é imprescindível que: “a) se tenha o que

dizer; b) se tenha uma razão para dizer o que se tem a dizer; c) se tenha para quem dizer o que se tem a dizer; d) o locutor se constitua como tal, enquanto sujeito que diz o que diz para quem diz; e) se escolham as estratégias para realizar (a), (b), (c) e (d)” (GERALDI, 1991, p. 137).

Em meados de abril de 2023, iniciei como secretária do Posto Aplicador do Exame Celpe-Bras que se localiza na UFRGS, auxiliando a Prof.^a Dr.^a Juliana Roquele Schoffen nos preparativos para o exame e realizando as inscrições dos examinandos. Foi um momento de muito aprendizado onde pude estar mais perto do exame e dos examinandos, percebendo ainda mais o construto do Celpe-Bras sendo posto em prática.

Em conjunto com todos esses aprendizados, se dava continuidade no grupo de pesquisa o projeto “Níveis de proficiência certificados pelo exame Celpe-Bras: descrição do nível Avançado Superior”, com o objetivo de descrever melhor os níveis de proficiência no exame, em específico o nível mais alto de proficiência do Celpe-Bras. Para alcançar o objetivo posto, foi iniciado o processo de compilação de um corpus que inicialmente seria composto apenas dos textos de nível Avançado Superior, e seriam utilizados para análises qualitativas.

Porém, o projeto ganhou maiores proporções, almejando construir um corpus de 15.344 textos das edições de 2015-2, 2016-1, 2016-2 e 2017-1 contendo produções escritas de todos os níveis de proficiência, o CorCel (SCHOFFEN, et al., em preparação), como também abriu espaço para produzir-se análises quantitativas a partir da Linguística de Corpus (LC). Atualmente este corpus está no final do processo de digitação, chegando no número almejado, e seguindo para a parte de revisão dos textos digitados e etiquetamento a partir do protocolo criado (SCHOFFEN, et al., 2024). Apesar de a compilação do CorCel não estar finalizada ainda, alguns trabalhos já foram produzidos com o intuito de auxiliar na descrição dos níveis utilizando a LC, como Divino (2021; 2024), Hanauer (2023) e Sostruznik (2023). Estes estudos já mostram quanto a LC pode auxiliar na caracterização mais detalhada dos níveis de proficiência em produções escritas do Celpe-Bras, auxiliando professores, pesquisadores e elaboradores do exame.

Este presente trabalho busca, de forma dialógica (BAKHTIN, 2003), responder aos enunciados criados anteriormente e contribuir para a descrição dos níveis de proficiência do exame, como também busca, ao detalhar uma das tarefas do Celpe-Bras, identificar as características presentes nos textos que são importantes para diferenciar as notas 2 e 5 da parte escrita do exame, auxiliar os examinandos e os professores de língua portuguesa como

língua adicional, como também para a elaboração de materiais didáticos preparatórios para o exame.

Com isso em mente, surgiu a pergunta de pesquisa norteadora deste trabalho: **Quais padrões lexicais caracterizam os textos avaliados com nota 2 e com nota 5 na Tarefa 3 da edição 2016-2 do Celpe-Bras?** com duas subdivisões: a) Como a extensão e os mecanismos lexicais utilizados nos textos de nota 2 e nota 5 da Tarefa 3 da edição 2016-2 do Celpe-Bras são importantes para a caracterização dos níveis de proficiência? e b) Como os textos avaliados com as notas 2 e 5 na Tarefa 3 da edição 2016-2 do Celpe-Bras recuperam as informações do material de insumo?

De modo a responder esta pergunta de pesquisa e suas subdivisões, o presente texto está dividido em seis capítulos. No primeiro, apresento o exame Celpe-Bras, sua noção de proficiência e a organização da Parte Escrita, e analiso a Tarefa 3 de 2016-2 e seus parâmetros de avaliação. No segundo, trago alguns pressupostos teóricos importantes para embasar a minha pesquisa: discuto sobre a Linguística de Corpus (LC) na área de avaliação de proficiência e apresento alguns trabalhos importantes realizados anteriormente sobre o Celpe-Bras à luz da LC. No terceiro capítulo, apresento os objetivos do trabalho, o corpus utilizado, o *software* e as ferramentas de análise utilizadas, e detalho os procedimentos metodológicos utilizados na pesquisa. No quarto capítulo, analiso os dados obtidos e, no quinto capítulo, discuto as análises apresentadas. Por fim, no capítulo seis, apresento as considerações finais, mostrando também algumas limitações desta pesquisa e ideias para trabalhos futuros.

1 EXAME CELPE-BRAS

No início dos anos 90, via-se a necessidade de desenvolver uma prova unificada que atestasse a proficiência em língua portuguesa dos estudantes que faziam parte do Programa de Estudantes-Convênio de Graduação (PEC-G) e Pós-Graduação (PEC-PG). Neste cenário, com a Portaria nº 1.787 no ano de 1994, foi instituído o Celpe-Bras, que teve sua primeira edição em 1998. Esta primeira edição alcançou 127 examinandos em 8 Postos Aplicadores, sendo um deles a UFRGS, através do Ministério da Educação (MEC), que ficou responsável pelo exame até o primeiro semestre de 2009.

Desde o segundo semestre de 2009, a responsabilidade do exame é do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). O Celpe-Bras conta hoje com mais de 130 Postos Aplicadores do exame, entre eles postos nacionais e internacionais em 42 países (DIVINO, 2024).

No exterior, o resultado do exame pode ser usado para comprovação de proficiência em Português como Língua Adicional (PLA); e, nacionalmente, o Celpe-Bras é usado de forma obrigatória para ingressar em cursos de graduação através do PEC-G e PEC-PG, para validação de diplomas de alguns profissionais estrangeiros que planejam trabalhar no Brasil e como uma das possibilidades de comprovação de proficiência para fins de naturalização.

De acordo com o Documento Base do Exame Celpe-Bras (BRASIL, 2020), o Celpe-Bras tornou-se, com o passar dos anos, “um sólido instrumento de avaliação de proficiência, de fortalecimento da política linguística brasileira” (p. 13) e também produz “efeitos retroativos positivos sobre o ensino de língua portuguesa para estrangeiros e sobre a formação de professores nessa área” (p. 13), já que um dos objetivos ao criar o exame era justamente oportunizar mudanças no ensino e aprendizado de português como língua adicional (SCARAMUCCI, 2008).

Schlatter et al (2005) conceituam efeito retroativo na área de avaliação como o efeito que os exames causam no ensino, direcionando os conteúdos aprendidos em sala de aula, por exemplo, para os conteúdos que serão cobrados nas provas de larga escala.

Além desses impactos no ensino e aprendizagem de língua, o exame também influencia na vida dos examinandos, podendo ter efeitos positivos ou negativos. Por exemplo, um efeito positivo na vida do examinando pode ser conceder-lhe a oportunidade de ingressar no ensino superior, uma bolsa de estudos ou um trabalho de qualidade; já um efeito negativo seria impedir o indivíduo de ingressar na faculdade desejada, ou de não obter o trabalho que queria e precisava (SHOHAMY, 2001; DIVINO, 2021). Por isso a importância de investir em

pesquisas sobre o exame e seus parâmetros de avaliação, auxiliando na credibilidade do exame e na coerência do seu construto. Por estes motivos, o Celpe-Bras é considerado um exame de alta relevância (SCHLATTER et al., 2009).

A Figura 1 apresenta o número de examinandos que realizaram o exame todos os anos desde que começou a ser aplicado.

Figura 1 - Número de examinandos anuais do Exame Celpe-Bras

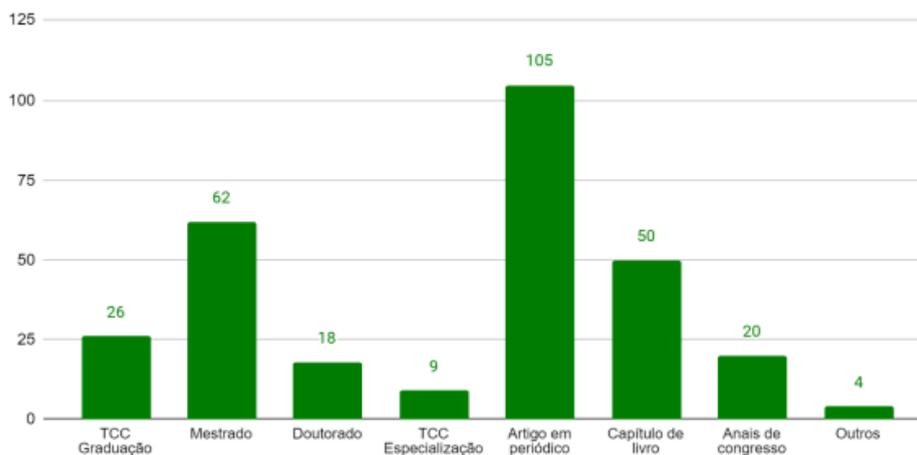


Fonte: Acervo Celpe-Bras. Disponível em: <https://www.ufrgs.br/acervocelpebras/dados-celpe-bras/>. Acesso em 01 jul. 2024.

A Figura 1 apresenta o número anual de examinando do Celpe-Bras, mostrando o aumento na procura pelo exame ao longo dos anos e também o grande declínio das inscrições em 2020 por conta da pandemia do COVID-19¹. Outro aspecto que mostra a relevância do exame é o grande número de pesquisas já realizadas sobre o Celpe-Bras. Na Figura 2 pode-se verificar este número e as suas divisões por tipo de trabalho.

Figura 2 - Total de produções sobre o Celpe-Bras por tipo de publicação (julho 2023)

¹ Em 2020, 2021 e 2022, o exame teve uma aplicação anual em razão das restrições da pandemia.



Fonte: Acervo Celpe-Bras. Disponível em: <https://www.ufrgs.br/acervocelpebras/dados-celpe-bras/>. Acesso em 01 jul. 2024.

A Figura 2 mostra o número total de produções feitas sobre o exame, divididas também por tipo de texto. Essa Figura está atualizada até julho de 2023 e, até essa data, foram publicados 26 TCCs de graduação, 62 dissertações de mestrado, 18 de doutorado, 9 TCCs de especialização, 105 artigos em periódicos, 50 capítulos de livros, 20 anais de congresso e mais 4 que não se enquadram nas outras categorias. Isso comprova a relevância do exame tanto para estudos acadêmicos como para a vida de muitos candidatos que são impactados pelo Celpe-Bras.

1.1 A NOÇÃO DE PROFICIÊNCIA DO EXAME

Diferentemente das práticas de ensino e avaliação estruturalistas da língua que se tinha até a década de 80, a comissão nomeada para o desenvolvimento do Celpe-Bras viu a necessidade de criar uma certificação que abrangesse a língua portuguesa em uso, onde os examinandos precisassem usá-la em práticas sociais letradas, já que o público-alvo da época eram estudantes que desejavam ingressar em universidades brasileiras. Com isso, o objetivo era propor para o ensino parâmetros que “se voltasse para oportunidades de uso da língua, visando à participação dos educandos em diferentes situações de comunicação em que o português fosse a língua de socialização entre os participantes” (SCHLATTER, 2014, p. 1).

Portanto, o exame avalia o desempenho do examinando em diversas práticas sociais de uso da língua portuguesa, buscando nivelar a sua proficiência, já que esta é sempre relativa, dependendo da necessidade de uso futuro da língua (BRASIL, 2020). Por isso, para o Celpe-Bras, a proficiência depende do contexto de uso em que está situada, em determinada prática social.

O exame certifica quatro níveis de proficiência, são eles: Intermediário, Intermediário Superior, Avançado e Avançado Superior. O Celpe-Bras, em oposição a outros exames de proficiência, não visa no Avançado Superior uma proficiência de um falante nativo ideal, que é sempre idealizada, pois a proficiência pode “variar de acordo com o gênero, o tema em questão, o interlocutor, o contexto, o barulho no ambiente, o estresse da situação, entre outros fatores” (BRASIL, 2020). Por ter a visão de que a proficiência é relativa, o construto do exame não prevê apenas um único nível de proficiência, mas vários níveis que levam em conta a situação de uso da língua com suas especificidades em diversos contextos. Dessa forma, “a proficiência é sempre definida localmente, por ser situada em contextos de uso, em determinada prática social” (BRASIL, 2020). A proficiência então, neste exame, se resume a saber usar a língua em diferentes níveis de comunicação.

Dessa forma, trata-se, então, de um exame de desempenho que avalia a capacidade de se engajar, de construir enunciados adequados em situações diversas de uso da língua portuguesa para desempenhar ações múltiplas no mundo (BRASIL. MEC, 2006; SCHOFFEN, 2009). Essa visão de proficiência leva em conta não só o domínio de formas gramaticais e itens lexicais, mas também a prática de uso da língua de forma contextualizada. Isso vai ao encontro da teoria bakhtiniana dos gêneros do discurso, à qual o exame Celpe-Bras se filia, que afirma que “cada enunciado é único e individual, mas cada domínio de uso de linguagem cria tipos relativamente estáveis de enunciados, que são chamados de gêneros do discurso” (BAKHTIN, 2003, p. 262). Esses enunciados sempre contam com “uma situação comunicativa específica, por alguém com um papel específico nessa situação, e são endereçados a interlocutores com propósitos específicos e também produzidos em um contexto” (SCHLATTER ET AL., 2009, p. 105-106). Além disso, os enunciados são processos dialógicos e axiológicos, o que, segundo Schoffen (2009, p. 82), significa sempre uma “resposta a enunciados anteriores a partir de uma rede de valores estabelecida por eles e em relação a eles. É o evento do enunciado que dará condições de atribuir-se sentido a ele, e não as palavras em si”. Nesse sentido, um enunciado é sempre uma resposta crítica e ativa a um enunciado anteriormente produzido, não existindo então um enunciado sem resposta, sendo sempre dialógico (SCHOFFEN, 2009).

Ademais, para Bakhtin (2003, p. 297), “cada enunciado é pleno de ecos e ressonâncias de outros enunciados com os quais está ligado pela identidade da esfera de comunicação discursiva”, isto é, segundo Souza (2002, p. 108), “[é] no interior do gênero do discurso determinado que o enunciado concreto ocupa uma posição definida em relação a um determinado tema”, sendo então o gênero do discurso um fator importante para definir a

interlocução em um enunciado. Dessa forma, o que constitui de fato um enunciado não é só o *que* é dito, mas também *como* é dito, os aspectos utilizados para chegar-se ao texto material que também torna-se um enunciado que será respondido futuramente por outros enunciados (SCHOFFEN, 2009).

Para verificar o nível de proficiência do examinando, o exame conta com duas partes a serem realizadas pelos participantes de forma presencial nos Postos Aplicadores, a Parte Oral (PO) e a Parte Escrita (PE). A PO tem o objetivo de avaliar a produção oral dos examinandos e consiste em uma interação que “parte de situações de comunicação próximas às reais, procurando simular, na medida do possível, conversas do dia a dia” (BRASIL, 2020, p. 41). Essas situações de comunicação são individuais, onde cada examinando tem seu horário específico para a prova, e duram em média 20 minutos. A PO é realizada pelo examinando, o avaliador-interlocutor, que tem a função de interlocutor, e o avaliador-observador, que observa a interação e faz a sua avaliação. A PO acontece em duas etapas que estão apresentadas no Quadro 1 abaixo:

Quadro 1 - Etapa 1 e 2 da Parte Oral do Celpe-Bras

| Etapa | Conteúdo da interação | Práticas envolvidas | Tempo |
|--------------|---|---|--|
| 1 | Conversa sobre interesses pessoais do examinando com base nas informações do formulário de inscrição | Compreensão oral e produção oral | 5 minutos |
| 2 | Conversa sobre tópicos do cotidiano e de interesse geral com base em três elementos provocadores ² | Compreensão oral, compreensão escrita e produção oral | 15 minutos (cinco minutos para cada elemento provocador) |

Fonte: Documento-Base do Exame Celpe-Bras (2020, p. 43)

A PO acontece em duas etapas: uma conversa inicial com o examinando sobre os interesses pessoais baseado nas informações do formulário de inscrição do exame e uma conversa sobre tópicos do cotidiano com base nos elementos provocadores escolhidos pelo avaliador-interlocutor, precisando escolher 3 dos elementos provocadores. Durante a interação, o avaliador-interlocutor tem uma grade holística (nota única para todos os critérios

² Os elementos provocadores são materiais utilizados na PO para contextualizar um determinado tema, apresentando informações e/ou pontos de vista com potencial para alimentar a interação sobre um determinado assunto (BRASIL, 2020). Normalmente são textos curtos de diferentes gêneros discursivos que circularam no Brasil.

de avaliação), e o avaliador-observador tem uma grade analítica (diferentes notas para cada um dos critérios de avaliação) para atribuir as suas notas.

A PE, que este trabalho se propõe a estudar, será melhor explicada na próxima subseção.

1.2 A PARTE ESCRITA

Na PE do Celpe-Bras, o examinando dispõe de três horas para responder a quatro tarefas, nas quais são avaliadas as habilidades de compreensão (oral e textual) e produção textual de forma integrada (BRASIL, 2020). A avaliação de habilidades de forma integrada é utilizada pelo Celpe-Bras já que na vida real as habilidades de compreensão e produção acontecem de forma conjunta (SCARAMUCCI; QUITZAU, 2018), o que condiz com o construto do exame de se manter fiel a situações reais de comunicação.

Para responder a essas tarefas, que têm o objetivo de se aproximar de situações concretas de uso da língua portuguesa, o participante precisa se colocar no lugar de um enunciador específico, com um propósito social que leva a uma ação, e deve direcionar seu texto a um ou mais interlocutores especificados no enunciado da tarefa (BRASIL, 2020). Com essas informações, o examinando deve construir uma produção escrita a partir do material de insumo, que depende da habilidade de compreensão avaliada. Neste sentido, a Tarefa 1 terá como material de insumo um vídeo e a Tarefa 2 um áudio, objetivando avaliar a compreensão oral e a produção escrita; nas Tarefas 3 e 4, tem-se como material de insumo textos escritos, avaliando assim a compreensão de leitura e a produção escrita do participante. Os materiais de insumo disponibilizados na prova são materiais de circulação midiática brasileiros, sendo retirados de “jornais, revistas, panfletos, informativos, livros, redes sociais, mídia eletrônica em geral, entre outros” (BRASIL, 2020, p. 58).

Para o exame, conceitua-se tarefa como uma espécie de “convite” para o examinando “usar a língua em diversos contextos, desempenhando papéis com variados propósitos e distintos interlocutores, produzindo textos de uma série de gêneros discursivos, que circulem em diversos suportes” (BRASIL, 2020, p. 31). Cada tarefa terá uma situação específica de comunicação que exigirá do examinando a utilização de diferentes habilidades da língua em uso, o que corrobora a visão de proficiência do exame, explicitada na subseção 1.1. Dessa forma, cada tarefa deverá ser cumprida a partir de seu enunciado, que explicita um evento comunicativo, para o qual o examinando deve ajustar seu texto, a partir das condições de produção, aos propósitos específicos da tarefa, estabelecendo a relação de interlocução. Se os objetivos forem bem explicitados no enunciado da tarefa, o examinando terá uma boa

compreensão do que deve fazer, interpretando corretamente a tarefa (SCARAMUCCI, 1999). No entanto, o desafio das tarefas da PE está em o examinando ter o conhecimento ou não sobre o gênero apresentado para conseguir se comunicar de forma ativa, pois, segundo Bakhtin (2003),

(...) quanto melhor dominamos os gêneros tanto mais livremente os empregamos, tanto mais plena e nitidamente descobrimos neles a nossa individualidade (onde isso é possível e necessário), refletimos de modo mais flexível e sutil a situação singular da comunicação; em suma, realizamos de modo mais acabado o nosso livre projeto de discurso. (BAKHTIN, 2003, p. 284)

Dessa forma, para o examinando atingir um determinado nível de proficiência na língua portuguesa, ele precisa também mostrar seu conhecimento das esferas da atividade humana em que os textos circulam na sociedade brasileira, não sendo suficiente saber apenas aspectos gramaticais da língua.

As condições de produção presentes nos enunciados das tarefas também irão nortear a avaliação dos textos produzidos pelos examinandos, não sendo o foco apenas o uso adequado da gramática e do léxico da língua portuguesa. Schlatter (2009) confirma isso ao dizer que

a adequação das ações produzidas pelo uso de uma língua é (e, nessa perspectiva, deve ser) avaliada através de critérios que constituem e constroem o contexto comunicativo apresentado pela própria tarefa (domínio temático, propósitos de compreensão e de produção textual, posição enunciativa, interlocutor, formato, suporte) (Schlatter et al., 2009, p. 106).

A especificação do gênero discursivo na tarefa estabelece uma expectativa ao ouvinte/leitor em relação ao enunciado que será produzido no momento que um gênero é posto, pois, se em uma das tarefas da PE do exame Celpe-Bras é pedido para escrever uma carta aberta ao prefeito da cidade reivindicando um determinado problema, o leitor não esperará identificar características de um texto de blog, ou de redes sociais, mas sim características específicas do gênero carta aberta que circula na sociedade brasileira.

Na próxima subseção, será apresentada a Tarefa 3 de 2016-2 como exemplo de tarefa do exame Celpe-Bras.

1.2.1 Exemplo de tarefa

A tarefa trazida como exemplo, e que também será a tarefa analisada nesta pesquisa, é a Tarefa 3 da edição 2016-2 do Celpe-Bras. No enunciado da tarefa presente na Figura 3 abaixo, o examinando recebe a informação de que, para cumprir a tarefa, ele deve se colocar

no lugar de **um dono de blog de alimentação saudável**, determinando o enunciador, que, após ler a notícia (texto de insumo) sobre **o cadastro de hortas em São Paulo**, decidiu **implementar esta iniciativa na sua cidade**, determinando o contexto comunicativo. O examinando deve então **escrever um texto de blog para pedir a participação dos leitores no cadastramento das hortas**, precisando **explicar o funcionamento do projeto**, explicitando o gênero discursivo, o interlocutor e o propósito comunicativo.

Figura 3 - Enunciado da Tarefa 3 da Edição de 2016-2

2016/2 **Celpe** **Bras** Certificado de Proficiência em Língua Portuguesa para Estrangeiros

Tarefa 3 | Cadastro de hortas Página 6

Você tem um *blog* sobre alimentação saudável e, depois de ler a notícia sobre o cadastro de hortas em São Paulo, decidiu implementar essa iniciativa em sua cidade. Escreva um texto em seu *blog* para pedir a participação dos leitores no cadastramento, explicando o funcionamento e a utilidade das informações coletadas.

Fonte: Caderno de Questões Celpe-Bras, edição 2016-2. Disponível no Acervo Celpe-Bras

O enunciado da tarefa, presente na Figura 3, indica logo no início que, para concluir a tarefa, o examinando deve ler o material de insumo, que é em formato de texto, a notícia que se chama “Alimentos orgânicos: à procura de hortas em plena São Paulo”, que pode ser observada na Figura 4. A notícia, publicada no site do jornal *Estadão*, apresenta uma iniciativa de cadastro de hortas que está acontecendo em São Paulo. Também apresenta o site *Wikiversidade* onde estão sendo feitos os cadastros, as idealizadoras do projeto e explicita os benefícios que estas hortas podem trazer para a sociedade, ressaltando a importância da iniciativa. Para fechar o texto, incentiva-se a implementação da iniciativa em outras cidades e se explica como é fácil aderir ao cadastramento. Para complementar o texto escrito, há uma fotografia de uma mulher em uma horta com prédios de cidades ao fundo, dando maior veracidade ao que foi falado na reportagem.

Figura 4 - Texto de Insumo da Tarefa 3 da Edição de 2016-2

À procura de hortas em plena São Paulo



Até esta manhã, eram 25 hortas cadastradas, sendo 20 de caráter “educativo” – ou seja, comunitárias e escolares – e 5 “comerciais” – quer dizer, voltadas à produção e venda de alimentos. O que elas têm em comum? Estão todas dentro do município de São Paulo. Várias – olhem só! – em densas áreas urbanas, como nos bairros de Vila Madalena, Vila Mariana, Pompeia, Mooca, Cambuci, Casa Verde, Sapopemba, Vila Nova Esperança, Jardim Ângela...

Já as “comerciais” inserem-se sobretudo na zona rural paulistana. Desde 2014 a cidade de São Paulo “ganhou” uma área rural, ou melhor, porções do território paulistano situadas principalmente nos extremos sul e leste da capital passaram a ser reconhecidas como “rurais”.

Esses dados vêm de uma pesquisa colaborativa em pleno andamento. Quem souber da existência de alguma horta urbana e orgânica sendo cultivada em algum “rincão”

da metrópole pode entrar na plataforma Wikiversidade e inserir os dados ali. Tem de preencher o nome da horta; se ela é educativa ou comercial; a região em que está; o bairro; o endereço e, por fim, o *link* para contato.

O cadastramento dessas hortas urbanas foi iniciado pela dupla Regiane Nigro, que trabalha no Instituto Kairós, em São Paulo, e a jornalista e uma das fundadoras do grupo Hortelões Urbanos, Claudia Visoni. Ambas têm forte atuação em duas dessas hortas comunitárias, a da Praça das Corujas, na Vila Madalena, e também a Horta do Ciclista, na Praça do Ciclista, região da Avenida Paulista. Regiane atua também, pelo Kairós, com os produtores da Associação de Agricultores da Zona Leste, que contempla bairros do extremo leste paulistano, como São Mateus, Guaianases, Cidade Tiradentes e São Miguel Paulista.

A ideia de cadastrar todas as hortas paulistanas – embora a plataforma também

permita a inclusão de cultivos urbanos em todo o país – já era antiga, conta ao blog Regiane Nigro. E foi finalmente posta em prática há cerca de 15 dias, para servir como base de uma reunião que Claudia Visoni teria com pessoas de outros países interessadas no tema hortas urbanas. “Foi este o motivador para o projeto começar a andar”, conta Regiane.

Se inicialmente, há cerca de cinco anos, quando a “onda” do resgate urbano da natureza começou aqui em São Paulo, as hortas cultivadas em plena metrópole eram raras, agora pode-se dizer que elas já têm animado bastante gente disposta a inserir mais verde na selva de pedra. Na maioria dessas hortas, dá-se preferência ao cultivo orgânico.

Regiane conta que a plataforma é importante também para estimular os produtores comerciais orgânicos a incluir suas hortas ali e se tornarem mais conhecidos – afinal, muitos deles fornecem, em sistema de compra coletiva, alimentos para a metrópole, entre outras formas de comercialização.

Atualmente, nos “rincões” paulistanos, um dos grupos mais organizados de agricultores – e que já estão lá no cadastro – é o da Cooperapas, no extremo sul da capital, região de Parelheiros. Outra grande utilidade deste cadastro é atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade. Claudia Visoni, ao chamar, num grupo do Facebook, as pessoas a colaborarem com o cadastramento das hortas, comenta que o registro é “importante para estudos acadêmicos, para os voluntários acharem uma horta perto de casa, para possíveis apoiadores (dos agricultores profissionais) e também para quem não está no Facebook”. E continua: “Quem puder, ajude a incluir as hortas que ainda não estão na tabela. Quem quiser criar tabelas para outras cidades, é só fazer. A plataforma Wiki é totalmente colaborativa e fácil de editar”.

Disponível em: <http://vida-estilo.estadao.com.br/blogs/alimentos-organicos/a-procura-de-hortas-em-plena-sao-paulo/>. Acesso em: 5 ago. 2016 (adaptado).

Fonte: Caderno de Questões Celpe-Bras, edição 2016-2. Disponível no Acervo Celpe-Bras

Essa tarefa, conforme Mendel (2017), é autêntica, pois “a produção solicitada ao examinando configura-se como uma resposta adequada às possibilidades de ação apontadas no texto de insumo” (p. 45), visto que no texto de insumo há a informação de que hortas de outras cidades também podem ser cadastradas na plataforma, e o enunciado corrobora com a mesma ideia ao solicitar que o examinando incentive seus leitores a participarem do cadastro de hortas na sua cidade. Esta é “uma proposta de produção escrita que consiste em uma resposta autêntica para o texto de insumo” (MENDEL, 2017, p. 45), aproximando a tarefa de uma situação real de uso da língua.

Scaramucci et al (2019), em sua pesquisa, pontuam que esta tarefa de 2016-2 é uma tarefa adequada, coerente e representativa do construto do exame, pois há de forma clara no material de insumo da tarefa a integração de habilidades (escrita e leitura) como uma exigência para o cumprimento da tarefa, ou seja, o examinando precisa ler o texto de insumo, entender e selecionar as informações que irão auxiliar no cumprimento da tarefa; a linguagem utilizada no enunciado é muito clara e sem ambiguidades, além de que explicita a situação de produção de forma assertiva, sem deixar dúvidas.

Em seu estudo para descrever as tarefas da PE do exame Celpe-Bras nas edições de 1998 a 2017, Schoffen et al (2018), classificaram esta tarefa de 2016-2 dentro da temática “políticas e cidadania”, que engloba “participação e patrocínio para projetos comunitários, voluntariado” (p. 20), além de se adequar à esfera de atuação “comunitária”, pois essa esfera é composta “por interações que envolvem atividades entre pessoas ou organizações que compartilham valores e objetivos comuns” (SCHOFFEN, et al., 2018, p. 15), e é isso que acontece no blog de alimentos saudáveis do qual o enunciador é dono. Também foi incluído no propósito de “incentivar” os leitores, que são o grupo social do enunciador, com o gênero do discurso sendo o “artigo” no suporte “blog” que pede uma construção de interlocução “pública, assimétrica descendente”, ou seja, são interações constituídas no âmbito social por meio de um texto publicado em suporte público, “escrito para interlocutor que está em posição inferior na hierarquia social” (SCHOFFEN, et al., 2018, p. 31).

Em sua pesquisa, Mendel (2017) apresenta a resposta esperada pela Tarefa 3 de 2016-2, que segue abaixo:

Figura 5 - Resposta Esperada na Tarefa 3 da Edição de 2016-2

| | |
|-------------------------|--|
| Enunciador | blogueiro |
| Gênero discursivo | texto para blog sobre alimentação saudável |
| Propósito | pedir a participação dos leitores no cadastramento de hortas urbanas, explicando o funcionamento e a utilidade das informações coletadas |
| Interlocutor | leitores do blog |
| Informações solicitadas | funcionamento do cadastramento de hortas urbanas (terceiro parágrafo do texto de insumo): <i>Quem souber da existência de alguma horta urbana e orgânica sendo cultivada em algum “rincão” da metrópole pode entrar na plataforma Wikiversidade e inserir os dados ali. Tem de preencher o nome da horta; se ela é educativa ou comercial; a região em que está; o bairro; o endereço e, por fim, o link para contato;</i> utilidade das informações coletadas no cadastramento (dois últimos parágrafos do texto de insumo): <i>a plataforma é importante também para estimular os produtores comerciais orgânicos a incluírem suas hortas ali e se tornarem mais conhecidos / Outra grande utilidade deste cadastro é atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade / o registro é “importante para estudos acadêmicos, para os voluntários acharem uma horta perto de casa, para possíveis apoiadores (dos agricultores profissionais) e também para quem não está no Facebook</i> |

Fonte: Mendel (2017, p. 44)

Além de esperar que o examinando se coloque no lugar de um blogueiro e escreva seu texto em um blog sobre alimentação saudável para os leitores deste mesmo blog, é esperado que o examinando mobilize e recontextualize algumas informações do texto de insumo. Para Mendel (2017), o gênero discursivo da tarefa, artigo de blog, possibilita que o examinando mobilize certos trechos do texto com “o propósito de incentivar os leitores a participarem do cadastramento de hortas urbanas” (MENDEL, 2017, p. 44) e também, como um bom blogueiro, “demonstrar que essa prática está ‘na moda’” (MENDEL, 2017, p. 45).

Uma hipótese seria de que, de acordo com Mendel (2017), ao analisar o texto de insumo e a proposta da tarefa, pode-se prever que haverá demasiada cópia por parte dos examinandos, principalmente daqueles com níveis de proficiência mais baixos, pois tanto o gênero do discurso usado no texto de insumo como o que o examinando deve produzir são semelhantes em alguns aspectos, principalmente por serem ambos da esfera de alimentos saudáveis. No entanto, Mendel (2017) afirma que a cópia por si só do texto ou o resumo do material não seriam suficientes para cumprir com o propósito da tarefa, já que a notícia tem o objetivo de *divulgar* a iniciativa do cadastramento de hortas urbanas, e o texto que deve ser produzido pelo examinando tem o propósito comunicativo de *incentivar* os leitores do blog a participarem do cadastramento de hortas (SCHOFFEN, et al., 2018), sendo necessário o

examinando escolher as informações mais relevantes para cumprir a tarefa, como o funcionamento do cadastro de hortas e a utilidade do projeto.

Mendel (2017) também discute sobre a avaliação integrada nas tarefas do exame, e verifica que na Tarefa 3 de 2016-2 é solicitado ao examinando produzir um texto sobre um assunto bastante específico, pois dificilmente um examinando saberia escrever sobre o cadastro de hortas sem ler previamente nenhum material de apoio. Portanto, nesta tarefa, Mendel (2017) confirma que “a compreensão do material de insumo demonstra-se essencial para a realização do propósito comunicativo” (MENDEL, 2017, p. 46), o que vai ao encontro do entendimento sobre a integração de habilidades.

Outra hipótese que Mendel (2017) aponta para esta tarefa é a de que o seu enunciado apresenta múltiplas possibilidades de interpretação, já que nele é explicitado que o enunciador do texto quer implementar a iniciativa na sua cidade, mas não explicita qual cidade seria. Isso deixa aberto para o examinando decidir o local da implementação, podendo ser em outros países, como também para decidir utilizar a própria São Paulo, que daí já não seria mais um contexto de implementação, mas sim um objetivo de expandir a iniciativa (MENDEL, 2017). Segundo Mendel (2017, p. 46), “[ao] não especificar em qual cidade o examinando se encontra e decide implementar a iniciativa, portanto, a tarefa abre espaço para outras leituras”. Isso abre espaço para que os examinandos lancem mão de diversos mecanismos para construir a relação de interlocução e cumprir com o propósito da tarefa.

A pesquisa de Scaramucci et al (2019) analisou 80 textos de nota 5 e levantou a hipótese que, nesta tarefa, “o gênero e a interlocução parecem prevalecer sobre a integração de habilidades e o cumprimento do propósito” (SCARAMUCCI, et al., 2019, p. 402). Essa afirmação deriva do fato de que se percebeu na análise dos textos que alguns examinandos não selecionaram e articularam tantas informações chaves do texto como era esperado, mas mesmo assim foram avaliados com nota 5 pela sua forte relação de interlocução e adequação ao gênero pedido.

Ao analisar a recontextualização de informações do texto de insumo nas produções escritas dos examinandos, essa pesquisa (SCARAMUCCI, et al., 2019) também trouxe a problemática de que, ao analisar os textos de nota 5, percebeu-se que os avaliadores deixaram algumas coisas que pareciam importantes para o cumprimento da tarefa de lado: “Os examinandos lidaram de maneiras diferentes com a notícia lida e a iniciativa já existente em São Paulo. Muitos mencionaram a leitura da notícia como a fonte da ideia para fazer a mesma coisa em sua cidade”, o que era o esperado nesta tarefa, porém “outros contaram como funciona o cadastro em São Paulo para só depois propor o novo cadastro”, o que não seria tão

distante de um dos propósitos da tarefa, mas alguns examinandos “ignoraram completamente a notícia e o cadastro já existente, como se a ideia toda tivesse partido do próprio autor do blog.” (SCARAMUCCI, et al, 2019, p. 403). Apesar dessas diferenças, todos receberam nota 5, comprovando que não há especificações suficientes sobre a avaliação de integração de habilidades (SCARAMUCCI, et al, 2019), e isso comprova a hipótese de Scaramucci et al (2019) anteriormente apresentada de que há um peso maior na avaliação para as produções que constroem uma boa relação de interlocução e trazem características pertencentes ao gênero escolhido para a tarefa.

Na pesquisa de Sirianni (2020), que analisou textos que receberam nota 1 e nota 2 nas quatro tarefas da edição 2016-2, a autora afirma que os textos de nota 2 têm problemas para compreender bem a tarefa, porém conseguem selecionar bem as informações do material de insumo para cumprir os propósitos solicitados, apesar de haver bastante cópia nos textos. Outra característica apresentada por Sirianni (2020) é de que os textos de nota 2 contam com um número considerável de inadequações ortográficas, principalmente relacionadas a acentos inadequados e falta de pontuação para dar coerência aos textos. Além disso, apesar de, por vezes, fazerem isso de forma inconsistente, os textos de nota 2 tentam construir uma relação de interlocução e tentam se adequar ao gênero proposto, conseguindo realizar esses propósitos de forma parcial.

1.2.2 Avaliação da Parte Escrita

A avaliação da PE é desenvolvida por uma equipe de professores(as) e pesquisadores(as) da área de Português como Língua Adicional (PLA). Nesta avaliação, os textos são desidentificados e digitalizados para manter a confiabilidade do exame (BRASIL, 2020). Assim, cada tarefa é avaliada de forma anônima por dois avaliadores de forma independente, que utilizam uma grade de avaliação holística em conjunto com uma resposta esperada da tarefa (a exemplo da resposta esperada da Tarefa 3 da edição 2016-2, apresentada anteriormente), que explicita aspectos particulares de cada tarefa da edição, atribuindo uma nota para cada texto em uma escala de 0 (zero) a 5 (cinco) pontos (BRASIL, 2020, p. 72-73).

Segundo o Documento Base do exame Celpe-Bras (2020, p. 72-73), se as notas convergirem, a nota final da tarefa é o resultado da média aritmética entre as notas dos dois avaliadores. Se houver discrepância (diferirem em 2 (dois) pontos ou mais), a tarefa precisará ser reavaliada por um terceiro avaliador e, caso essa nota convirja com a de um dos primeiros avaliadores, a nota final da tarefa será a média entre as duas notas convergentes, excluindo-se a outra divergente; caso a nota do terceiro avaliador seja igualmente distante das demais notas

atribuídas, prevalecerá a nota desse terceiro avaliador. Para a nota final da PE, calcula-se a média aritmética entre as notas finais das quatro tarefas.

1.2.3 Parâmetros de Avaliação da Parte Escrita

Com o intuito de avaliar os textos segundo a perspectiva de proficiência adotada pelo Celpe-Bras, que implica a compreensão de que cada texto é único e irrepetível e, dessa forma, precisa ser avaliado em sua singularidade (SCHOFFEN, 2009; BRASIL, 2020), o Celpe-Bras opta pelo uso de uma grade de avaliação holística, destinando uma única nota para cada texto. Esta grade holística é um parâmetro de avaliação comum, uma base, já que para cada tarefa existe uma resposta esperada (como a da Figura 5) que também contém observações específicas para a correção (SCARAMUCCI, et al, 2019), que precisam ser levadas em consideração na hora da avaliação.

Conforme a proposta de Schoffen (2009), os parâmetros de avaliação integram os eixos chamados Adequação Contextual e Adequação Linguístico-Discursiva, sendo utilizados de forma conjunta na avaliação (BRASIL, 2020) e também avaliados nos textos dos examinandos de forma integrada. Cada nota tem a sua descrição de competências atingidas pelo examinando, como também corresponde a um nível de proficiência. O Quadro 2 apresenta os parâmetros de avaliação utilizados na PE do Celpe-Bras:

Quadro 2 - Parâmetros de Avaliação da Parte Escrita

| Nota | Descrição das competências atingidas pelos examinandos em cada nível de proficiência na Parte Escrita |
|------|---|
| 5 | Configura adequadamente a relação de interlocução no gênero discursivo proposto na tarefa, realizando a ação solicitada. Recontextualiza apropriadamente e de maneira autoral as informações necessárias para cumprir o propósito interlocutivo de forma consistente. Eventuais inadequações ou equívocos não comprometem a configuração da interlocução. Produz um texto autônomo, claro e coeso, em que os recursos linguísticos acionados são apropriados para configurar a relação de interlocução no gênero solicitado, e possíveis inadequações raramente comprometem a fluidez da leitura. |
| 4 | Configura a relação de interlocução no gênero discursivo proposto na tarefa, realizando a ação solicitada. Recontextualiza apropriadamente as informações necessárias para cumprir o propósito interlocutivo, mas possíveis equívocos ou incompletudes podem fragilizar, em momentos localizados, a consistência da interlocução. Os recursos linguísticos acionados são apropriados para configurar a relação de interlocução no gênero proposto, construindo um texto claro e coeso em que possíveis inadequações podem comprometer, em momentos localizados, a fluidez na leitura. |
| 3 | Configura a relação de interlocução no gênero discursivo proposto na tarefa, realizando a ação solicitada, ainda que a consistência da relação de interlocução possua algumas falhas. Pode recontextualizar de forma pouco articulada e/ou equivocada ou não recontextualizar informações necessárias para cumprir o propósito dentro do contexto de produção solicitado. Os recursos |

| | |
|---|--|
| | linguísticos acionados são apropriados, podendo apresentar limitações ou inadequações que podem prejudicar, em alguns momentos, a configuração da interlocução no gênero proposto. Problemas de clareza e coesão podem ocasionar, em alguns momentos, dificuldades na leitura. |
| 2 | Configura a relação de interlocução de forma pouco consistente, realizando superficialmente a ação solicitada. Pode estabelecer uma relação de interlocução próxima à solicitada, não cumprir propósito(s) menor(es) e/ou apresentar problemas na construção do gênero. Pode apresentar trechos do texto que remetem a um gênero diferente, comprometendo a relação de interlocução. A relação entre o propósito do texto e a interlocução configurada não é clara ou não é totalmente adequada. Pode não recontextualizar informações que seriam necessárias para a configuração adequada da interlocução ou não articular claramente essas informações. Equívocos de compreensão podem comprometer parcialmente o cumprimento do propósito. Os recursos linguísticos acionados são limitados e/ou inadequados, podendo prejudicar parcialmente a configuração da relação de interlocução no gênero solicitado. Problemas de clareza e coesão podem ocasionar, em diferentes momentos, dificuldades na leitura. |
| 1 | Configura com problemas recorrentes ou não configura a relação de interlocução solicitada, realizando muito superficialmente ou não realizando a ação solicitada. Remete-se ao tema, mas pode não considerar o contexto de produção e não construir o gênero discursivo proposto ou apresentar problemas recorrentes na sua construção. Não recontextualiza informações suficientes para o cumprimento do propósito comunicativo, considerando a relação de interlocução configurada. OU Pode apresentar equívocos graves e/ou frequentes de compreensão que comprometem o cumprimento do propósito. Os recursos linguísticos acionados são muito limitados e/ou inadequados, o que prejudica substancialmente o cumprimento do propósito e a configuração da relação de interlocução, comprometendo a construção do gênero solicitado. Problemas frequentes de clareza e coesão ocasionam, em vários momentos, problemas na leitura. |
| 0 | Não configura, ou configura de forma equivocada, a relação de interlocução, não realizando a ação solicitada. OU Trata de outro tema. OU Demonstra problemas generalizados de compreensão, impedindo o cumprimento do propósito e a configuração da relação de interlocução E/OU Limita-se a reproduzir o(s) texto(s)-base(s), sem marcas de autoria. OU Ignora completamente os texto(s)-base(s). E/OU Problemas generalizados de clareza e coesão e/ou inadequações linguísticas impedem a configuração da relação de interlocução no gênero solicitado, comprometendo a compreensão geral do texto. OU A produção é insuficiente para a avaliação. |

Fonte: Documento-Base do Exame Celpe-Bras (2020, p. 39)

Os parâmetros de avaliação da parte escrita informam as notas possíveis que o examinando pode receber na PE e também a descrição de cada nota. A nota 5, por exemplo, é descrita como a nota mais alta que se pode alcançar, porém, conforme o construto do exame, não almeja um falante nativo “ideal”, mas deixa claro que mesmo recebendo essa nota, o examinando pode produzir **eventuais equívocos ou inadequações**, mas que **não comprometem a construção da relação de interlocução** e nem a **fluidez da leitura**. Além disso, informa que, nos textos de nota 5, o examinando realiza a tarefa proposta, **recontextualizando** as informações de **maneira autoral**, cumprindo o **propósito interlocutivo**, e também produzindo **um texto autônomo, claro e coeso**, utilizando apropriadamente os **recursos linguísticos**.

Já na nota 2, o examinando demonstra alguns problemas na construção do texto, construindo de **forma pouco consistente a relação de interlocução** e **realizando de forma**

superficial a ação solicitada na tarefa. O texto não cumpre com todos os propósitos da tarefa, apresentando **problemas na construção do gênero do discurso**. A descrição da nota 2 explicita também que os textos podem apresentar características de **gêneros diferentes**, como também **não contextualizar informações**, ou não articulá-las com clareza, que são necessárias para cumprir com a tarefa. Os **recursos linguísticos** dos textos de nota 2 **são limitados ou inadequados**, e traz também **problemas de clareza e coesão** no texto, dificultando a leitura e compreensão.

Com essas descrições, pode-se ter uma noção maior das diferenças que uma nota tem da outra, auxiliando o avaliador a atribuir a nota correspondente ao texto produzido por cada examinando. Após apresentar o construto e funcionamento do Celpe-Bras, na seção seguinte serão apresentados alguns pressupostos teóricos importantes para a pesquisa.

2 PRESSUPOSTOS TEÓRICOS

Após apresentar o Celpe-Bras, explicitando como o exame funciona, sua noção de proficiência, como a PE se articula e os parâmetros de avaliação da PE, nesta seção serão apresentados alguns pressupostos teóricos importantes para as análises que serão realizadas neste trabalho, sendo eles: Linguística de Corpus, as relações entre a LC e a área de avaliação de proficiência e estudos anteriormente produzidos sobre o Celpe-Bras à luz da LC.

2.1 LINGUÍSTICA DE CORPUS

Segundo Sardinha (2000), a LC 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.” (SARDINHA, 2000, p. 325), como também dedica-se “à exploração da linguagem através de evidências empíricas” (SARDINHA, 2000, p. 325). Além das pesquisas que focam no estudo das descrições a respeito do léxico e dos aspectos gramaticais, a LC tem interesse e pode ser utilizada como ferramenta de pesquisa em áreas diversas, como ensino, aprendizado e avaliação de línguas, análise do discurso, estilística literária, linguística forense, entre outras (MCCARTHY; O’KEEFE, 2010 apud DIVINO, 2024).

Um corpus pode ser definido como

Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SANCHEZ, 1995, pp. 8-9 apud SARDINHA, 2000, p. 338)

Segundo Sardinha (2000), este conceito é completo, pois deixa claro alguns aspectos que são essenciais para um corpus: a origem (dados autênticos), o propósito (finalidade de ser parte de um estudo linguístico), a composição (conteúdo escolhido de forma criteriosa), a formatação (dados legíveis por computador), a representatividade (de uma língua ou variedade) e a extensão (precisa ser vasto para ser representativo).

Com isso posto, Sardinha (2000) afirma que existem três corpora eletrônicos, todos da língua inglesa, que são um marco na história da LC por sua influência nos estudos: Corpus Brown (1964), Birmingham Corpus (1987) e BNC (1995). Segundo Sardinha (2000), o Corpus Brown foi o início das pesquisas de corpora em 1964 com 1 milhão de palavras no

corpus do inglês americano escrito; já o Birmingham Corpus, de 1987, deu continuidade ao Corpus Brown e ultrapassou a marca de 1 milhão de palavras do inglês britânico; e o BNC leva a marca, em 1995, de ter sido o primeiro a conter 100 milhões de palavras do inglês britânico escrito e falado.

Na LC, existem duas abordagens principais: a pesquisa *baseada em corpus* (*corpus-based research*) e a pesquisa *guiada por corpus* (*corpus-driven research*). A pesquisa baseada em corpus é utilizada para confirmar ou negar hipóteses criadas previamente pelos pesquisadores, podendo comparar produções escritas em níveis de proficiência em língua adicional diferentes de textos produzidos por falantes nativos (CALLIES; DÍEZ-BEDMAR; ZAYTSEVA, 2014 apud DIVINO, 2024). A pesquisa guiada por corpus, diferentemente da anterior, não cria hipóteses anteriores às análises, e é *guiada* a partir do que os testes no *software* escolhido irão mostrar a respeito dos níveis de proficiência (CALLIES; DÍEZ-BEDMAR; ZAYTSEVA, 2014 apud DIVINO, 2024). Neste trabalho, a pesquisa foi *guiada por corpus*, pois não foram criadas hipóteses anteriores, mas as análises foram guiadas a partir dos dados achados nos testes feitos no *software Sketch Engine*, sempre atenta ao que poderia aparecer de relevante nas análises. A partir disso, hipóteses foram criadas e respondidas a partir dos dados obtidos.

Outro aspecto importante é que, apesar de as pesquisas que utilizam da LC serem de natureza quantitativa, também há a possibilidade de se utilizar, em conjunto, uma abordagem qualitativa. Nesse caso, os dados obtidos de forma quantitativa são contextualizados, possibilitando o reconhecimento de padrões do uso da língua entre textos de um mesmo corpus. Estudos que unem essas duas abordagens na LC são capazes de aumentar as análises, observando aspectos lexicais, gramaticais e discursivos diversos (KENNEDY, 2001). Essas duas abordagens foram utilizadas em conjunto neste trabalho, resultando em análises complexas e contextualizadas.

2.1.1 Linguística de Corpus e avaliação de língua

Segundo Cushing (2022), a descrição da aprendizagem de língua de diferentes níveis de proficiência é algo muito promissor na LC. Desde o final do século XX, vê-se um maior interesse em utilizar a Linguística de Corpus no campo de avaliação de língua. Barker (2010) confirma os benefícios dessa união ao dizer que um corpus de língua adicional pode trazer contribuições diversas para o exame, como auxiliar na definição do construto do exame, validar o instrumento avaliativo utilizado, descrever os níveis de proficiência e possibilitar

avaliações mais automatizadas, juntamente com ferramentas para fornecer *feedback* (CUSHING, 2022; DIVINO, 2024).

Gablasova (2020) afirma que o uso de LC na avaliação de língua adicional é muito útil, pois pode-se observar características linguísticas mais padronizadas e outras mais únicas de domínio da língua fornecidas por um grupo específico de falantes. Além de que, pesquisas sobre níveis de proficiência utilizando corpora são importantes para a validade do exame, sendo validade a relação entre o que o teste pretende avaliar e o que de fato avalia (SCHLATTER et al, 2005), confirmando este aspecto ou não. Esse tipo de pesquisa também pode refinar os descritores de proficiência, o que pode contribuir para que os elaboradores de exames de proficiência em língua adotem abordagens da LC para pensarem as avaliações (CALLIES; GÖTZ, 2015; DIVINO, 2021).

As contribuições da Linguística de Corpus em exames de avaliação de língua, principalmente de língua adicional, são diversas e, por isso, na próxima subseção, serão apresentados alguns estudos que utilizaram a abordagem da LC para pesquisas de avaliação de língua adicional, sendo o foco no exame Celpe-Bras.

2.2 ESTUDOS SOBRE O CELPE-BRAS À LUZ DA LINGUÍSTICA DE CORPUS

Desde o início do século, vêm sendo feitas pesquisas focadas na descrição dos níveis de proficiência do exame Celpe-Bras utilizando as produções escritas dos examinandos ou alunos em cursos preparatórios para a prova. Pode-se citar Sidi, 2002; Schoffen, 2009; Gomes, 2009; Evers, 2013; Dell’Isola; Tosatti, 2015; Sirianni, 2016; 2020; Kunrath, 2019; Mendel, 2019; Divino, 2021; Tosatti, 2021; Vicentini, 2022; Hanauer, 2023; Schoffen; Divino, 2023; Sostruznik, 2023; e Divino, 2024. Em sua maioria, são pesquisas com análises qualitativas com focos diferentes, como análise dos parâmetros de avaliação (contribuindo para a sua melhor descrição) e de mecanismos utilizados para construir os gêneros do discurso como também a busca pela presença da autoria nas produções escritas dos examinandos (MENDEL, 2019).

As pesquisas quantitativas com o uso da LC para as análises têm sido mais recentes, mas já apresentam resultados interessantes para o exame na descrição dos níveis de proficiência, sendo utilizados os *softwares* Coh-Metrix-Port e o Sketch Engine. Esses trabalhos aprofundaram as discussões acerca da importância da recuperação do conteúdo informacional (KUNRATH, 2019; DIVINO, 2021; 2024; HANAUER, 2023), da autoria nos textos dos examinandos a partir de recursos linguísticos utilizados que não estão presentes no material de insumo (DIVINO, 2021; 2024; HANAUER, 2023), da extensão dos textos

(EVERS, 2013; DIVINO, 2021; 2024; HANAUER, 2023), da relação entre o aumento no nível de proficiência com a utilização de conjunções (SOSTRUZNIK, 2023), dos recursos utilizados para dar coesão na produção textual (KUNRATH, 2019) e da presença de cópia do material de insumo nos textos dos examinandos, independentemente do nível de proficiência (DIVINO, 2024).

Com este panorama geral posto, este trabalho também tem o objetivo de contribuir para as discussões já feitas, auxiliando na descrição dos níveis de proficiência e fortalecendo mais os laços entre avaliação de língua e Linguística de Corpus.

3 METODOLOGIA

Nesta seção de metodologia, serão apresentados os objetivos deste trabalho, como também as perguntas de pesquisa, o Corpus do Celpe-Bras (CorCel), o *software* utilizado para a análise dos dados (*Sketch Engine*) e, por fim, serão detalhados os procedimentos utilizados para estas análises.

3.1 OBJETIVOS DE PESQUISA

Este trabalho busca verificar as características dos textos avaliados com nota 2 e nota 5 na Tarefa 3 de 2016-2 do Celpe-Bras, a fim de contribuir para uma descrição mais detalhada dos níveis de proficiência avaliados na Parte Escrita do exame. Com base na literatura da área e em pesquisas anteriormente produzidas, principalmente Mendel (2017), e nas análises empreendidas neste trabalho, busca-se responder à seguinte pergunta de pesquisa e suas subdivisões:

- 1. Quais padrões lexicais caracterizam os textos avaliados com nota 2 e com nota 5 na Tarefa 3 da edição 2016-2 do Celpe-Bras?**
 - a) Como a extensão e os mecanismos lexicais utilizados nos textos de nota 2 e nota 5 da Tarefa 3 da edição 2016-2 do Celpe-Bras são importantes para a caracterização dos níveis de proficiência?
 - b) Como os textos avaliados com as notas 2 e 5 na Tarefa 3 da edição 2016-2 do Celpe-Bras recuperam as informações do material de insumo?

Com as análises e reflexões realizadas neste trabalho, espera-se contribuir para os estudos do Celpe-Bras a partir da Linguística de Corpus e também para a descrição dos níveis de proficiência certificados pelo exame. O corpus de pesquisa e os procedimentos de análise serão apresentados nas subseções a seguir.

3.2 CORCEL: O CORPUS DO CELPE-BRAS

De acordo com Kennedy (2001), a compilação de um corpus é feita a partir de uma compilação de textos, descrevendo e analisando aspectos de uma determinada língua, e isso pode ser realizado a partir de produções escritas de exames de proficiência. Segundo Divino (2024), existem alguns projetos de compilação de corpora de textos escritos por examinandos ou alunos de cursos preparatórios de língua portuguesa, como o Corpus de Aquisição de L2 (CAL2) (CLUL, 2012), composto por 1.607 textos, totalizando 314.817 palavras; o Corpus de Português Língua Estrangeira/Língua Segunda (COPLE2) (MENDES et al., 2016),

contendo 966 textos; e o Corpus de Produções Escritas de Aprendentes de PL2 (PAEPL2) (MARTINS et al., 2019), contendo textos de 458 alunos de diferentes níveis de proficiência.

Com o objetivo de compilar um corpus com 15.315 produções textuais de examinandos das edições 2015-1, 2016-1, 2016-2 e 2017-1, o Corpus do Celpe-Bras (CorCel) (SCHOFFEN, et al., em preparação) é um projeto em andamento que está sendo desenvolvido pelo Grupo Avalia, atuante na Universidade Federal do Rio Grande do Sul (UFRGS). Em cada uma dessas edições há quatro tarefas e seis subcorpora de textos que receberam as notas de 0 a 5 pelos avaliadores. Cada subcorpora contém até 200 textos que foram selecionados aleatoriamente no site Invertexto³ dentre os textos que receberam a mesma nota, e depois foram digitados e revisados. Quando uma determinada nota de alguma tarefa não chegava ao número esperado de 200 textos, incluiu-se todos os textos avaliados naquela nota, o que ocasionou que alguns subcorpora contenham um número menor de textos comparado a outros.

Os textos originais foram disponibilizados em cópia digitalizada pelo Inep, juntamente com planilhas de Excel informando as notas que cada produção escrita recebeu de cada um dos avaliadores da tarefa, sem nenhuma informação sobre quem eram os avaliadores e nem os examinandos. Como não se tem a informação de quem são os examinandos que produziram os textos, conforme Divino (2024), o termo “corpus de aprendizes” não pode ser utilizado neste caso, já que, apesar de não serem indivíduos que têm o português como primeira língua, “não é possível saber se esses examinandos são estudantes de português ou não, pois o fato de não serem falantes nativos não os torna automaticamente aprendizes, apenas falantes de uma [língua adicional]” (DIVINO, 2024, p. 78).

O corpus desta pesquisa é composto por textos produzidos por examinandos na Tarefa 3 da edição de 2016-2 do exame Celpe-Bras, e faz parte do CorCel (Schoffen, et al., em preparação). Os subcorpora utilizados nesta pesquisa foram nomeados de corcel20162t3n5 e corcel20162t3n2, sendo “t” referente à tarefa à qual pertencem os textos e “n” à nota atribuída.

O critério para a escolha desta tarefa foi motivado pela intenção de dialogar com estudos já feitos anteriormente sobre esta tarefa (MENDEL, 2017), a partir da LC. O corpus é composto por 200 textos nota 5 e 200 textos nota 2 que atingiram estas notas como média final da tarefa, independente das notas atribuídas de forma individual pelos avaliadores.

³ Link: <https://www.invertexto.com/lista-aleatoria>

No protocolo de digitação, as assinaturas deveriam ser substituídas pelo nome Bibiana e sobrenome, caso tivesse, Cambará, ambos em homenagem à personagem feminina do romance *O Tempo e o Vento*, de Érico Veríssimo. Essa escolha foi feita para manter a anonimização dos participantes, e também como uma forma de auxiliar em análises futuras que focassem no número de assinaturas feitas pelos examinandos em gêneros do discurso que pedem esse uso, já que essa decisão de padronizar o nome permite que se saiba todos os textos que utilizaram uma assinatura.

Os textos utilizados neste trabalho não sofreram nenhum tipo de normalização da ortografia, pois o corpus ainda está em processo de digitação e revisão, e pretende-se fazer algumas anotações nestes textos para auxiliar em análises futuras. A não normalização da ortografia pode se mostrar um empecilho em algumas análises e não dar tanta confiabilidade aos dados. Na pesquisa de Hanauer (2023), ao fazer o cálculo do *Type-Token-Ratio* (TTR), os textos de nota 2 se mostraram com uma riqueza lexical maior do que os de nota 5, o que era o oposto do esperado. Isso pode ter acontecido pelo fato de os textos utilizados na pesquisa não terem sido normalizados ortograficamente, fazendo com que os textos de nota 2 tivessem mais palavras diferentes (*types*) do que os de nota 5, já que cada palavra escrita diferente, por mais semelhante que seja, é contada de forma separada pelo *Sketch Engine*. Hanauer (2023) e Divino (2021; 2024) afirmam que, ao utilizar a ferramenta *keywords* (que busca as palavras mais típicas de um corpus) do *Sketch Engine*, aparecem muitas inadequações ortográficas nos textos de nota 2, o que dificultou as análises ao comparar as palavras mais relevantes do corpus, as *keywords*, com as palavras do material de insumo. Conforme Hanauer (2023), as inadequações ortográficas também podem interferir no valor do *log-likelihood*, pois ele verifica em qual corpus uma palavra é mais utilizada de forma significativa e, como nos textos de nota 2 os examinandos podem ter escrito uma mesma palavra de formas diferentes e o *Sketch Engine*, como explicado anteriormente, conta cada palavra escrita de uma forma como uma palavra diferente, acaba que o valor do *log-likelihood* sempre tende a apontar mais significância nos textos de nota 5. Fazer a normalização ortográfica dos textos, segundo Divino (2024), auxiliaria para que as análises fossem mais precisas e também confiáveis, o que é um objetivo futuro do Grupo Avalia para o CorCel.

As pesquisas de Divino (2021; 2024), Hanauer (2023) e Sostruznik (2023) utilizaram textos do CorCel para suas análises. Este estudo também utilizou o Corpus do Celpe-Bras para suas análises.

3.3 PROCEDIMENTOS PARA AS ANÁLISES

Para as análises foi utilizado o *Sketch Engine* (SkE) e suas ferramentas: *keywords*, *wordlist*, *n-grams* e *concordance*. Além disso, também foi utilizado o cálculo *Type-Token-Ratio* (TTR). Nesta subseção, será explicado e detalhado os procedimentos adotados em cada etapa das análises com as diferentes ferramentas escolhidas.

3.3.1 O Sketch Engine

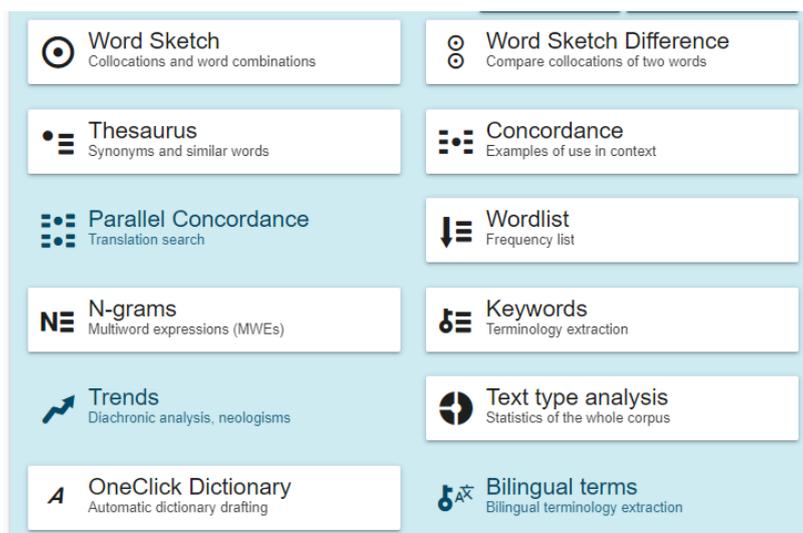
Em consonância com Divino (2021; 2024), Hanauer (2023) e Sostruznik (2023), o *software* utilizado para as análises foi o *Sketch Engine*⁴, uma ferramenta utilizada para análise de corpora (KILGARRIFF et al., 2004). Com o intuito de fornecer amostras autênticas dos idiomas, o programa conta com 800 corpora em mais de 100 idiomas, incluindo o Português. Além disso, o usuário do *software* pode criar o seu próprio corpus adicionando seus próprios arquivos no programa ou então utilizando arquivos da internet. Para utilizar o programa, pode-se utilizar a versão gratuita ou então adquirir uma licença para utilizar outros recursos pagos, sendo a versão paga a utilizada neste trabalho⁵.

Ao fazer o cadastro e entrar no programa, o usuário pode escolher utilizar um corpus já existente do programa ou então fazer upload do seu próprio corpus em formato *zip*. Dentro do *Sketch Engine*, após escolher um corpus, o usuário pode utilizar alguns recursos, como o *word sketch*, que gera um perfil lexical de uma palavras ao mostrar termos que mais frequentemente co-ocorrem com ela; o *thesaurus*, que permite encontrar sinônimos e antônimos de uma palavra, como também palavras relacionadas semanticamente; o *concordance*, que mostra ocorrências de uma palavras ou frase em contexto; a *wordlist*, que gera uma lista de palavras mais frequentes no corpus analisado; a *keywords*, que gera uma lista com as palavras mais típicas de um corpus; e o *n-grams*, que é a sequência contígua de 2 a 6 palavras, gerando uma lista com as sequências de maior frequência. Na Figura 6, pode-se observar o painel (*dashboard*) do programa com as funcionalidades disponíveis.

⁴ Disponível em: <https://www.sketchengine.eu>

⁵ A versão paga utilizada neste trabalho foi fornecida pelo PPGLetras-UFRGS.

Figura 6 - Dashboard do Sketch Engine



Fonte: Sketch Engine

Para esta pesquisa, foram utilizados os recursos que se mostraram mais relevantes para responder a pergunta de pesquisa. São eles: o *keywords*, *n-grams*, *wordlist* e *concordance*, que serão apresentados a seguir.

3.3.1.1 Keywords

A ferramenta *keywords* pode ser utilizada para comparar dois corpora, documentos ou textos, sendo um o corpus de referência, que é utilizado para comparação, e um o corpus de estudo/foco, que é o corpus cujas palavras-chave se deseja explorar, para identificar o que é diferente no corpus de estudo comparado com o corpus de referência. Na lista de *keywords* são elencados não necessariamente os termos que mais aparecem no corpus de estudo, mas sim palavras que são chaves, ou seja, típicas do corpus de estudo em relação ao corpus de referência. O resultado desse recurso pode ser dividido em até três resultados: itens de uma única palavra (*single-words*), itens de várias palavras (*multi-word terms*) e itens de multipalavras-chave (*n-grams*), podendo estes últimos serem de 2 a 6 *tokens* (palavras) em sequência.

Para realizar as análises no programa, precisou-se escolher um corpus de referência para comparar com o corpus de estudo, e o escolhido foi o Portuguese Trends, que é um corpus de português disponível no SkE que contém 374.409.743 *tokens* e 1.461.503 *types*; é construído de amostras de textos jornalísticos da internet sobre temáticas diversas (esportes, educação, cultura, etc), sendo atualizado diariamente com novos textos (DIVINO, 2024).

Os critérios escolhidos para analisar os dados foram as padrões da ferramenta, que é a opção de não fazer distinção das palavras escritas com letra maiúscula e minúscula (A=a), entrando na mesma contagem, como também para aparecer somente alfanuméricos (letra, número e hífen). Na configuração da ferramenta *keywords*, foi selecionada a opção de *word* (palavra). Os resultados encontrados foram comparados com o material de insumo, enunciado e texto da tarefa, e também comparados entre as notas. Na Figura 7, pode-se observar os critérios utilizados no programa.

Figura 7 - Critérios selecionados na ferramenta *Keywords*

The screenshot displays the configuration interface for the 'Keywords' tool, organized into several sections:

- Focus subcorpus:** Set to 'none (the whole corpus)'.
- Reference corpus:** Set to 'Portuguese Trends'.
- Reference subcorpus:** Set to 'none (the whole corpus)'.
- Focus on:** A slider set to '1' (neutral), between 'rare' and 'common'.
- Minimum frequency:** Set to '1'.
- Maximum frequency:** Set to '0'.
- Maximum items:** Set to '1000'.
- Checkboxes (all checked):**
 - A = a ?
 - At least one alphanumeric ?
 - Only alphanumeric ?
- Unchecked checkboxes:**
 - Include nonwords ?
 - Exclude these words: ?
 - From list ?
- Identify keywords:** Checked.
 - Keywords settings:**
 - Attribute: 'word'
 - Matching regex: '.*'
- Identify terms:** Checked.
 - Terms settings:**
 - Matching regex: '.*'
- Identify n-grams:** Unchecked.
 - N-grams settings:**
 - Attribute: 'word'
 - Matching regex: '.*'
 - N-gram length: 2, 3, 4, 5, 6 (3 is selected)

Fonte: Elaborado pela autora.

Após escolher os critérios, o usuário é direcionado para outra página do programa onde aparecerá uma lista com as palavras mais típicas do corpus de estudo, como na Figura 8 abaixo.

Figura 8 - Resultados da ferramenta *Keywords*

SINGLE-WORDS ✓ MULTI-WORD TERMS ✓ N-GRAMS ✕

reference corpus: Portuguese Trends (Items: 3,751)

| | Word | Frequency per million | Document frequency | Relative DOCF |
|---|---------------|-----------------------|--------------------|---------------|
| 1 | hortas | 19,778.91 | 200 | 100.00 % ... |
| 2 | wikiversidade | 3,582.16 | 149 | 74.50 % ... |
| 3 | bibiana | 1,559.97 | 71 | 35.50 % ... |
| 4 | cadastramento | 4,756.95 | 130 | 65.00 % ... |
| 5 | visoni | 1,232.57 | 59 | 29.50 % ... |
| 6 | horta | 7,665.05 | 187 | 93.50 % ... |
| 7 | cambará | 1,290.35 | 60 | 30.00 % ... |
| 8 | regiane | 1,348.12 | 68 | 34.00 % ... |
| 9 | nigro | 1,309.61 | 67 | 33.50 % ... |

Fonte: Elaborado pela autora.

Nessa lista, referente ao subcorpus de nota 5, aparece a palavra (*word*), a frequência dessa palavra por milhão de palavras (*frequency per million*), a frequência dessa palavra nos documentos analisados, ou seja, quantos documentos tem a utilização dessa palavra (*document frequency*) e a porcentagem do uso dessa palavra nos documentos analisados (*relative DOCF*). Por exemplo, a palavra “horta” tem a frequência por milhão de 19,778.91 (aparece mais de 19 mil vezes a cada milhão de palavras) e aparece em 200 documentos do corpus de estudo, o que, em porcentagem, condiz com 100% dos textos de nota 5 analisados. Também há outros recursos disponíveis que podem ser visíveis nesta lista, vai depender do foco de cada pesquisa.

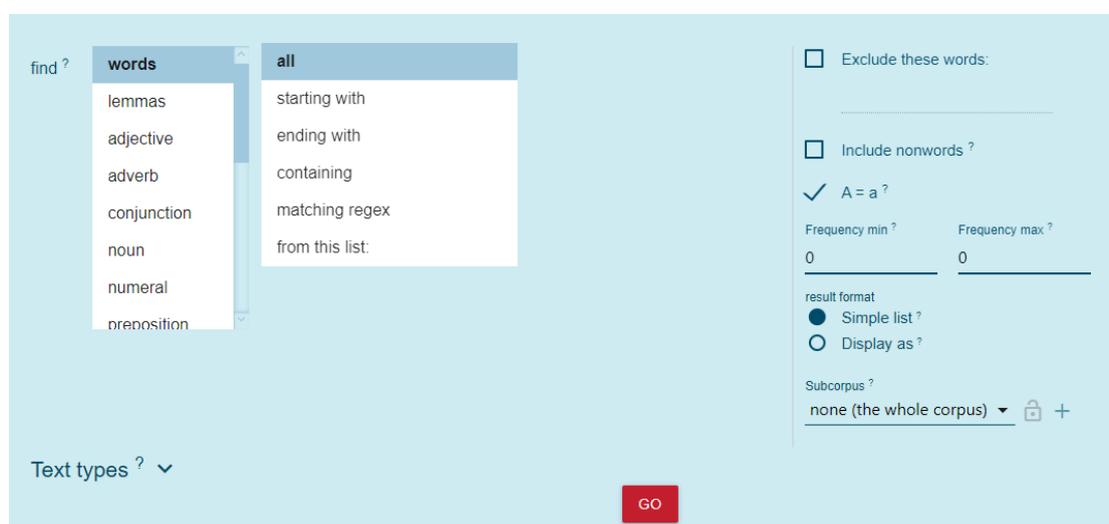
A partir das análises, se fez necessário verificar as *keywords* do material de insumo. Para isso, foi transcrito em um documento do Google Docs o enunciado da tarefa e o texto de insumo, nomeando-o corcel20162t3mi (“mi” de material de insumo). O subcorpus foi colocado no SkE e analisado na ferramenta *keywords*, verificando as palavras mais típicas do corpus de estudo comparado ao corpus de referência Portuguese Trends. Nesta parte da pesquisa, além de verificar a porcentagem de uso desta palavra no corpus de estudo, também foi verificado esta porcentagem no corpus de referência, como também o número de textos que aparecem as *keywords* no corpus de referência.

3.3.1.2 Wordlist

A *wordlist*, diferentemente das *keywords*, trata da exploração de uma lista de frequência de palavras, ordenadas de maior a menor frequência no corpus selecionado, podendo contar com todas as palavras do corpus, a *words*, ou então uma categoria específica,

como uma das classes gramaticais. Neste trabalho, escolhemos listar todas as palavras (*word*), não buscando uma classe de palavras específica. Foi selecionado também que o resultado não fizesse distinção entre letras minúsculas e maiúsculas (A=a), contando como uma palavra só. Os resultados encontrados foram filtrados, sendo utilizado para as análises as palavras que aparecem em até 5% dos textos, depois 10%, 20% e 50%. Após esta etapa, os resultados foram comparados com o material de insumo, enunciado e texto da tarefa, e também comparados entre as notas. Na Figura 9, pode-se observar os critérios escolhidos na ferramenta.

Figura 10 - Critérios selecionados na ferramenta *Wordlist*



The screenshot shows the Wordlist tool interface with the following settings:

- find ?**: A dropdown menu with "words" selected. Other options include lemmas, adjective, adverb, conjunction, noun, numeral, and preposition.
- all**: A dropdown menu with "all" selected. Other options include starting with, ending with, containing, matching regex, and from this list.
- Exclude these words ?**: An unchecked checkbox.
- Include nonwords ?**: An unchecked checkbox.
- A = a ?**: A checked checkbox.
- Frequency min ?**: A text input field with "0".
- Frequency max ?**: A text input field with "0".
- result format**: Radio buttons for "Simple list ?" (selected) and "Display as ?".
- Subcorpus ?**: A dropdown menu with "none (the whole corpus)" selected, a lock icon, and a plus sign.
- Text types ?**: A dropdown menu.
- GO**: A red button at the bottom center.

Fonte: Elaborado pela autora.

Após escolher os critérios para a análise, o programa gera uma lista das palavras mais frequentes às menos frequentes, como mostra a Figura 10 abaixo.

Figura 11 - Resultados da ferramenta *Wordlist*

| | Word | Frequency [?] | DOCF [?] ↓ | Relative DOCF [?] ↓ |
|----|--------|------------------------|---------------------|------------------------------|
| 1 | de | 1,838 | 200 | 100.00 % ... |
| 2 | hortas | 1,027 | 200 | 100.00 % ... |
| 3 | a | 1,436 | 199 | 99.50 % ... |
| 4 | e | 1,521 | 199 | 99.50 % ... |
| 5 | para | 976 | 198 | 99.00 % ... |
| 6 | que | 1,135 | 197 | 98.50 % ... |
| 7 | o | 1,194 | 197 | 98.50 % ... |
| 8 | é | 718 | 193 | 96.50 % ... |
| 9 | da | 514 | 192 | 96.00 % ... |
| 10 | em | 809 | 192 | 96.00 % ... |
| 11 | horta | 398 | 187 | 93.50 % ... |
| 12 | na | 507 | 186 | 93.00 % ... |

Fonte: Elaborado pela autora.

Na lista gerada, referente ao subcorpus de nota 5, temos a coluna das palavras (*word*), a da frequência de uso, que mostra quantas vezes a palavra aparece no corpus (*frequency*), a quantidade de documentos que tem essa palavra (*DOCF*) e a porcentagem de textos do corpus em que a palavra aparece (*relative DOCF*). Para exemplificar, a palavra “de” é usada 1,838 vezes no corpus, aparecendo em 200 textos, o que corresponde a 100% dos textos de nota 5.

3.3.1.3 N-grams

A ferramenta *n-grams* é utilizada para produzir listas de sequência de palavras, podendo ser uma sequência de 2 até 6 *tokens*. Os *n-grams* também são chamados de expressões multipalavras ou pacotes lexicais. A lista é gerada a partir do maior percentual de ocorrência nos textos do corpus até o menor número, sendo excluídos da lista os *n-grams* de baixa frequência.

Ao escolher os critérios para gerar a lista, o usuário pode escolher um atributo, como *word* (palavra) e *lemma* (lema), que são as mais utilizadas no programa. *Lemma*, diferente de *word*, é a forma gráfica da palavra no dicionário, por exemplo, em uma frase onde tem a palavra “sou”, o programa contará o “sou” junto com o “ser” e todas as outras declinações do verbo “ser”. Já a *word* conta cada uso de forma separada.

Os critérios selecionados foram os de procurar por palavras (*word*), incluindo as *nonwords* (não palavras, como os sinais de pontuação). Também foi selecionado para não

fazer distinção entre letras maiúsculas e minúsculas (A=a), e, num primeiro momento, foi buscado por *n-grams* de 4 itens seguidos, depois 5 itens e, por fim, 6 itens que tivessem no mínimo 10% de frequência nos textos. Os resultados encontrados foram comparados com o material de insumo e agrupados por grupo de informação de acordo com o trecho do material de insumo ao qual faziam referência. Também foram comparados posteriormente os resultados de cada subcorpora. Na Figura 11, pode-se observar os critérios escolhidos na ferramenta.

Figura 11 - Critérios selecionados na ferramenta *N-grams*

The image shows the configuration interface for the N-grams tool. The settings are as follows:

- N-gram length:** 4 (selected from 2, 3, 4, 5, 6)
- Attribute:** word
- Frequency min:** 2
- Frequency max:** 0
- Subcorpus:** none (the whole corpus)
- Additional criteria:**
 - Nest n-grams ?
 - Include nonwords ?
 - A = a ?
 - Exclude these words: ?
 - Key n-grams ?
 - Additional criteria menu:
 - all
 - starting with letters
 - ending with letters
 - containing letters
 - starting with word
 - containing word
 - ending with word
 - matching regular expression

Fonte: Elaborado pela autora.

Após escolher os critérios para a análise, o programa gera uma lista de frequência a partir dos critérios selecionados, como na Figura 12 abaixo.

Figura 12 - Resultados da ferramenta *N-grams*

5–6-grams, word (lowercase) (items: 2,942 , total frequency: 13,047)

| | N-gram | Frequency ? | DOCF ? | Relative DOCF ? |
|----|---------------------------------|-------------|--------|-----------------|
| 1 | o link para contato . | 67 | 67 | 33.50% ... |
| 2 | o nome da horta , | 66 | 66 | 33.00% ... |
| 3 | se ela é educativa ou | 64 | 64 | 32.00% ... |
| 4 | se ela é educativa ou comercial | 62 | 62 | 31.00% ... |
| 5 | ela é educativa ou comercial | 62 | 62 | 31.00% ... |
| 6 | preencher o nome da horta | 54 | 54 | 27.00% ... |
| 7 | a região em que está | 45 | 45 | 22.50% ... |
| 8 | é educativa ou comercial , | 44 | 44 | 22.00% ... |
| 9 | nome da horta , se | 43 | 43 | 21.50% ... |
| 10 | e o link para contato | 41 | 41 | 20.50% ... |
| 11 | o bairro , o endereço | 38 | 38 | 19.00% ... |

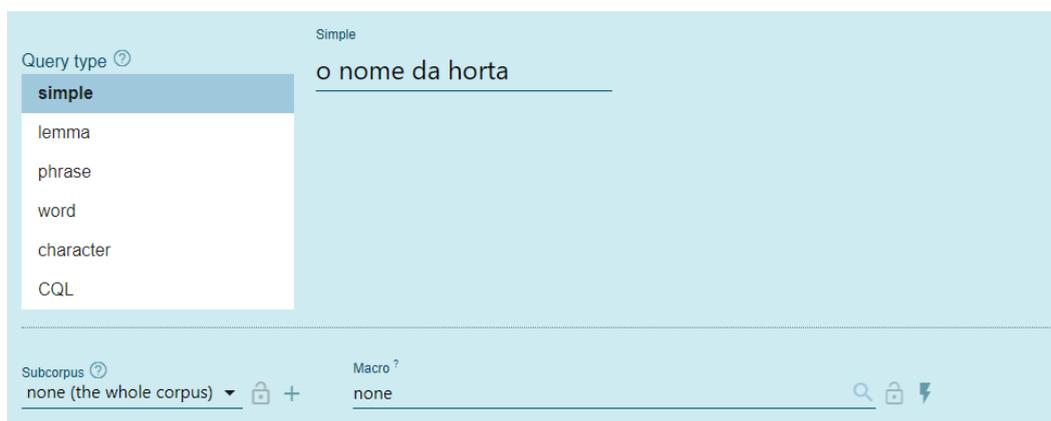
Fonte: Elaborado pela autora.

Nesta lista referente ao subcorpus de nota 5 aparecem os trechos com maior frequência nos textos com extensão de 5 e 6 itens. Ao lado dos *n-grams*, a ferramenta mostra a frequência com que esse pacote lexical aparece no corpus - quantas vezes aparece (*frequency*), em quantos textos aparece esse trecho (*DOCF*) e a porcentagem de textos em que aparece (*relative DOCF*). Por exemplo, o trecho “o link para contato.” aparece 67 vezes e em 67 textos, resultando em 33.50% do subcorpus. Da mesma forma que nas demais ferramentas do *software* existem outros recursos que podem ser visualizados na lista, nos *n-grams* também tem alguns outros dados disponíveis para análise.

3.3.1.4 Concordance

A última ferramenta utilizada foi a *concordance*, que busca palavras ou frases nas linhas de concordância. Neste recurso do *software*, pode-se encontrar exemplos de uma palavra, lema, frase ou estrutura gramatical inserida no contexto em que foi utilizada. O critério utilizado foi o tipo de consulta simples (*simple*) e foram analisados os pacotes lexicais com maior frequência nos subcorpora a partir dos resultados obtidos com a ferramenta *n-grams*. Os resultados obtidos foram analisados, agrupados e comparados entre as duas notas. Na Figura 13, pode-se observar os critérios escolhidos para o pacote lexical “o nome da horta”.

Figura 13 - Critérios selecionados na ferramenta *Concordance*



Fonte: Elaborado pela autora.

Após escolher o que se deseja procurar em contexto, o programa gera uma lista com trechos dos textos do corpus onde aparece a palavra, lema, frase ou estrutura gramatical escolhida. Na Figura 14 abaixo, pode-se verificar os primeiros resultados referentes ao pacote lexical “o nome da horta”.

Figura 14 - Resultados da ferramenta *Concordance*

| | Details | Left context | KWIC | Right context |
|----|---------------------------------|---|------------------------|---|
| 1 | <input type="checkbox"/> doc#0 | AMENTO APRESENTA REQUISITOS IMPORTANTES COMO: | O NOME DA HORTA | , SE TEM CARÁTER EDUCATIVO OU COMERCIAL, A REGIÃO |
| 2 | <input type="checkbox"/> doc#1 | orma e inserir os dados ali. </s><s>É muito fácil, é só preencher | o nome da horta | , se ela é educativa ou comercial, a região em que está, o bairro |
| 3 | <input type="checkbox"/> doc#3 | as hortas no site wikiversidade inserindo os dados delas como | o nome da horta | , o tipo de horta, o bairro onde fica, seu endereço e o link para c |
| 4 | <input type="checkbox"/> doc#9 | de. </s><s>Ali vocês podem inserir os dados como por exemplo | o nome da horta | , o caráter, isto é se ela é educativa ou comercial; a localização |
| 5 | <input type="checkbox"/> doc#12 | mediante a plataforma wikiversidade; ali são inseridos os dados; | o nome da horta | , se ela é educativa ou comercial; a região em que está, o bairro |
| 6 | <input type="checkbox"/> doc#13 | na wikiversidade e inserir os dados ali. </s><s>Tem que colocar | o nome da horta | , e de que caráter é, a região onde está, o bairro, o endereço e |
| 7 | <input type="checkbox"/> doc#14 | dados da horta que você vai identificar. </s><s>Tem que inserir | o nome da horta | , categorizar se ela é educativa ou comercial, identificar a regiã |
| 8 | <input type="checkbox"/> doc#15 | cidade, pode registrá-la na plataforma wikiversidade, inserindo | o nome da horta | , se é educativa ou comercial, a região e o bairro onde se encor |
| 9 | <input type="checkbox"/> doc#16 | im que entrar na pagina wikiversidade e preencher dados como | o nome da horta | e a localização dela. </s><s>Obrigado pela presença no meu blc |
| 10 | <input type="checkbox"/> doc#22 | ento. </s><s>Vocês têm que editar a plataforma wiki, preencher | o nome da horta | , indicar se é uma horta educativa ou comercial, a região em qu |
| 11 | <input type="checkbox"/> doc#23 | na Wikiversidade e inserir os dados da sua horta; preenchendo | o nome da horta | ; se a mesma é educativa ou comercial, qual é á região em que |
| 12 | <input type="checkbox"/> doc#25 | trar, precisa-se acessar a plataforma wikiversidade e preencher | o nome da horta | ; definir se ela é educativa ou comercial; a região em que está; c |
| 13 | <input type="checkbox"/> doc#26 | l "wikinatureza" e inserir os dados ali. </s><s>Tem de preencher | o nome da horta | ; a região em que está, o bairro; o endereço e o link para contat |

Fonte: Elaborado pela autora.

Após gerada a lista pela ferramenta, o próximo passo da análise foi verificar quais palavras ocorriam antes do termo escolhido, o *left context* (contexto do lado esquerdo) e posteriormente o *right context* (contexto do lado direito). Após isso, foram comparados os

resultados encontrados entre os subcorpora. Dessa forma, a análise utilizada nesta parte da pesquisa foi mista, sendo tanto qualitativa quanto quantitativa.

3.3.2. Type-Token-Ratio (TTR)

Conforme Granger e Wynne (1999), o cálculo *Type-Token-Ratio* mostra a variação lexical a partir do número presente em cada corpora de palavras diferentes, isto é, evidência qual corpus apresenta um maior número de palavras diferentes em relação ao número total de palavras do corpus (SKETCH ENGINE, s. d.). Este cálculo foi utilizado em consonância com os trabalhos de Divino (2021; 2024), Hanauer (2023) e Sostruznik (2023), que o utilizaram para verificar a riqueza lexical de cada corpora. A fórmula é

$$\frac{\textit{types}}{\textit{tokens}} \times 100$$

O número de *types* é o número de palavras diferentes encontradas em um corpus e o número de *tokens* é o número total de palavras dentro do corpus. Também a fórmula pede para ser multiplicado por 100 o valor dos *types* dividido pelos *tokens* para que seja possível ler em porcentagem. Por isso, segundo a página do *Sketch Engine* (SKETCH ENGINE, s. d.), é importante que o cálculo seja feito em corpora de mesmo ou similar tamanho de palavras para que não haja uma discrepância muito grande entre os resultados, sendo assim mais coerente o resultado.

Para isso, foram utilizados os valores de *types* e *tokens* disponibilizados pelo *Sketch Engine*. Para acessar essa informação, o usuário deve clicar em “*corpus info*” no *dashboard* do programa, e será direcionado para uma página que contém várias informações sobre o corpus, como na Figura 15:

Figura 15 - Página do *Corpus Info*

corcel20162t3n5r user/divinoluiza/tcc20162t3n5r • created April 29, 2024 at 5:35:03 PM

MANAGE CORPUS MANAGE SUBCORPORA COMPARE CORPORA TEXT TYPE ANALYSIS

GENERAL INFO

Language: Portuguese

CORPUS DESCRIPTION & BIBLIOGRAPHY

TAGSET

WORD SKETCH GRAMMAR

TERM GRAMMAR

COUNTS

| | |
|-----------|--------|
| Tokens | 51,924 |
| Words | 45,848 |
| Sentences | 2,302 |
| Documents | 200 |

TEXT TYPES

TEXT TYPE ANALYSIS

| | |
|----------------------------|-------|
| <doc> (3) | 200 |
| File ID , doc.id | 200 |
| File name , doc.filename | 200 |
| Folder , doc.parent_folder | 1 |
| <g> (0) | 5,808 |
| <s> (0) | 2,302 |

LEXICON SIZES

| | |
|--------------|-------|
| word? | 4,823 |
| tag | 204 |
| lempos? | 2,910 |
| gender_lemma | 2,761 |

COMMON TAGS

| | |
|-------------|----|
| adjective | A* |
| adverb | R* |
| conjunction | C* |
| determiner | D* |

Fonte: Sketch Engine

Para o TTR, o que interessa é o quadro *counts* que mostra o número de *words*, destacado em verde, que corresponde ao número de *tokens* do corpus. O quadro *lexicon sizes* também é importante para o TTR, pois mostra o número de *word* que corresponde ao número de *types* do corpus. Neste exemplo, que é o corpus de nota 5, o número de *tokens* corresponde a 45,848 e o número de *types* corresponde a 4,823. O “*tokens*” presente no quadro *counts* não corresponde ao número de *tokens* do TTR, pois no SkE essa contagem abrange os sinais de pontuação e também números, por isso não foi utilizado no cálculo.

4 ANÁLISE DOS DADOS

Nesta seção, serão apresentadas as análises realizadas com o objetivo de responder a pergunta de pesquisa e as suas subdivisões. Cada subseção será destinada a uma das ferramentas utilizadas: o *Type-Token-Ratio* (TTR), *keywords*, *wordlist*, *n-grams* e *concordance*.

4.1 EXTENSÃO DOS SUBCORPORA E TYPE-TOKEN-RATIO (TTR)

O primeiro objetivo deste trabalho é verificar a extensão dos corpora e também a sua riqueza lexical através do TTR. Verificou-se o número de textos de cada subcorpora, o número de *types*, *tokens* e sentenças, que podem ser observados no Quadro 3 abaixo:

Quadro 3 - Extensão dos subcorpora

| Extensão dos subcorpora | corcel2016t3n2 | corcel2016t3n5 |
|-------------------------|----------------|----------------|
| Textos | 200 | 200 |
| <i>Types</i> | 4.083 | 4.823 |
| <i>Tokens</i> | 31.479 | 45.848 |
| Sentenças | 1.386 | 2.302 |

Fonte: Elaborado pela autora.

No Quadro 2, pode-se verificar que o subcorpora corcel2016t3n2 contém 200 textos, totalizando 4.083 *types* e 31.479 *tokens*, além de 1.386 sentenças no total. No subcorpora corcel2016t3n5, o número de textos é o mesmo, 200 textos, mas há um maior número de *types* e *tokens*, sendo 4.823 e 45.848, respectivamente. As sentenças deste subcorpora também são em maior número, totalizam 2.302 ao todo. Os dois subcorpora apresentam extensões diferentes, apresentando uma diferença maior no número de *tokens* dos subcorpora. Como explicitado na metodologia deste trabalho (seção 4), não houve a normalização da ortografia nos textos utilizados nesta pesquisa, o que pode ocasionar distorção na contagem do número de *types*, principalmente nos textos de nota 2 (já que as diferentes grafias de uma mesma palavra contam como diferentes *types*).

Em conformidade com os resultados de Divino (2021), Hanauer (2023) e Sostruznik (2023), os textos de nota 5 analisados neste trabalho são, em média, mais extensos do que os textos de nota 2, apresentando maior número de *types*, *tokens* e sentenças. Estes resultados apontam que, quanto maior o nível de proficiência do examinando, mais extensas serão as

produções escritas (concordando, assim, com os resultados de DIVINO, 2021; HANAUER, 2023; SOSTRUZNIK, 2023).

Também calculou-se a média de *types*, *tokens* e sentenças por texto em cada subcorpora. Os resultados podem ser observados no Quadro 4 abaixo.

Quadro 4 - Média de *types*, *tokens* e sentenças por texto em cada subcorpora

| | corcel20162t3n2 | corcel20162t3n5 |
|----------------------------------|------------------------|------------------------|
| Média de <i>types</i> por texto | 20,41 | 24,11 |
| Média de <i>tokens</i> por texto | 157,39 | 229,24 |
| Média de sentenças por texto | 6,93 | 11,51 |

Fonte: Elaborado pela autora.

Observa-se que, ao analisar o Quadro 4, o corcel20162t3n2 apresenta uma média de 20,41 *types* por texto, 157,39 *tokens* por texto e 6,93 sentenças por texto; já o corcel20162t3n5 apresenta uma média de 24,11 *types* por texto, 229,24 *tokens* por texto e 11,51 sentenças por texto. Esses resultados mostram que os textos do subcorpus corcel20162t3n5 têm, em média, um número maior de *types*, *tokens* e sentenças por texto do que os do subcorpus corcel20162t3n2, o que deixa mais evidente que os textos que receberam nota 5 nesta tarefa são, em média, maiores em extensão comparados com os textos que receberam nota 2.

Ao calcular o *Type-Token-Ratio*, os resultados obtidos estão presentes no Quadro 5 abaixo.

Quadro 5 - Resultado do Type-Token-Ratio

| Extensão dos subcorpora | corcel20162t3n2 | corcel20162t3n5 |
|--------------------------------|------------------------|------------------------|
| Type-Token-Ratio (TTR) | 12,97% | 10,51% |

Fonte: Elaborado pela autora.

O corcel20162t3n2 referente aos textos de nota 2 teve um percentual de 12,97% no cálculo TTR; já no corcel20162t3n5, que é referente à nota 5, o percentual foi de 10,51%. Isso mostra que os textos que receberam nota 2 no exame apresentam maior riqueza lexical (utilizam mais palavras diferentes em relação ao número total de palavras) do que os textos que receberam nota 5. Esse resultado é contrário ao estudo de Divino (2021), onde os textos

de nota 5 tiveram um percentual de TTR maior, o que seria o esperado, mas vai ao encontro dos resultados obtidos por Hanauer (2023), onde os textos de nota 2 também mostraram uma maior riqueza lexical, e Divino (2024), onde os textos de nota 1, intitulados n1, mostraram ter uma maior riqueza lexical do que as demais notas.

Em seu trabalho, Hanauer (2023) destaca que as inadequações ortográficas encontradas em textos menos proficientes podem influenciar no cálculo TTR, visto que o *software* SkE computa como um novo *type* todas as palavras que são escritas de maneiras diferentes, mesmo que, por exemplo, seja a mesma palavra, mas com um acento a menos, como “orgânicos” ser escrito “organicos”. Isso altera o número de *types*, podendo aumentar o valor que indica a riqueza lexical do corpus. Divino (2024) confirma os resultados de Hanauer (2023), afirmando que o número de “palavras diferentes” de um corpus não indica, necessariamente, um maior repertório linguístico para cumprir os propósitos das tarefas. Provavelmente, é o que acontece também nos dados analisados no presente trabalho. Ao analisarmos as *keywords*, na seção seguinte, poderemos ter uma visão melhor das inadequações ortográficas presentes nos corpora, mesmo não sendo este o foco desta pesquisa.

4.2 KEYWORDS

A ferramenta *keywords*, como já apresentado anteriormente, fornece uma lista com as palavras mais típicas do corpus escolhido como corpus de estudo com base em um corpus de referência que, nesta pesquisa, foi o Portuguese Trends. No Quadro 6, são apresentadas as primeiras 15 *keywords* de cada subcorpus, informando também o percentual de textos em que elas aparecem (Relative DOCF). Apesar de ter a porcentagem do Relative DOCF ao lado de cada palavra, essa lista não está organizada pela frequência de uso, mas sim por sua relevância no subcorpus, na mesma ordem em que aparece no SkE. Para melhor visualização, foram destacadas com a cor azul as palavras que se repetem nos dois subcorpora.

Quadro 6 - 15 primeiras *keywords* e Relative DOCF (%) de cada subcorpora

| corcel20162t3n2 | | corcel20162t3n5 | |
|-----------------|-------------------|-----------------|-------------------|
| Termo | Relative DOCF (%) | Termo | Relative DOCF (%) |
| hortas | 98,50% | hortas | 100% |
| wikiversidade | 43,00% | wikiversidade | 74,50% |
| saudavel | 18,00% | bibiana | 35,50% |

| | | | |
|----------------|--------|----------------|--------|
| cadastroamento | 48,50% | cadastroamento | 65,00% |
| bibiana | 21,00% | visoni | 29,50% |
| visoni | 17,50% | horta | 93,50% |
| cultivos | 36,50% | cambará | 30,00% |
| horta | 66,00% | regiane | 34,00% |
| regiane | 25,50% | nigro | 33,50% |
| organico | 9,50% | cultivos | 25,50% |
| importancia | 11,00% | wiki | 17,00% |
| cambará | 16,50% | educativa | 62,00% |
| nigro | 22,50% | orgânicos | 61,00% |
| ciudade | 7,00% | incluírem | 16,00% |
| incluierem | 10,00% | kairós | 13,50% |

Fonte: Elaborado pela autora.

Comparando as duas colunas, mais da metade das palavras aparece como palavras mais relevantes nos dois subcorpora, como “hortas”, que aparece em 100% dos textos nota 5 e 98,50% dos textos nota 2; “wikiversidade”, que aparece em maior quantidade nos textos de nota 5, sendo 74,50%, e na nota 2 em 43,00%; “cadastroamento”, com maior incidência na nota 5, 65,00%, e 48,50% na nota 2; “visoni” (n5 29,50% e n2 17,50%), “regiane” (n5 34,00% e n2 25,50%) e “nigro” (n5 33,50% e n2 22,50%) também aparecem em ambas as listas, sendo o nome/sobrenome das idealizadoras do projeto; “horta” aparece em quase todos os textos de nota 5, 93,50%, porém na nota 2 aparece em apenas 66,00% dos textos; “cultivos” já mostra um comportamento diferentes dos demais, ocorrendo mais nos textos de nota 2, 36,50%, do que nos de nota 5, 25,50%.

Ao analisar o material de insumo, todas essas palavras que aparecem nos dois subcorpora estão presentes no material de insumo e mostram-se relevantes para o cumprimento da tarefa na medida que trazem informações importantes sobre o projeto do cadastroamento de hortas, como as palavras “hortas” e “horta”, que dizem respeito ao assunto principal da tarefa; os termos “wikiversidade” e “cadastroamento” são relacionados com a plataforma onde devem ser cadastradas as hortas, o que se mostra uma informação relevante da tarefa; “visoni”, “regiane” e “nigro” são os nomes das idealizadoras do projeto, mas, como mostra a sua baixa porcentagem em ambos os subcorpora, não se mostra tão relevante para o cumprimento da tarefa; e “cultivos” que também tem relação com o assunto de hortas, mas

não se mostra como uma informação imprescindível para cumprir a tarefa, já que sua porcentagem é baixa em ambos os subcorpora.

Além disso, há algumas palavras diferentes que aparecem nas listas se comparadas entre si. Na nota 5, aparecem os termos “wiki” (17,00%), “educativa” (62,00%), “orgânicos” (61,00%), “incluírem” (16,00%) e “kairós” (13,50%), sendo “educativa” e “orgânicos” as mais frequentes nos textos. Comparadas com o material de insumo, todas elas também estão presentes.

Na nota 2, aparecem os termos “saudavel” (18,00%), “organico” (9,50%), “importancia” (11,00%), “ciudade” (7,00%) e “incluirem” (10,00%). Todas estas palavras que aparecem como *keywords* no corpus de nota 2 mas não aparecem no corpus de nota 5 têm alguma inadequação ortográfica, na sua maioria por falta de um acento na palavra. Se “organico” e “incluirem” fossem normalizadas, por exemplo, seriam idênticas às palavras da nota 5, “orgânicos” e “incluírem”. Apesar de se mostrarem relevantes em comparação com o corpus de referência, justamente por não aparecerem no corpus de referência, a porcentagem dessas palavras com inadequação ortográfica não chega a 20% dos textos. Conforme Hanauer (2023) e Divino (2024), os subcorpora de notas mais baixas apresentam mais palavras com inadequações ortográficas, o que confirma a hipótese de que essas inadequações modificam o resultado do TTR.

Outras duas palavras que aparecem nas duas listas de *keywords* são “bibiana” e “cambará”. Essas duas, como já explicado anteriormente, fazem parte do processo de anonimização dos nomes escritos nos textos, substituindo as assinaturas pelo nome Bibiana e os sobrenomes, caso tivesse, por Cambará. No caso desta tarefa específica, no subcorpus de nota 2 a palavra “bibiana” aparece em 21,00% dos textos, e “cambará” em 16,50%. No subcorpus de nota 5, “bibiana” aparece em 35,50% dos textos e “cambará” em 30,00% dos textos. Apesar de incidir mais na nota 5, o percentual de uso dessas palavras (que representam a assinatura dos textos) é relativamente baixo, não parecendo ser um termo de extrema importância para o cumprimento da tarefa. Essa não necessidade de assinatura está relacionada ao gênero do texto solicitado (um texto para blog), que não costuma demandar assinatura.

Para expandir os resultados e verificar em uma escala maior quais palavras utilizadas pelos examinados nas notas 2 e 5 estão presentes no material de insumo e quais não estão, decidiu-se analisar as 100 primeiras *keywords* e, desta vez, ordená-las por Relative DOCF, do mais frequente ao menos frequente. O Quadro 7 apresenta todas as palavras, entre as 100 primeiras *keywords*, que não estão no material de insumo. Decidiu-se também marcar as

palavras com inadequações ortográficas para, conforme Divino (2024), evidenciar as diferentes grafias de uma mesma palavra.

Quadro 7 - Keywords que não estão presentes no material de insumo ordenados pelo Relative DOCF (%) de cada subcorpora

| corcel20162t3n2 | corcel20162t3n5 |
|---------------------|----------------------|
| bibiana (21%) | bibiana (35,50%) |
| saudavel (18%) | cambará (30%) |
| tambem (17,50%) | orgánicos (9%) |
| cambará (16,50%) | noticia (8%) |
| importancia (11%) | participem (7%) |
| alem (10,50%) | prezados (6,50%) |
| incluïrem (10%) | alí (6,50%) |
| organico (9,50%) | facil (6,50%) |
| saude (9%) | convïdo (6,50%) |
| organicos (8%) | utilidades (6%) |
| voluntarias (8%) | voluntarios (6%) |
| facil (8%) | incluïrem (6%) |
| atenciosamente (8%) | tambem (5,50%) |
| ciudade (7%) | aquí (5,50%) |
| barrio (7%) | possiveis (5%) |
| orgánicos (6,50%) | importancia (4,50%) |
| metropole (6%) | comunitarias (4,50%) |
| interesante (6%) | cadastre (4,50%) |
| comunitarias (6%) | cadastrarem (4,50%) |
| academicos (6%) | alem (4,50%) |
| municipio (6%) | países (4%) |
| productores (5,50%) | saudavel (4%) |
| saudaveis (5,50%) | beneficios (4%) |
| voluntarios (5,50%) | cadastrem (4%) |
| organica (5%) | voluntarias (4%) |
| voces (5%) | cadastrando (4%) |
| países (5%) | idéia (4%) |
| convïdo (4,50%) | municipio (3,50%) |
| beneficios (4,50%) | podem-se (3%) |
| alí (4,50%) | orgánico (3%) |

| | |
|---------------------|--------------------|
| carater (4,50%) | existencia (3%) |
| kairos (4,50%) | territorio (3%) |
| prezados (4,50%) | qué (3%) |
| orgánicos (4%) | brasilía (2,50%) |
| comercio (4%) | cultivam (2,50%) |
| productos (4%) | ajudemos (2,50%) |
| asim (4%) | academicos (2,50%) |
| interesadas (4%) | prencher (2,50%) |
| estam (3,50%) | comentarios (2%) |
| quimicos (3,50%) | quizer (2%) |
| día (3,50%) | interesadas (2%) |
| organicas (3%) | algúm (2%) |
| invito (3%) | orgánicas (2%) |
| orgánica (3%) | invito (2%) |
| areas (3%) | començo (1,50%) |
| pedido (3%) | |
| existencia (3%) | |
| propria (3%) | |
| idea (3%) | |
| interesados (2,50%) | |
| alimentação (2,50%) | |
| datos (2,50%) | |
| voluntario (2,50%) | |
| cadastra (2,50%) | |
| possiveis (2,50%) | |
| familias (2,50%) | |
| creação (2%) | |
| bolivia (2%) | |
| ecologicas (2%) | |
| importança (2%) | |
| alimentação (2%) | |
| projeto (1,50%) | |
| través (1,50%) | |
| proprias (1%) | |

Fonte: Elaborado pela autora.

Ao analisar as palavras e o percentual presente no Quadro 7, os dados não se mostram tão relevantes por conta da baixa porcentagem de frequência nos textos, tendo algumas palavras com frequência de 1%, o que não é relevante para as análises. Até a *keyword* com maior frequência, “bibiana” com 35,50%, mostra-se com um percentual baixo de presença nos textos.

A lista do subcorpus de nota 5 é de 45 palavras que não estão presentes no material de insumo, o que é menor do que a lista do subcorpus de nota 2, de 64 palavras, 19 palavras a mais do que os textos de nota 5. Apesar desse número elevado de palavras que não estão presentes no material de insumo, apenas 13 palavras da lista de nota 5 não contêm inadequações ortográficas, que são “bibiana”, “cambará”, que dizem respeito às assinaturas, e “participem”, “prezados”, “convido”, “utilidades”, “cadastre”, “cadastrarem”, “cadaستم”, “cadastrando”, “podem-se”, “cultivam” e “ajudemos”, que têm relação com a temática proposta na tarefa. Na nota 2 esse número cai ainda mais, sendo apenas 5 palavras da lista de keywords de fora do material de insumo que não contêm inadequações ortográficas, que são “bibiana” e “cambará”, “convido” e “prezados”, que mostram relação com a proposta da tarefa. A palavra “atenciosamente” também está na lista, mas parece demonstrar um problema de construção de gênero, pois é uma palavra utilizada em cartas, mas normalmente não é utilizada em textos de blog. No entanto, como já comentado anteriormente, não são dados relevantes para descrever os subcorpora, pois as porcentagens de ocorrência são baixas.

Como a ferramenta *keywords* procura as palavras mais típicas em um corpus de estudo comparado com o corpus de referência, a lista traz muitas palavras com inadequações ortográficas porque de fato são acontecimentos raros, não fazendo parte do corpus de referência, e isso se confirma ao analisar a porcentagem de frequência no corpus nas palavras com inadequações ortográficas.

Esses resultados parecem apontar para o fato de que a Tarefa 3 da edição de 2016-2 é uma tarefa que não exige um vocabulário muito especializado e rico do examinando, que pode cumprir a tarefa com palavras de uso mais frequente, diferentemente do que foi encontrado na pesquisa de Divino (2024) com a tarefa 4 de 2015-2. Essa afirmação vem do fato de que, ao analisar os resultados da ferramenta *keywords*, as palavras que se mostram típicas no corpus de estudo, não são palavras especializadas, com a exceção de algumas mais específicas da temática da tarefa, como “hortas”, “cadastramento”, “wikiversidade” e “cultivos”. Essa presença de poucas palavras-chave também justifica o grande número de inadequações ortográficas nas listas de *keywords*, pois, por ser uma inadequação ortográfica,

é de fato algo raro, não sendo esperado que muitos examinandos cometam as mesmas inadequações nas mesmas palavras.

Com o objetivo de verificar esta hipótese, foi analisado e colocado no SkE o corpus corcel20162t3mi, que diz respeito ao texto e ao enunciado do material de insumo. No Quadro 8, estão as informações referente à extensão do corpus.

Quadro 8 - Extensão do corpus corcel20162t3mi

| Extensão do corpus | corcel20162t3mi |
|--------------------|-----------------|
| <i>Types</i> | 355 |
| <i>Tokens</i> | 631 |
| Sentenças | 23 |

Fonte: elaborado pela autora.

O corcel20162t3mi se constitui do enunciado e do material de insumo (Anexo 1), de 355 *types*, 631 *tokens* e 23 sentenças. Este corpus foi analisado a partir da ferramenta *keywords* com o Portuguese Trends como corpus de referência, com o objetivo de encontrar as palavras mais típicas do corpus de estudo comparado com o corpus de referência. Para fazer um recorte, no Quadro 9 estão as 15 primeiras *keywords* do corpus corcel20162t3mi⁶. Além das palavras encontradas, foram adicionadas no quadro a frequência dessa palavra nos corpora de estudo e de referência (*frequency*), bem como o número de textos em que esta palavra está presente no corpus de referência (DOCF).

Quadro 9 - 15 primeiras *keywords* do corpus corcel20162t3mi

| Itens | Frequency (focus) | Frequency (reference) | DOCF (reference) |
|----------------------|-------------------|-----------------------|------------------|
| hortas | 12 | 1226 | 727 |
| visoni | 3 | 6 | 3 |
| regiane | 5 | 691 | 426 |
| kairós | 2 | 156 | 74 |
| cooperapas | 1 | 0 | 0 |
| wikiversidade | 1 | 0 | 0 |
| hortelões | 1 | 4 | 4 |
| nigro | 2 | 842 | 397 |
| rincões | 1 | 145 | 139 |

⁶ No Anexo 2 está presente a lista das *keywords* na íntegra.

| | | | |
|----------------------|---|------|------|
| cadastramento | 3 | 1835 | 1353 |
| inserem-se | 1 | 167 | 166 |
| colaborarem | 1 | 228 | 223 |
| wiki | 1 | 238 | 162 |
| corujas | 1 | 239 | 144 |
| cambuci | 1 | 245 | 187 |

Fonte: Elaborado pela autora.

No Quadro 9, os números referentes ao corpus de estudo (*focus*) são baixos, pois o corpus de estudo é composto somente pelo enunciado da tarefa e o texto de insumo. O que chama a atenção é que mesmo as *keywords*, que deveriam ser as palavras mais típicas comparadas com o corpus de referência (*reference*), apresentam um número elevado no Portuguese Trends. Por exemplo, a palavra “hortas” aparece 1.226 vezes no corpus de referência; “cadastramento” também aparece com um número alto de frequência no corpus de referência, sendo 1.835 vezes. Esses resultados parecem indicar, confirmando a hipótese levantada anteriormente, que as palavras presentes no material de insumo não são palavras específicas da tarefa. Não há a presença de muitas palavras incomuns no material de insumo desta tarefa, e por isso aparecem poucas palavras-chaves no corpus de referência no Quadro 9.

Também há algumas palavras que de fato apresentam pouquíssima frequência no corpus de referência, como “visoni”, com apenas 6 vezes, “cooperapas”, com 0 vezes, “wikiversidade”, com 0 vezes também, e “hortelões”, com 4 vezes, sendo de fato palavras típicas do corpus de estudo. Porém, os demais itens aparecem com bastante frequência no corpus de referência e, ao olhar a lista de *keywords* presente no Quadro 9, a porcentagem de uso no corpus de referência vai só aumentando.

4.3 WORDLIST

Com o objetivo de verificar quais informações do material de insumo foram recuperadas pelos examinandos, foi utilizada a ferramenta *wordlist* nos dois subcorpora. Como existe em cada subcorpora uma grande quantidade de palavras, decidiu-se por restringir, conforme Divino (2024), as análises para as palavras que estivessem presentes em pelo menos 5% dos textos de cada subcorpora. No Quadro 10, pode-se observar os resultados encontrados, informando o número total de *types* que aparecem em ao menos 5% dos textos, mostrando quantas dessas palavras estão presentes no material de insumo, sendo elas idênticas ou com alguma inadequação ortográfica, e também quantas palavras não estão

presentes no material de insumo, indicando quantas apresentam inadequações ortográficas e quantas não. Para além dos números brutos, foram adicionadas as porcentagens aproximadas do que aqueles números representam com base no total de palavras encontradas em cada subcorpus.

Quadro 10 - Quantidade de palavras que aparecem no mínimo em 5% dos textos em cada subcorpora

| Subcorpus | Total de <i>types</i> | Estão presentes no material de insumo | | Não estão presentes no material de insumo | |
|------------------------|-----------------------|---------------------------------------|-------------------------------|---|-------------------------------|
| | | Idênticas | Com inadequações ortográficas | Sem inadequações ortográficas | Com inadequações ortográficas |
| corcel20162t3n5 | 516 (100%) | 240 (46,51%) | 10 (1,93%) | 266 (51,55%) | 0 (0%) |
| corcel20162t3n2 | 405 (100%) | 240 (59,25%) | 24 (5,92%) | 135 (33,33%) | 6 (1,48%) |

Fonte: Elaborado pela autora.

No subcorpus corcel20162t3n5, foram encontrados 516 *types* com no mínimo 5% de frequência nos textos, sendo 240 idênticas às palavras presentes no material de insumo (46,51%) e apenas 10 com inadequações ortográficas que também estão no material de insumo (1,93%); ao olhar para as palavras que não estão presentes no material de insumo e são encontradas em ao menos 5% dos textos, identificou-se 266 *types* (51,55%) e nenhuma com inadequações ortográficas (0%). Esse resultado já mostra que os textos de nota 5 utilizaram, levando em conta a frequência mínima em 5% dos textos, mais palavras que não estão presentes no material de insumo, apesar de o número ser bem próximo entre palavras do material de insumo e palavras que não estão no material de insumo, sendo a diferença de apenas 5,04%. Isso parece ter relação com o fato de que, como visto na subseção das *keywords*, as palavras do material de insumo não são palavras tão específicas, o que pode facilitar o processo de paráfrase das informações presentes no material, diminuindo assim as cópias e o uso de palavras idênticas ao material de insumo.

No subcorpus corcel20162t3n2, foram encontrados 405 *types*, 111 *types* a menos do que no subcorpus de nota 5, com frequência em no mínimo 5% dos textos, sendo 240 idênticas às palavras do material de insumo (59,25%) e 24 com inadequações ortográficas que também estão presentes no material de insumo (5,92%); as palavras que não estão

presentes no material de insumo são, diferentemente do subcorpus de nota 5, em número menor do que as que estão presentes, sendo 135 *types* (33,33%) e 6 com inadequação ortográfica (1,48%). Na nota 2 o resultado mostra que, com frequência em, no mínimo, 5% dos textos, os examinandos utilizaram mais as palavras que estão presentes no material de insumo, com uma diferença de 25,92%.

Ao comparar os resultados obtidos nos dois subcorpora, pode-se afirmar que os textos de nota 5, apesar de utilizarem muitas palavras do material de insumo, utilizam mais palavras que estão fora do material de insumo. O contrário acontece nos textos de nota 2, mostrando que foram utilizadas mais palavras presentes no material de insumo. Sobre as inadequações ortográficas, os textos de nota 2 apresentam mais inadequações, porém as porcentagens são baixas se comparado ao número total de *types* presente na análise. Diferentemente dos textos de nota 5, os textos de nota 2 utilizaram também palavras com inadequações ortográficas que não estão presentes no material de insumo, o que é esperado em níveis mais baixos de proficiência.

Dando seguimento nos resultados encontrados, no Quadro 11 está um resumo do que foi encontrado ao aumentar a porcentagem de frequência das palavras para no mínimo 10%.

Quadro 11 - Quantidade de palavras que aparecem no mínimo em 10% dos textos em cada subcorpora

| | | Estão presentes no material de insumo | | Não estão presentes no material de insumo | |
|-----------------|-----------------------|---------------------------------------|-------------------------------|---|-----------------------------|
| Subcorpus | Total de <i>types</i> | Idênticas | Com inadequações ortográficas | Sem inadequação ortográfica | Com inadequação ortográfica |
| corcel20162t3n5 | 304 (100%) | 194 (63,81%) | 0 (0%) | 110 (36,18%) | 0 (0%) |
| corcel20162t3n2 | 236 (100%) | 178 (75,42%) | 3 (1,27%) | 53 (22,45%) | 2 (0,84%) |

Fonte: Elaborado pela autora.

No subcorpus corcel20162t3n5, foram encontrados 304 *types* com frequência em, no mínimo, 10% dos textos, sendo 194 idênticas às palavras do material de insumo (63,81%), nenhuma com inadequações ortográficas (0%). Ao verificar as palavras que não estão presentes no material de insumo, foram encontradas 110 palavras (36,18%), também nenhuma com inadequações ortográficas (0%). Esses resultados se mostram diferentes dos

resultados obtidos quando analisamos as palavras com frequência em no mínimo 5% dos textos, verificando-se, aqui, o uso de mais palavras provenientes do material de insumo. Esse resultado mostra que as palavras que se repetem em um número maior de textos (pelo menos 10% dos textos) são, em sua maioria, provenientes do material de insumo.

No subcorpus corcel20162t3n2, foram encontrados 236 *types* com frequência em no mínimo 10% dos textos, 68 *types* a menos do que no subcorpus referente à nota 5. Desses 236 *types*, 178 estão presentes no material de insumo (75,42%) e 3 apresentam inadequação ortográfica (1,27%). Analisando as palavras que não estão presentes no material de insumo, identificou-se 53 palavras (22,45%) e mais 2 que apresentaram inadequações ortográficas (0,84%). Diferentemente do que foi analisado no quadro 10, ao verificar os *types* com frequência em no mínimo 10% dos textos, o número de inadequações ortográficas diminuiu bastante, principalmente nas palavras provenientes do material de insumo. Isso era esperado, visto que, para aparecer na listagem, a palavra precisaria ter sido escrita com a mesma inadequação por, ao menos, 10% dos examinandos.

Sobre as inadequações ortográficas, os textos que receberam nota 5 mostram um número nulo para inadequações ortográficas com frequência mínima em 10% dos textos; os textos de nota 2, apesar de ser um número bem baixo, mostram ainda algumas inadequações ortográficas que estão presentes tanto em palavras do material de insumo como em palavras que não estão no material de insumo. Isso parece confirmar que há mais palavras com inadequações ortográficas no subcorpus de nota 2 do que no subcorpus de nota 5, como também que, com o aumento da porcentagem mínima de frequência, as inadequações ortográficas vão ficando cada vez mais raras.

Aumentando um pouco mais a porcentagem mínima de frequência, no Quadro 12 pode-se verificar os resultados encontrados de *types* com frequência em no mínimo 20% nos textos de cada subcorpora.

Quadro 12 - Quantidade de palavras que aparecem no mínimo em 20% dos textos em cada subcorpora

| | | Estão presentes no material de insumo | | Não estão presentes no material de insumo | |
|-----------|----------------|---------------------------------------|-------------------------------|---|-----------------------------|
| Subcorpus | Total de types | Idênticas | Com inadequações ortográficas | Sem inadequação ortográfica | Com inadequação ortográfica |
| | | | | | |

| | | | | | |
|------------------------|---------------|-----------------|-----------|----------------|-----------|
| corcel20162t3n5 | 166 (100%) | 136 (81,92%) | 0 (0%) | 30 (18,07%) | 0 (0%) |
| corcel20162t3n2 | 124 (100%) | 109 (87,20%) | 0 (0%) | 15 (12,09%) | 0 (0%) |

Fonte: Elaborado pela autora.

No subcorpus corcel20162t3n5 encontrou-se 166 *types* com frequência em no mínimo 20% dos textos, sendo 136 idênticas ao material de insumo (81,92%), e 30 palavras utilizadas que não estão presentes no material de insumo (18,07%) . Da mesma forma que na análise das palavras com no mínimo 10% de frequência, percebe-se o maior uso de palavras presentes no material de insumo, sendo bem baixa a porcentagem de palavras utilizadas de fora do material de insumo. Isso parece demonstrar que, ao passo que o número mínimo de frequência aumenta nas análises, o número de palavras que não estão no material de insumo diminui.

No subcorpus corcel20162t3n2 foram encontrados 124 *types* com frequência em no mínimo 20% dos textos, 42 a menos do que no corpus com textos de nota 5. Desses 124 *types*, 109 estão presentes no material de insumo (87,20%) e 15 não estão presentes (12,09%). Diferentemente dos resultados obtidos com frequência em no mínimo 5% e 10% dos textos, neste caso não há palavras com inadequações ortográficas. Ao comparar-se o resultado dos dois subcorpora, percebe-se que, nas frequências aqui analisadas, em ambos são utilizadas mais palavras provenientes do material de insumo, sendo pouquíssimas as palavras utilizadas que não estão presentes no material. Esses dados apontam para a diminuição do número de palavras que não estão no material de insumo conforme a porcentagem mínima de frequência vai aumentando, o que já era esperado, visto que é natural que as palavras do material de insumo se repitam em um maior número de textos produzidos pelos examinandos.

Para finalizar as análises utilizando a ferramenta *wordlist*, foram analisados os *types* com frequência em no mínimo 50% dos textos. O Quadro 13 apresenta os resultados obtidos com esta porcentagem.

Quadro 13 - Quantidade de palavras que aparecem no mínimo em 50% dos textos em cada subcorpora

| | | |
|--|--|--|
| | Estão presentes no material de insumo | Não estão presentes no material de insumo |
|--|--|--|

| Subcorpus | Total de types | Idênticas | Com inadequações ortográficas | Sem inadequação ortográfica | Com inadequação ortográfica |
|-----------------|----------------|--------------|-------------------------------|-----------------------------|-----------------------------|
| corcel20162t3n5 | 60 (100%) | 57 (95%) | 0 (0%) | 3 (5%) | 0 (0%) |
| corcel20162t3n2 | 31 (100%) | 31 (100%) | 0 (0%) | 0 (0%) | 0 (0%) |

Fonte: Elaborado pela autora.

No subcorpus corcel20162t3n5, foram encontrados 60 *types* com frequência em no mínimo 50% de frequência dos textos, sendo 57 idênticas ao material de insumo (95%) e 3 que não estão presentes no material de insumo (5%). Esses resultados mostram que as palavras utilizadas em no mínimo 50% dos textos estão, em sua grande maioria, presentes no material de insumo, sendo quase nulo o número de palavras utilizadas de fora do material.

No subcorpus corcel20162t3n2, foram encontradas apenas 31 *types* com frequência em no mínimo 50% dos textos, 29 a menos do que no subcorpus de nota 5. Desses 31 *types*, todos estão presentes no material de insumo, não havendo nenhuma palavra com frequência em no mínimo 50% dos textos de fora do material de insumo. Esses resultados mostram apenas palavras presentes no material de insumo foram utilizadas em no mínimo 50% dos textos, o que é diferente do que tinha sido analisado até então nesse subcorpus.

Ao comparar os resultados obtidos nos dois subcorpora, percebe-se em ambos um baixo número de palavras com frequência em no mínimo 50% dos textos. Isso parece ter relação com o fato de que, como visto na subseção das *keywords*, as palavras do material de insumo não são palavras específicas, o que facilita o processo de parafraseamento das informações presentes no material, e conseqüentemente diminui o uso de palavras iguais entre os examinandos, já que cada examinando pode parafrasear uma palavra do material de insumo de forma diferente.

Além disso, os resultados obtidos com a ferramenta *wordlist* parecem mostrar que, assim como nos resultados de Divino (2024), o número de *types* encontrados em no mínimo 5% dos textos é muito pequeno nos dois subcorpora se comparado com o número de *types* total de cada subcorpus, e essa quantidade vai diminuindo à medida que é aumentado o percentual de frequência mínima. Isso significa que a grande maioria das palavras utilizadas não se repete em nem 5% dos textos de nenhum dos subcorpora, o que indica que os examinandos usaram, nesta tarefa, um vocabulário bastante rico e diversificado que não se repete muito nos diferentes textos, sendo algo mais subjetivo de cada texto. No entanto, nos

textos de nota 2, esse vocabulário rico pode também estar relacionado com as inadequações ortográficas que, como vimos nas *keywords* e nos resultados obtidos na *wordlist*, são mais frequentes no subcorpus de nota 2.

Apesar desse número pequeno de *types* repetidos em ao menos 5% dos textos, o subcorpus corcel20162t3n5 mostrou que as palavras com frequência em no mínimo 5% dos textos eram, em sua maioria, de fora do material de insumo, o que, como já mencionado, pode ter relação com o possível parafraseamento de informações, como também com o uso de um vocabulário mais rico. Da mesma forma, havia a presença de um número baixo de inadequações ortográficas nas palavras pertencentes ao material de insumo. Esse resultado, no entanto, mudou ao se alterar a frequência mínima, já que não havia ocorrências de palavras com inadequações ortográficas e as palavras usadas em no mínimo 10%, 20% e 50% dos textos eram, em sua maioria, pertencentes ao material de insumo, o que já era esperado, pois é natural que as palavras do material de insumo se repitam em um maior número de textos produzidos pelos examinandos do que as palavras de fora do material de insumo.

O subcorpus corcel20162t3n2 manteve a porcentagem maior no número de palavras que estão presentes no material de insumo em todas as frequências mínimas utilizadas. Nas frequências de 5% e 10%, há a presença de palavras com inadequação ortográfica tanto pertencentes ao material de insumo como de fora do material, o que parece ser uma característica dos textos de nota 2. Nas frequências de 20% e 50%, não havia mais a presença de palavras com inadequações ortográficas, o que já era esperado, pois as inadequações ortográficas, mesmo nos textos de nota 2, não acontecem de forma idêntica numa frequência tão alta como 20% e 50% dos textos. Isso confirma a hipótese de a repetição de palavras com a mesma inadequação em vários textos ser incomum.

Além disso, a análise também evidenciou que há inadequações ortográficas em ambos os subcorpora, não sendo uma característica específica dos texto de nota 2, apesar de haver mais frequência de palavras com inadequações ortográficas neste subcorpus.

4.4 N-GRAMS

Nesta subseção, com o objetivo de expandir as análises sobre a recuperação de informações do material de insumo, serão apresentados os resultados obtidos com a ferramenta *n-grams*. Como já mencionado anteriormente, foram selecionados os critérios de procurar por *word*, não distinguindo palavras escritas com letras maiúsculas ou minúsculas (A=a). Estão contadas nos resultados as não-palavras (pontuação e sinais diversos), isto é, uma vírgula, por exemplo, contará como uma palavra individual. Em um primeiro momento,

foram analisados os *n-grams* com 4 itens em sequência com frequência em, no mínimo, 10% dos textos do subcorpus, após 5 itens e, por fim, 6 itens. No Quadro 14 abaixo, pode-se observar quantas vezes cada sequência de itens foi encontrada em cada um dos subcorpora.

Quadro 14 - Quantidade de itens em sequência que aparecem no mínimo em 10% dos textos em cada subcorpora⁷

| Subcorpus | <i>N-grams</i> de 4 itens presentes em no mínimo 10% dos textos | <i>N-grams</i> de 5 itens presentes em no mínimo 10% dos textos | <i>N-grams</i> de 6 itens presentes em no mínimo 10% dos textos |
|-----------------|---|---|---|
| corcel20162t3n5 | 100 | 72 | 52 |
| corcel20162t3n2 | 54 | 27 | 12 |

Fonte: Elaborado pela autora.

No Quadro 14, é possível verificar que o corcel20162t3n5 apresenta um total de 100 *n-grams* com sequência de 4 elementos com frequência em, no mínimo, 10% dos textos, sendo a sequência “o nome da horta” a que apresenta maior frequência (52%). O número de *n-grams* de 5 elementos diminui um pouco, totalizando 72 *n-grams*, sendo “o link para contato .” a sequência com maior porcentagem (33,50%). Esse número baixa mais ainda quando olhamos para os *n-grams* de 6 palavras em sequência, sendo 52 *n-grams* com a sequência mais frequente “se ela é educativa ou comercial” com 31%.

No subcorpus corcel20162t3n2, diferentemente do esperado, o número de *n-grams* em sequência é menor do que no subcorpus de nota 5. Foram encontrados 54 *n-grams* com sequência de 4 itens (46 a menos do que no corcel20162t3n5), sendo “inclusão de cultivos urbanos” o *n-grams* com maior porcentagem (21,50%); 27 *n-grams* com a sequência de 5 palavras (45 a menos do que o corcel20162t3n5), sendo “a inclusão de cultivos urbanos” o *n-grams* com maior frequência, (19%); e apenas 12 *n-grams* com sequência de 6 palavras (40 a menos do que o corcel20162t3n5), com o trecho “na plataforma wikiversidade e inserir os” (15,50%) sendo o de maior porcentagem.

Em um primeiro momento, foram analisados de forma mais minuciosa os *n-grams* com 4 elementos em sequência do subcorpus corcel20162t3n5. Foram encontrados na lista 100 sequências de 4 elementos com no mínimo 10% de frequência nos textos. A partir desta lista, foi agrupado as sequências por grupo de informação presente no texto de insumo. No

⁷ As listas completas com todas as sequências encontradas em cada subcorpora estão presentes na seção de anexos (Anexo 3, 4, 5, 6, 7 e 8).

Anexo 3, está presente o quadro na íntegra com todos os grupos de *n-grams* agrupados e suas porcentagens. No Quadro 15, estão os grupos de informação que puderam ser agrupados a partir do material de insumo por ordem que apareceram na lista do SkE.

Quadro 15 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n5 com 4 elementos em sequência

| |
|--|
| 1) souber da existência de alguma horta urbana e orgânica (...) pode entrar na plataforma wikiversidade e inserir os dados ali . (...) tem <i>que</i> preencher o nome da horta (;) (;) se ela é educativa ou comercial (;) (;) a região em que está (;) (;) o bairro (;) (;) o endereço e (;) o link para contato (52 itens) |
| agrupamento de 59 n-grams de 4 elementos em sequência |
| 2) estimular os produtores comerciais orgânicos a incluírem suas hortas (...) e se tornarem mais conhecidos (14 itens) |
| agrupamento de 6 n-grams de 4 elementos em sequência |
| 3) atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade . (15 itens) |
| agrupamento de 12 n-grams de 4 elementos em sequência |
| 4) o cadastro de hortas em São Paulo (,) (.) e (8 itens) |
| agrupamento de 7 n-grams de 4 elementos em sequência |
| 5) na cidade de São Paulo (,) (.) (6 itens) |
| agrupamento de 4 n-grams de 4 elementos em sequência |
| 6) . além disso , (4 itens) |
| agrupamento de 1 n-grams de 4 elementos em sequência |
| 7) resgate urbano da natureza (...) inserir mais verde na (8 itens) |
| agrupamento de 2 n-grams de 4 elementos em sequência |
| 8) uma horta perto de casa (5 itens) |
| agrupamento de 2 n-grams de 4 elementos em sequência |
| 9) cadastrar todas as hortas (4 itens) |
| agrupamento de 1 n-grams de 4 elementos em sequência |
| 10) produção e venda de alimentos (5 itens) |
| agrupamento de 2 n-grams de 4 elementos em sequência |
| 11) o cadastramento de hortas (4 itens) |
| agrupamento de 1 n-grams de 4 elementos em sequência |
| 12) regiane nigro e claudia visoni (5 itens) |
| agrupamento de 2 n-grams de 4 elementos em sequência |
| 13) em nossa cidade . (4 itens) |

| |
|--|
| agrupamento de 1 n-grams de 4 elementos em sequência |
|--|

Fonte: Elaborado pela autora.

Ao analisar o Quadro 15, pode-se perceber que o maior trecho presente é o do grupo de informação número 1, com 52 itens. Ao comparar com o material de insumo, verifica-se que quase todo o parágrafo está presente no grupo 1, o que indica que foi a parte presente mais recuperada nos textos que receberam nota 5. É justificável a cópia ou paráfrase deste parágrafo, pois ele fala sobre o site Wikiversidade e como fazer o cadastro das hortas no site, que são informações essenciais para o cumprimento do propósito da tarefa. Neste trecho, também há a presença de alguns itens que não estão na sequência do parágrafo no material de insumo, o *que* na parte “tem *que* preencher o nome da horta”, como também as vírgulas (,) que foram usadas em alguns casos ao invés de ponto e vírgula (;). Isso pode ser um indício de que os examinandos que receberam nota 5 nos seus textos estão parafraseando esta parte do texto, o que mostra um domínio sobre a língua e faz sentido com esta tarefa já que as palavras do material de insumo são mais frequentes na língua, facilitando o uso de paráfrases.

Os grupos 2 e 3 são os próximos trechos mais longos, com 14 e 15 itens. Também é uma cópia justificável, pois o grupo 2 fala sobre um dos benefícios que o cadastro de hortas traz, que é estimular os produtores orgânicos a incluírem suas hortas no site para assim se tornarem mais conhecidos; o grupo 3 é referente a atrair mais pessoas para fazerem o cadastro de hortas no site e conhecerem esse projeto. Os demais grupos já são menores em extensão e dizem respeito ao cadastro de hortas (8, 9 e 11) em São Paulo (4 e 5), a ação de inserir mais verde nas cidades (7), a produção e venda de alimentos (10) e sobre as fundadoras do projeto (12). O grupo 13 é composto de apenas 1 *n-gram* e é um final de uma frase do material de insumo, por isso, não demonstra um sentido sozinha.

Os grupos 5, 6, 11 e 12 não estão presentes no material de insumo como uma sequência. As informações desses grupos estão presentes no material de insumo, mas não na forma da sequência que apareceu na lista de *n-grams*. Esse resultado parece dar mais indícios de que os examinandos parafraseiam algumas informações do material de insumo, usando algumas palavras-chaves pertencentes ao material para cumprir com os propósitos da tarefa.

No Quadro 16 estão os grupos de informação encontrados nas sequências de 5 itens em no mínimo 10% dos textos que receberam nota 5. O quadro na íntegra, com todos os grupos de *n-grams* encontrados e suas porcentagens, está presente no Anexo 4 ao final do trabalho.

Quadro 16 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n5 com 5 elementos em sequência

| |
|---|
| 1) da existência de alguma horta urbana e orgânica (...) entrar na plataforma wikiversidade e inserir os dados ali (...) tem <i>que</i> preencher o nome da horta (:), (,) se ela é educativa ou comercial (,) (;) a região em que está (,) (;) o bairro (,) (;) o endereço e (,) o link para contato . (50 itens) |
| agrupamento de 52 n-grams de 5 elementos em sequência |
| 2) atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade . (15 itens) |
| agrupamento de 11 n-grams de 5 elementos em sequência |
| 3) na cidade de São Paulo (5 itens) |
| agrupamento de 1 n-grams de 5 elementos em sequência |
| 4) estimular os produtores comerciais orgânicos a (...) e se tornarem mais conhecidos (11 itens) |
| agrupamento de 3 n-grams de 5 elementos em sequência |
| 5) produção e venda de alimentos (5 itens) |
| agrupamento de 1 n-grams de 5 elementos em sequência |
| 6) uma horta perto de casa (5 itens) |
| agrupamento de 1 n-grams de 5 elementos em sequência |
| 7) cadastro de hortas em são paulo (6 itens) |
| agrupamento de 2 n-grams de 5 elementos em sequência |
| 8) regiane nigro e claudia visoni (5 itens) |
| agrupamento de 1 n-grams de 5 elementos em sequência |

Fonte: Elaborado pela autora.

No Quadro 16, o maior trecho encontrado é o grupo de informação 1, com 50 itens. É o mesmo grupo de informação do quadro 15, o que era esperado, já que, como explicado anteriormente, é um parágrafo do material de insumo que se espera que seja copiado pelos examinandos por conter informações importantes para o cumprimento da tarefa. Também há a presença do *que* neste trecho e vírgulas (,) no lugar do ponto e vírgula (;), como ocorre no Quadro 15.

Os próximos grupos com maior número de itens são o 2 e 4, com 15 e 11 itens. Também são os mesmos trechos presentes no Quadro 15 que falam sobre incentivar mais pessoas a participarem do projeto e também sobre estimular os produtores orgânicos a se tornarem mais conhecidos através do projeto, porém são trechos menores, com algumas palavras a menos. Os demais grupos contêm menos itens e suas informações têm relação com

produção e venda de alimentos (5), a cidade de São Paulo (3), com as hortas (6) e seu cadastramento (7), e os nomes das fundadoras do projeto (8).

Como na análise do Quadro 15, neste quadro também há alguns grupos de informação que não estão presentes no material de insumo: o grupo 3 e o 8. São informações que aparecem no material de insumo, mas não estão escritas exatamente como aparece na lista de *n-grams*. Isso, como falado anteriormente, parece ser um sinal de que os examinandos que escreveram textos que receberam nota 5 parafrasearam algumas informações de modo que não ficaram exatamente iguais ao material de insumo.

Para finalizar as análises do corcel20162t3n5, no Quadro 17 estão presentes os grupos de informação encontrados com 6 itens em sequência. O quadro na íntegra, com todos os grupos de *n-grams* encontrados e suas porcentagens, está presente no Anexo 5 ao final do trabalho.

Quadro 17 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n5 com 6 elementos em sequência

| |
|--|
| 1) da existência de alguma horta urbana e orgânica (...) entrar na plataforma wikiversidade e inserir os dados (...) tem <i>que</i> preencher o nome da horta (,) (;) se ela é educativa ou comercial (,) (;) a região em que está (,) (;) o bairro (,) (;) o endereço e o link para contato . (48 itens) |
| agrupamento de 41 n-grams de 6 elementos em sequência |
| 2) atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade . (15 itens) |
| agrupamento de 9 n-grams de 6 elementos em sequência |
| 3) cadastro de hortas em são paulo (6 itens) |
| agrupamento de 1 n-grams de 6 elementos em sequência |

Fonte: Elaborado pela autora.

Diferentemente dos outros quadros, este apresenta um número bem menor de grupos de informação, apenas 3. Como nos outros quadros, este também apresenta o grupo 1 como o que contém o maior número de itens, 48. É o mesmo trecho das análises feitas anteriormente, que fala sobre o site Wikiversidade e como realizar o cadastro de hortas. Do Quadro 15 para esta, baixou em 4 o número de itens do grupo 1; do Quadro 16, baixou em 2 itens. O grupo 2 diz respeito a incentivar as pessoas a cultivarem suas hortas, e o grupo 3, o menor trecho, fala sobre o cadastro de hortas em São Paulo especificamente. Como já visto nos outros quadros, aqui também há indícios de paráfrases no grupo de informação 1, com o uso do *que* e das vírgulas (,).

Esses resultados parecem demonstrar que os textos do subcorpus corcel20162t3n5 recuperam informações do material de insumo e trechos de informação longos, mesmo que não seja de forma idêntica, já que apenas uma sequência de 4 elementos está presente em mais da metade dos textos (52%). Alguns grupos de informações não estavam presentes de forma idêntica no material de insumo, o que dá indícios de que os examinandos parafrazearam algumas informações. Como já discutido anteriormente, as palavras do material de insumo desta tarefa não são palavras específicas, o que facilita o parafraseamento delas pelos examinandos. Apesar disso, a maioria dos trechos encontrados na ferramenta *n-grams* se mostram necessários para o cumprimento da tarefa, como a citação do site Wikiversidade e como ele funciona, o cadastro de hortas e o incentivo de participação e conhecimento do projeto. Todos esses trechos foram encontrados nas sequências de 4, 5 e 6 elementos. Porém, alguns grupos de informação não parecem ser necessários para o cumprimento da tarefa, como os que falam de São Paulo, das fundadoras do projeto e também os trechos que ficaram demasiados curtos para se mostrarem relevantes para as análises.

Para verificar as diferenças entre os subcorpora, o olhar agora se volta para os textos que receberam nota 2. No Quadro 18, estão apresentados os grupos por informação encontrados a partir de sequências de 4 elementos. O quadro na íntegra com todos os grupos de *n-grams* encontrados e suas porcentagens está presente no Anexo 6 ao final do trabalho.

Quadro 18 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n2 com 4 elementos em sequência

| |
|--|
| 1) é atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade . (16 itens) |
| agrupamento de 10 n-grams de 4 elementos em sequência |
| 2) a inclusão de cultivos urbanos em todo o (8 itens) |
| agrupamento de 5 n-grams de 4 elementos em sequência |
| 3) existência de alguma horta urbana e (...) pode entrar na plataforma wikiversidade e inserir os dados ali (...) preencher o nome da horta , se ela é educativa ou comercial , a região em que está (...) o link para contato . (39 itens) |
| agrupamento de 21 n-grams de 4 elementos em sequência |
| 4) para estimular os produtores comerciais orgânicos a (...) e se tornarem mais conhecidos (12 itens) |
| agrupamento de 6 n-grams de 4 elementos em sequência |
| 5) de cadastrar todas as hortas (5 itens) |

| |
|---|
| agrupamento de 2 n-grams de 4 elementos em sequência |
| 6) pela dupla regiane nigro (...) em são paulo , (8 itens) |
| agrupamento de 2 n-grams de 4 elementos em sequência |
| 7) o cadastro de hortas em são paulo (7 itens) |
| agrupamento de 2 n-grams de 4 elementos em sequência |
| 8) resgate urbano da natureza (...) mais verde na selva de pedra (7 itens) |
| agrupamento de 3 n-grams de 4 elementos em sequência |
| 9) produção e venda de alimentos (5 itens) |
| agrupamento de 2 n-grams de 4 elementos em sequência |
| 10) cidade de são paulo (4 itens) |
| agrupamento de 1 n-grams de 4 elementos em sequência |

Fonte: Elaborado pela autora.

Os grupos de informação encontrados no corcel20162t3n2 com 4 elementos em sequência é menor do que os que foram encontrados no corcel20162t3n5 com 4 elementos em sequência, resultando em 3 grupos a menos. Como apresentado anteriormente, o número de sequências de *n-grams* encontradas nos textos de nota 2 é quase a metade dos encontrados nos textos de nota 5, por isso era esperada a redução dos grupos de informação.

Ao analisar o Quadro 18, verifica-se que o maior trecho encontrado nos textos de nota 2 é o grupo 3, com 39 itens. Esse trecho é o mesmo encontrado nos textos de nota 5 como o maior, que fala sobre o site Wikiversidade e como fazer o cadastro de hortas. O que chama a atenção neste trecho é que nos textos de nota 2 ele se encontra de forma mais incompleta, não tendo sido copiado na íntegra (as incompletudes são marcadas pelas reticências entre parênteses: (...)). Por exemplo, nos textos de nota 5 e no material de insumo da tarefa, aparece, em sequência com as informações de como fazer o cadastro no blog que precisa ser informado no site Wikiversidade, “*o bairro*” onde a horta está localizada. É uma informação relevante para o cumprimento da tarefa, mas que não está presente nos textos de nota 2.

Os grupos 1 e 4 são os próximos maiores, com 16 e 12 itens. O grupo 1 fala sobre atrair mais pessoas para o projeto, incentivando o cultivo orgânico; já o grupo 4 fala sobre incentivar os produtores orgânicos e os tornarem mais conhecidos, que é um dos benefícios do cadastramento de hortas. Os demais grupos são referentes à inclusão das hortas nas cidades (2), ao cadastro de hortas (5 e 7), à São Paulo especificamente (6, 7 e 10), o que parece ser um trecho não necessário para o cumprimento da tarefa, e também sobre uma das fundadoras do projeto (6), que também não parece ser um trecho relevante do material de

insumo. Também há grupos que falam sobre a necessidade do resgate urbano da natureza que as hortas podem promover (8), e a produção e venda de alimentos.

Todos os trechos encontrados fazem parte do material de insumo, apesar de algumas informações não serem tão relevantes para o cumprimento da tarefa, como os grupos que falam sobre a cidade de São Paulo especificamente e sobre apenas uma das fundadoras do projeto.

No Quadro 19, estão os resultados encontrados no corcel20162t3n2 com 5 elementos em sequência que aparecem em no mínimo 10% dos textos. O quadro na íntegra, com todos os grupos de *n-grams* encontrados e suas porcentagens, está presente no Anexo 7 ao final do trabalho.

Quadro 19 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n2 com 5 elementos em sequência

| |
|--|
| 1) a inclusão de cultivos urbanos em todo o (8 itens) |
| agrupamento de 4 n-grams de 5 elementos em sequência |
| 2) existência de alguma horta urbana e (...) pode entrar na plataforma wikiversidade e inserir os dados ali (...) preencher o nome da horta (...) se ela é educativa ou comercial (...) a região em que está (...) o link para contato . (33 itens) |
| agrupamento de 13 n-grams de 5 elementos em sequência |
| 3) é atrair mais pessoas interessadas em se tornarem (...) no cultivo orgânico em plena cidade (14 itens) |
| agrupamento de 6 n-grams de 5 elementos em sequência |
| 4) para estimular os produtores comerciais orgânicos (6 itens) |
| agrupamento de 2 n-grams de 5 elementos em sequência |
| 5) de cadastrar todas as hortas (5 itens) |
| agrupamento de 1 n-grams de 5 elementos em sequência |
| 6) produção e venda de alimentos (5 itens) |
| agrupamento de 1 n-grams de 5 elementos em sequência |

Fonte: Elaborado pela autora.

Diferentemente do Quadro 18, neste quadro há 6 grupos de informação, sendo o grupo 2 o maior entre todos, com 33 itens. Apesar de ser um trecho grande que pertence ao material de insumo, ele se mostra novamente fragmentado, ao passo que faltam alguns itens para ficar mais semelhante ao material de insumo, mais do que os apresentados no Quadro 18. O grupo 3 também é um trecho mais longo, com 14 itens, que tem relação com atrair mais pessoas

para conhecer o projeto e ajudar nesta causa. Os demais grupos são menores em extensão e por isso ficam em sua maioria incompletos de sentido, mas falam sobre a inclusão das hortas (1), os produtores orgânicos (4), do cadastramento (5) e da venda e produção de alimentos (6).

Ao comparar com o Quadro 18, estes resultados parecem trazer trechos do texto de insumo que são relevantes para o cumprimento da tarefa, apesar da sua demasiada fragmentação de informações, deixando incompletas as orientações sobre o projeto.

Para finalizar as análises do subcorpus de nota 2, foram verificados os resultados obtidos com 6 elementos em sequência que aparecem em no mínimo 10% dos textos. O quadro na íntegra, com todos os grupos de *n-grams* encontrados e suas porcentagens, está presente no Anexo 8 ao final do trabalho.

Quadro 20 - Extensão dos trechos agrupados por grupo de informação do corcel20162t3n5 com 6 elementos em sequência

| |
|---|
| 1) pode entrar na plataforma wikiversidade e inserir os dados ali (...) se ela é educativa ou comercial (16 itens) |
| agrupamento de 6 n-grams de 6 elementos em sequência |
| 2) atrair mais pessoas interessadas em se tornarem (...) no cultivo orgânico em plena cidade (13 itens) |
| agrupamento de 3 n-grams de 6 elementos em sequência |
| 3) a inclusão de cultivos urbanos em todo o (7 itens) |
| agrupamento de 3 n-grams de 6 elementos em sequência |

Fonte: Elaborado pela autora.

No Quadro 20, está a presença de apenas 3 grupos de informação, o menor número de grupos comparado com as demais Tabelas do subcorpus corcel20162t3n2. O trecho de maior extensão é o do grupo 1, com 16 itens. Esse trecho está demasiado fragmentado, tendo pouquíssimas informações importantes para o cumprimento da tarefa. O grupo 2 tem 13 itens e tem relação com incentivar as pessoas a participarem do projeto; e o grupo 3 conta com 7 itens e é referente à inclusão de hortas.

Ao comparar os três quadros referentes aos textos de nota 2, todos trouxeram três grupos de informação principais: sobre o site Wikiversidade e como funciona o cadastro, sobre atrair as pessoas para participarem do projeto, e referente à inclusão das hortas. São trechos do texto de insumo importantes para o cumprimento da tarefa e por isso são

justificáveis de serem copiadas pelos examinandos. Porém, estão demasiado fragmentados, o que reduz o sentido dos trechos encontrados.

Os resultados de ambos os subcorpora parecem demonstrar que há mais recuperação de informações nos textos de nota 5 do que nos textos de nota 2. Também evidenciam que os textos de nota 5 recuperam as informações de forma mais semelhante dos trechos do material de insumo, apesar da presença de alguns elementos de fora do material de insumo; já os textos de nota 2 recuperam as informações de forma mais fragmentada, o que pode indicar a falta de habilidade de selecionar as informações mais importantes do material de insumo que irão auxiliar no cumprimento do propósito da tarefa. Também há, em ambos os subcorpora, a presença de informações que não são relevantes para o cumprimento da tarefa, como quando são referidas as fundadoras do projeto (ou apenas uma fundadora), a cidade de São Paulo especificamente, e os trechos que ficaram muito curtos para se mostrarem relevantes para as análises.

Os resultados do subcorpus corcel20162t3n5 parecem indicar que há a presença de paráfrases dos examinandos, já que há grupos de informação que não estão presentes de forma idêntica no material de insumo e também a presença de elementos que foram adicionados nos trechos copiados do material de insumo. Esses resultados fortalecem a hipótese de que há o uso de paráfrases e sinônimos por parte dos examinandos, já que o material de insumo utiliza palavras que não são tão especializadas, auxiliando o examinando a utilizar palavras diferentes, mas que tem sentidos similares aos do material de insumo.

4.5 CONCORDANCE

Com os resultados obtidos nas ferramentas *keywords*, *wordlist* e *n-grams*, decidiu-se utilizar a ferramenta *concordance*, que mostra a palavra ou trecho escolhido em contexto, para comprovar ou refutar a hipótese do uso de paráfrases nos textos dos subcorpora. Buscando fazer um recorte que fosse relevante para a análise, foi escolhido o trecho com maior porcentagem de frequência encontrado nos *n-grams* de cada subcorpus. Ao observar o Anexo 3, verifica-se que o trecho com maior porcentagem do corcel20162t3n5 é: “o nome da horta”, com 4 elementos em sequência e 52% de frequência nos textos de nota 5.

Ao abrir a ferramenta, aparece junto com os resultados o número de textos que contém esse trecho, que é 104 textos, isto é, mais da metade dos textos do subcorpora. Foi observada a frase onde o trecho “o nome da horta” está inserido, tanto a parte da frase que vem antes (o *left context*) como a parte da frase que vem depois do trecho (o *right context*). O quadro na íntegra com todos os resultados obtidos na ferramenta *concordance* com o trecho

“o nome da horta” está presente no Anexo 9. No Quadro 21, estão presentes de forma resumida os resultados encontrados no *left context*.

Quadro 21 - Resultados encontrados no *left context* do corcel20162t3n5

| Palavra idêntica ao material de insumo: <i>preencher</i> | |
|---|---|
| Derivações da palavra <i>preencher</i> | “preenchendo”, “preencha”, “preenchido”, “preenche” e “preenchem” |
| Palavras diferentes do material de insumo | “colocar”, “indicar”, “inserir”, “inserindo”, “preencher lá”, “como”, “informação”, “devem inserir são/aportar são”, “inserir os dados ali”, “preencher o formulário:”, “é que tenha”, “só com” e “por exemplo” |

Fonte: elaborado pela autora.

Ao analisar o *left context*, percebeu-se que em 55 textos foi utilizada a palavra “preencher”, que é o termo utilizado no material de insumo no trecho ao qual “o nome da horta” pertence. Isso é esperado, já que é a palavra que está no material de insumo. O que é interessante é que algumas outras palavras foram utilizadas no lugar de “preencher”. No Quadro 20, verifica-se que uma parte dessas palavras utilizadas nos textos de nota 5 foram derivações da palavra “preencher”, como “preenchendo” e “preenchido”; outras são palavras diferentes, mas que cumprem com o papel de indicar para o leitor os dados que ele precisa preencher no site Wikiversidade, como “colocar”, “indicar” e também alguns trechos, como “devem inserir são/aportar são” e “inserir os dados ali”. Essas variações já apontam para o uso de paráfrases e também sinônimos por parte dos examinandos para recuperar trechos importantes do material de insumo, mas sem copiar de forma idêntica. Apesar de o material de insumo ser composto por palavras mais frequentes na língua e conseqüentemente ser mais fácil o parafraseamento das informações, essa habilidade também evidencia o nível de proficiência do examinando, sendo condizente com os textos que receberam nota 5.

Ao analisar o *right context*, focou-se mais em procurar as formas com que os examinandos escreveram a parte “se ela é educativa ou comercial”, pois é o trecho que segue “o nome da horta” no material de insumo. No Quadro 22, estão presentes de forma resumida os resultados encontrados no *right context*.

Quadro 22 - Resultados encontrados no *right context* do corcel20162t3n5

| Palavra idêntica ao material de insumo: <i>se ela é educativa ou comercial</i> |
|---|
|---|

| | |
|---|---|
| <p>Trechos semelhantes, mas com a adição ou subtração de um ou mais elementos</p> | <p>“se é educativa ou comercial”, “escrevem se a horta é educativa ou comercial”, “se ela é comercial ou não”, “precisar se ela é educativa ou comercial”, “se ela é educativa, quer dizer comunitária e escolar, ou comercial”, “especificar se ela é educativa (quer dizer “comunitária” e “escolar”) ou comercial”, “escrever se ela é educativa ou comercial”, “descrevendo se é comercial ou educativa”, “e identificar a sua característica se é educativa ou comercial”, “se ela é de caráter educativo (...) ou de caráter comercial (...)”, “se ela é de caráter “educativa ou comercial”, “se ela é educativa - ou seja, comunitária ou escolar - ou comercial”, “se a horta é educativa ou comercial”, “colocar se ela é educativa ou comercial”, “especifica se ela é educativa ou comercial”, “definir se ela é educativa ou comercial”, “se a mesma é educativa ou comercial”, “indicar se é uma horta educativa ou comercial”, “categorizar se ela é educativa ou comercial”, “o caráter, isto é, se ela é educativa ou comercial” e “se tem caráter educativo ou comercial”</p> |
| <p>Trechos diferentes do material de insumo</p> | <p>“qual é a finalidade da mesma”, “a sua categoria”, “a finalidade (educativa ou comercial)”, “a classificação da horta, educativa ou comercial”, “objetivo da horta”, “a sua natureza (educativa ou comercial)”, “se ela está disponível para fins comerciais”, “tipo”, “o tipo de horta”, “o tipo dela (se é comercial ou se é educativa)”, “o tipo ao que pertence, se ela é educativa ou talvez comercial”, “tipo (comercial ou educativa)”, “o caráter dela” e “e de que caráter é”.</p> |

Fonte: elaborado pela autora.

Foi encontrada a sequência idêntica ao material em 49 textos. Nos demais, foram encontrados alguns trechos bem parecidos, mas faltando ou acrescentando uma ou mais palavras, como “se é educativa ou comercial”, sem o “ela”; “escrevem se a horta é educativa ou comercial”, inserindo o “escrevem” e “horta”; “se ela é educativa, quer dizer comunitária e escolar, ou comercial”, inserindo a explicação do que seria uma horta educativa (explicação que está presente em outro parágrafo do material de insumo); e “se tem caráter educativo ou comercial”, incluindo “tem caráter” na frase.

Também foram encontrados trechos diferentes da frase que pertence ao material de insumo, como “qual é a finalidade da mesma”, “a sua categoria”, “a finalidade (educativa ou comercial)” e “o tipo de horta”. Apesar de serem diferentes da frase do material de insumo, são formas de falar sobre os tipos de hortas que o texto de insumo apresenta, não sendo necessário copiar exatamente o trecho “se ela é educativa ou comercial” para trazer esta informação importante do material de insumo. No entanto, foi evidenciado também trechos que não trazem a informação “se ela é educativa ou comercial” logo após o trecho “o nome da horta”, acabando a frase após este trecho principal ou trazendo outras informações relevantes que seguem no material de insumo, como *a região em que está, o bairro, o endereço e o link para contato*. Esses resultados evidenciam o uso de paráfrases e sinônimos tanto no *left context* como no *right context*, confirmando as hipóteses criadas anteriormente no *n-grams* e *wordlist*.

Ao observar o Anexo 6, verifica-se que o trecho com maior porcentagem do corcel20162t3n2 é: “inclusão de cultivos urbanos”, com 4 elementos em sequência e 21,50% de frequência nos textos de nota 2. O quadro na íntegra com todos os resultados obtidos na ferramenta *concordance* com o trecho “inclusão de cultivos urbanos” está presente no Anexo 10. No Quadro 23, estão presentes de forma resumida os resultados encontrados no *left context*.

Quadro 23 - Resultados encontrados no *left context* do corcel20162t3n2

| Palavra idêntica ao material de insumo: <i>permita a</i> | |
|--|---|
| Palavras semelhantes ao material de insumo | “permitão a”, “permite a”, “permitir a”, “permita”, “permite também a”, “permita-se a”, “permita la” e “permitir” |

| | |
|---|---|
| Palavras diferentes do material de insumo | “é a”, “projeto de”, “para a”, “a”, “fazendo a”, “lograr a”, “crear a”, “propósito a”, “motivar a”, “aceite a”, e “ajudar na” |
|---|---|

Fonte: elaborado pela autora.

Ao analisar o *left context*, percebeu-se que, diferente dos textos de nota 5, o trecho presente no material de insumo “permita a” que é anterior ao trecho “inclusão de cultivos urbanos” só está presente em três textos. Isso indica que foram muito mais frequentes outros elementos lexicais para referenciar a este trecho do material de insumo do que a forma idêntica. Entende-se que a palavra “permita” é a mais importante se comparada com o “a” na frase. Por isso, a divisão dos resultados se baseia na palavra “permita”. Ao verificar os resultados na íntegra, percebe-se que algumas palavras foram utilizadas no lugar de “permita a” que são semelhantes, como “permite a” e “permitir a”; outras foram palavras diferentes comparado ao material de insumo, como “projeto de” e “fazendo a”, termos que até mudam o sentido da frase original. Como nos textos de nota 5, esses resultados mostram o uso de paráfrases e sinônimos por parte dos examinandos para recuperar informações importantes do material de insumo sem necessariamente copiá-las de forma idêntica.

No entanto, diferentemente dos resultados do corcel20162t3n5, há a presença de inadequações ortográficas, como “propósito a”, “crear a” e “permita la”, sendo a primeira com uma inadequação no acento e as outras duas com influência da língua espanhola.

Ao analisar o *right context*, focou-se mais em procurar as formas com que os examinandos escreveram a parte “em todo o país”, pois é o trecho que segue “inclusão de cultivos urbanos” no material de insumo. No Quadro 24, estão presentes de forma resumida os resultados encontrados no *right context*.

Quadro 24 - Resultados encontrados no *right context* do corcel20162t3n2

| Palavra idêntica ao material de insumo: <i>em todo o país</i> | |
|--|--|
| Trechos semelhantes, mas com a adição ou subtração de um ou mais elementos | “em tudo o país”, “orgânicos em todo o país”, “pelo país”, “em toda a cidade”, “em todo Brasil” e “em todo país” |

| | |
|--|--|
| Trechos diferentes do material de insumo | “é importante também para estimular”, “na maioria ao cultivo organico”, “na busqueda disso”, “e rurais no país”, “animando bastante gente disposta a inserir mais verde”, “foi iniciado pela dupla Regiane Nigro e Claudia Visoni”, “nas cidades”, “que poderia ser implemetada neste cidade”, “na maioria do país” e “estimulando os produtos comerciais e organicos” |
|--|--|

Fonte: elaborado pela autora.

Foi encontrada a sequência idêntica ao material em 24 textos. Nos demais, foram encontrados alguns trechos semelhantes, mas faltando ou acrescentando uma ou mais palavras, como “em tudo o país”, utilizando “tudo” invés de “todo”; “orgânicos em todo o país”, acrescentando “orgânicos” no trecho; e “em todo Brasil”, usando “Brasil” invés de “o país”. Além disso, também foram encontrados trechos diferentes do material de insumo, mas que tem relação com o trecho “em todo o país”, como “e rurais no país” e “na maioria do país”. Esses resultados também confirmam a hipótese do uso de paráfrases e sinônimos por parte dos examinandos.

No entanto, foram evidenciados trechos que são diferentes do material e não trazem a informação “em todo o país” logo após o trecho “inclusão de cultivos urbanos”, acabando a frase após o trecho principal ou trazendo outras informações que, comparado ao material de insumo, não tem relação com o trecho principal, como “é importante também para estimular”, mudando totalmente a ideia inicial da frase; “animando bastante gente disposta a inserir mais verde”, incluindo um outro trecho que também está presente no material de insumo, mas não segue o trecho principal; e “foi iniciado pela dupla Regiane Nigro e Claudia Visoni”, que fala sobre as idealizadoras do projeto, mas também foge do esperado. Isso parece afirmar a hipótese de que os textos que receberam nota 2 recuperam as informações do material de insumo de forma fragmentada, misturando informações diferentes do material de insumo. Esses resultados parecem apontar para essa fragmentação de informações como uma característica dos textos de nota 2, já que não acontecem nos textos de nota 5. Outro aspecto que chama a atenção e que não acontece nos textos de nota 5 é a presença de inadequações ortográficas nos trechos copiados do material de insumo, como “orgânicos”. Como já visto nas análises das demais ferramentas, é uma característica dos textos de nota 2 apresentar um

número maior de inadequações ortográficas nos textos, o que pode alterar o resultado das análises quantitativas.

Esses resultados evidenciam então que há a presença do uso de paráfrases e sinônimos tanto no *left context* como no *right context* pelos textos de nota 2, porém também afirmam a presença de informações fragmentadas, misturando informações distintas, o que não acontece nos textos de nota 5.

5 DISCUSSÃO DOS RESULTADOS

As análises realizadas com os subcorpora corcel20162t3n2 e corcel20162t3n5 buscaram responder a pergunta de pesquisa: **Quais padrões lexicais caracterizam os textos avaliados com nota 2 e com nota 5 na Tarefa 3 da edição 2016-2 do Celpe-Bras?**

Como pudemos perceber, os resultados obtidos através das análises deste trabalho evidenciam padrões lexicais importantes para a caracterização das notas e mostram que os textos de nota 2 são, em média, mais curtos do que os textos de nota 5, além de apresentarem um número elevado de inadequações ortográficas. A pesquisa não tinha como objetivo verificar a presença de inadequações ortográficas nos textos dos examinandos, mas, por conta da falta de normalização da ortografia, essas inadequações foram evidenciadas e se mostram presentes com frequência nos textos de nota 2. Além disso, percebeu-se que os textos de nota 2 utilizam mais palavras provenientes do material de insumo. As paráfrases e sinônimos também estão presentes nos textos de nota 2, mas são organizados e articulados de forma fragmentada, deixando o texto confuso e sem coerência em determinados momentos. Já os textos de nota 5 fazem uso de paráfrases e sinônimos bem articulados nas produções escritas, utilizando um vocabulário mais diverso e rico. Diferentemente do esperado, nos textos de nota 5 há a recuperação de trechos maiores do material de insumo do que nos textos de nota 2, o que parece ser uma característica dos textos de nota 5 nesta tarefa. No entanto, também verificou-se a presença de palavras com inadequações ortográficas nos textos de nota 5, porém em menor quantidade comparado aos textos de nota 2.

Para auxiliar na resposta da pergunta norteadora, apresentamos, abaixo, as respostas às duas subperguntas de pesquisa, bem como discussões que buscam sistematizar os resultados apresentados anteriormente e colaborar para uma descrição de padrões lexicais característicos dos textos avaliados com nota 2 e com nota 5 na Tarefa 3 da edição de 2016-2 do Celpe-Bras.

Como a extensão e os mecanismos lexicais utilizados nos textos de nota 2 e nota 5 da Tarefa 3 da edição 2016-2 do Celpe-Bras são importantes para a caracterização dos níveis de proficiência?

Com base nas análises, pode-se concluir que os textos pertencentes ao corcel20162t3n5 são, em média, mais extensos do que os textos que pertencem ao corcel20162t3n2, apresentando maior número de *types*, *tokens*, sentenças, como também na média de *types*, *tokens* e sentenças por texto. Esse resultado está em conformidade com Divino (2021), Hanauer (2023) e Sostruznik (2023), apontando também que quanto maior o

nível de proficiência do examinando, potencialmente mais extensas serão suas produções escritas. Com este resultado, esperava-se que o percentual do *Type-Token-Ratio* (TTR) fosse maior nos textos de nota 5, porém não se chegou a este resultado.

O cálculo TTR apresentou-se maior nos textos do corcel20162t3n2, que obteve um percentual de 12,97%, maior do que o percentual obtido pelo corcel20162t3n5, de 10,51%. Esses percentuais mostram que os textos que receberam nota 2 apresentam uma riqueza lexical maior, ou seja, utilizam mais palavras diferentes em relação ao número total de palavras do subcorpus, do que os textos de nota 5. Esse resultado é contrário ao estudo de Divino (2021), mas está em conformidade com os trabalhos de Hanauer (2023) e Divino (2024), onde textos com notas mais baixas apresentaram um percentual de riqueza lexical maior do que os textos com notas maiores. A hipótese para esses resultados, que depois foi confirmada ao analisar os resultados da ferramenta *keywords*, é de que os textos menos proficientes apresentam um maior número de inadequações ortográficas e isso influencia no cálculo TTR, já que o *software* SkE computa cada palavra escrita de formas diferentes como um novo *type*, sendo vista pelo programa como palavras distintas. Como não foi realizada a normalização da ortografia nos textos utilizados nesta pesquisa, as inadequações ortográficas provavelmente alteraram os resultados do cálculo TTR. Em virtude disso, concordamos com Divino (2024) e Hanauer (2023) quando afirmam que o número de “palavras diferentes” de um corpus não indica, necessariamente, um maior repertório linguístico para cumprir os propósitos das tarefas.

Os dados analisados referentes às primeiras 15 *keywords* evidenciaram palavras relevantes para o cumprimento da tarefa, como “hortas”, que aparece em 100% dos textos de nota 5 e em 98,50% nos textos de nota 2; “horta”, no singular, que está presente em 93,50% dos textos de nota 5, mas apresenta um percentual bem mais baixo nos textos de nota 2, estando presente em 66% dos textos; “wikiversidade” também se mostra importante, pois aparece em 74,50% dos textos de nota 5, mas aparece pouco nos textos de nota 2, apenas em 43%; “cadastramento”, aparecendo em 65% dos textos de nota 5 e em 48,50% dos textos de nota 2. Estas palavras estão todas presentes no material de insumo da tarefa, e mostram-se relevantes para o cumprimento da tarefa a partir do seu alto percentual de aparecimento nos textos, mas também por trazerem informações importantes, como as palavras “hortas” e “horta”, que dizem respeito ao assunto principal da tarefa, e os termos “wikiversidade” e “cadastramento”, que estão relacionados com a plataforma onde devem ser cadastradas as hortas. Essas palavras relevantes e suas porcentagens também distinguem as notas dos textos, já que essas palavras têm mais frequência nos textos de nota 5 do que nos textos de nota 2.

Evidenciou-se também nesta lista palavras com inadequações ortográficas, como “ciudade”, “saudavel”, “organico”, “incluierem” e “importancia”. Todas apareceram no subcorpus corcel20162t3n2, o que vai ao encontro da pesquisa de Sirianni (2020), que afirma que os textos de nota 2 desta tarefa apresentam problemas de inadequação ortográfica, principalmente com acentos inadequados nas palavras, o que foi evidenciado neste trabalho, e falta de pontuação adequada nas frases. Porém, é importante perceber que essas palavras com inadequações ortográficas estão presentes em um percentual baixo de textos, e por isso se mostram relevantes em comparação com o corpus de referência, Portuguese Trends, porque não aparecem no corpus de referência. Caso a ortografia dessas palavras fosse normalizada, os resultados da ferramenta *keywords* poderiam ser diferentes e possivelmente mais relevantes para o estudo.

Ao verificar as palavras que não estão no material de insumo das 100 primeiras *keywords* dos dois subcorpora, os dados obtidos não se mostraram relevantes pelo baixo percentual de frequência nos textos dos subcorpora, tendo até algumas palavras com 1% de frequência. No entanto, o que chama a atenção, mesmo que não seja o foco do trabalho, foram as inadequações ortográficas que apareceram não só no corcel20162t3n2, mas também no corcel20162t3n5. Isso mostra que há a presença de inadequações ortográficas tanto nos textos de nota mais alta como nos textos de nota mais baixa. No entanto, o maior número de ocorrências está presente nos textos de notas mais baixas, corroborando Hanauer (2023) e Divino (2024).

Ainda sobre as inadequações ortográficas, os dados provenientes da ferramenta *wordlist* evidenciaram a presença de inadequações ortográficas nos textos de nota 5, mas um número baixo de ocorrências somente presente na frequência mínima de 5%, já que nas demais frequências mínimas (10%, 20% e 50%) não há a presença de inadequações ortográficas nesta nota. O corcel20162t3n2 manteve a maior porcentagem de palavras presentes no material de insumo em todas as frequências mínimas. Evidenciou-se a presença de palavras com inadequações ortográficas tanto fora do material de insumo como pertencentes ao material nas frequências de 5% e 10%, o que parece ser uma característica dos textos de nota 2. Porém, essas inadequações ortográficas não aparecem mais nas frequências de 20% e 50%, o que era esperado, pois as inadequações ortográficas, mesmo nos textos de nota 2, não acontecem de forma idêntica em todos os textos, não aparecendo em frequências altas como 20% e 50% dos textos.

Com os resultados obtidos da ferramenta *keywords*, se levantou a hipótese de que as palavras presentes no material de insumo da tarefa não fossem palavras tão específicas e, por

isso, não apareceriam tanto na ferramenta, dando maior visibilidade para as inadequações ortográficas. Essa hipótese foi confirmada ao analisar o material de insumo na ferramenta *keywords*, mantendo o corpus de referência como o Portuguese Trends. Para fazer um recorte, foram analisadas as primeiras 15 *keywords*, verificando sua frequência no corpus de estudo (corcel20162t3mi) e no corpus de referência (Portuguese Trends).

Os resultados evidenciaram que as palavras presentes no material de insumo da Tarefa 3 da edição 2016-2 são palavras de uso mais frequente na língua, confirmando a hipótese, já que as palavras que seriam consideradas mais típicas do corpus de estudo comparado com o corpus de referência apresentam um percentual alto de ocorrências no corpus de referência, como “hortas”, que aparece 1.226 vezes no corpus de referência. Esses resultados mostram-se diferentes do que foi encontrado em Divino (2024), onde a Tarefa 4 da edição de 2015-2 tem palavras mais específicas no seu material de insumo, dificultando a não cópia idêntica das palavras pelo examinando. Esse resultado vai ao encontro das Recomendações para Elaboração da Tarefa 3 do exame (não publicado), ao passo que solicita que as produções das Tarefas 3 de qualquer edição devem ter um registro menos formal e de complexidade menor, como “solicitar que o examinando, para cumprir a tarefa, estabeleça comparação entre dois textos, ou entre um texto e um gráfico ou imagem, ou que use parte das informações de cada texto para responder ao enunciado da tarefa”.

Com base nas análises, a ferramenta *wordlist* evidenciou que o número de *types* encontrados em no mínimo 5% dos textos é muito baixo nos dois subcorpora se comparado com o número total de *types* de cada subcorpus, isto é, a grande maioria das palavras utilizadas não se repete em nem 5% dos textos de nenhum dos subcorpora. Esses resultados indicam que os examinandos usaram, nesta tarefa, um vocabulário mais diversificado que não se repete muito nos diferentes textos, sendo elementos linguísticos mais subjetivos de cada texto. Esse uso de vocabulário mais diversificado por parte dos examinandos pode ter relação com o material de insumo da tarefa ser constituído de palavras mais frequentes na língua, não sendo necessário utilizarem nos seus textos exatamente a palavra presente no material de insumo. Evidenciou-se também que o número de *types* vai diminuindo nas análises à medida que o percentual mínimo de frequência vai aumentando.

Como os textos avaliados com as notas 2 e 5 na Tarefa 3 da edição 2016-2 do Celpe-Bras recuperam as informações do material de insumo?

O corcel20162t3n5 mostrou que as palavras com frequência em no mínimo 5% dos textos eram, em sua maioria, de fora do material de insumo, o que parece indicar o possível

parafraseamento de informações, como também com a utilização de um vocabulário mais diverso. Essa presença de palavras que não estão presentes no material de insumo vai ao encontro do que Mendel (2019) afirma quando diz que os textos de nota 5 “selecionam informações não explicitadas e recursos linguísticos não mobilizados no material de insumo, configurando autoria na recontextualização de informações pelo modo singular com que articulam o que dizer e como dizer.” (MENDEL, 2019, p. 102), e esse “modo singular com que articulam o que dizer e como dizer” pode ser o uso de paráfrases e sinônimos para recuperar informações do texto. Esse resultado mudou a medida que a frequência mínima de aparecimento nos textos aumenta, sendo as palavras mais usadas nos textos de nota 5 em no mínimo 10%, 20% e 50% dos textos que pertencem ao material de insumo, o que era esperado, pois é natural que as palavras do material de insumo se repitam em um maior número de textos produzidos pelos examinandos do que as palavras que não estão no material de insumo.

Os resultados da ferramenta *n-grams* evidenciam que há mais recuperação de informação semelhante ao material de insumo nos textos de nota 5 do que nos textos de nota 2. Isso é contrário à hipótese de Mendel (2017) que, ao analisar o enunciado e o texto de insumo da sua tarefa, previa demasiada cópia por parte dos examinandos, principalmente aqueles de notas mais baixas, o que não acontece, sendo exatamente o oposto.

No corcel20162t3n5 também se evidenciou a presença de elementos que foram adicionados nos trechos copiados do material de insumo, o que parece indicar novamente a presença de paráfrases e sinônimos nos textos de nota 5. Os resultados também apontam para uma recuperação de informações mais fragmentada nos textos de nota 2, o que pode indicar a falta de proficiência para selecionar as informações mais importantes do material de insumo que auxiliariam no cumprimento dos propósitos da tarefa, como também uma dificuldade de organizar de forma coerente o texto, concordando com os resultados de Sirianni (2020).

A resposta esperada da Tarefa 3 da edição 2016-2 (não publicado) visava três trechos importantes para estarem presentes nos textos que cumpririam com o propósito de *pedir a participação dos leitores no cadastramento de hortas urbanas, explicando o funcionamento e a utilidade das informações coletadas*. A partir dos dados obtidos, concluiu-se que tanto os textos de nota 5 como os textos de nota 2 recuperaram os três trechos importantes presentes na resposta esperada da tarefa. Essa presença da recuperação de informações nos textos de nota 2 corrobora com os resultados de Sirianni (2020) quando diz que “os textos avaliados com nota 2 souberam selecionar informações necessárias para cumprir os propósitos das tarefas” (SIRIANNI, 2020, p. 108), pois, apesar de ser um nível baixo, é o primeiro nível de

certificação do exame e é necessário recuperar minimamente as informações necessárias para cumprir a tarefa.

Além dessas informações relevantes para o propósito da tarefa, também há a presença de informações que não são relevantes para o cumprimento da tarefa, tanto nos textos de nota 2 como nos textos de nota 5, o que vai ao encontro da pesquisa de Scaramucci et al (2019) que evidenciou nas suas análises a predileção dos avaliadores por uma relação de interlocução bem feita à recuperação de informações chaves do material de insumo, o que abre espaço para os textos de nota 5 trazerem informações que não são relevantes para o cumprimento da tarefa.

Ao utilizar a ferramenta *concordance*, evidenciou-se, no trecho “o nome da horta” o uso de paráfrases e sinônimos no *left context* do corcel20162t3n5, utilizando derivações do verbo “preencher”, como “preenchendo”, e também palavras mais diferentes, mas que cumprem com o papel de indicar para o leitor os dados que ele precisa preencher no site Wikiversidade, como “colocar”. Esses resultados indicam uma habilidade que o examinando tem de recuperar trechos importantes do material de insumo, mas sem copiá-los de forma idêntica, o que corrobora os resultados de Mendel (2019), que afirma que os textos de nota 5 recuperam informações do material de insumo, mas também recontextualizam essas informações a partir de recursos linguísticos próprios.

No *right context*, também foi encontrado o uso de paráfrases e sinônimos, evidenciando textos em que faltava ou acrescentava uma ou mais palavras no trecho “se ela é educativa ou comercial”, por exemplo, “se ela tem caráter educativo ou comercial”, incluindo o “tem caráter” na frase. Além disso, também se evidenciou trechos diferentes da frase que pertence ao material de insumo, mas que tem relação com o sentido da frase, como “qual é a finalidade da mesma” e “a sua categoria”. No entanto, foi evidenciado também trechos que não trazem a informação “se ela é educativa ou comercial” logo após o trecho “o nome da horta”, acabando a frase após este trecho principal ou trazendo outras informações relevantes que seguem no material de insumo, como *a região em que está, o bairro, o endereço e o link para contato*. Esses resultados provenientes dos textos de nota 5 estão previstos na descrição dessa nota dos parâmetros de avaliação da parte escrita, visto que, na descrição da nota 5, fala-se sobre “recontextualizar apropriadamente e de maneira autoral as informações necessárias para cumprir o propósito” (BRASIL, 2020, p. 39), já que esta “maneira autoral” pode ser construída no texto com o uso de paráfrases e sinônimos.

No subcorpus corcel20162t3n2, o trecho analisado no *concordance* foi “inclusão de cultivos urbanos”, que também apresenta o uso de paráfrases e sinônimos, mas de uma forma

fragmentada, misturando algumas informações do material de insumo com outras que também estão no material de insumo, mas não têm relação juntas. No *left context*, há a presença de derivações do verbo “permita a”, como “permitir a”, mas também o uso de palavras diferentes comparado ao material de insumo que mudam o sentido da frase original, como “projeto de”. Além disso, diferentemente do corcel20162t3n5, há a presença de inadequações ortográficas que antecedem o trecho principal, como “propósito a” e “crear a”.

Ao analisar o *right context*, esperava-se encontrar o trecho “em todo o país” ou variações do mesmo, o que de fato aconteceu em alguns casos, sendo possível evidenciar trechos semelhantes, como “em tudo o país” e “em todo Brasil”. Também foram encontrados trechos diferentes do material de insumo, como “na maioria do país”. Porém, nos textos de nota 2 foram evidenciados trechos que acabam a frase logo após o trecho principal ou que fogem do esperado, não tendo relação com as informações do trecho principal, como “animado bastante gente disposta a inserir mais verde”, misturando duas informações que estão presentes no material de insumo, mas que não têm relação direta uma com a outra. Esses resultados provenientes das análises dos textos de nota 2 estão previstos na descrição desta nota nos parâmetros de avaliação da parte escrita, pois na descrição da nota 2 fala-se sobre a recontextualização de informações, afirmando que o examinando pode “não articular claramente essas informações” (BRASIL, 2020, p. 39), deixando o texto confuso para o leitor.

Esses resultados, tanto do *left* quanto do *right context*, confirmam a hipótese de que os textos que receberam nota 2 recuperam as informações do material de insumo de forma fragmentada, misturando informações diferentes do material de insumo, o que vem ao encontro dos resultados de Sirianni (2020), que, ao analisar a nota 2 desta tarefa, diz que “o texto apresenta trechos bastante confusos, em que as informações do texto de insumo e do próprio examinando se misturam e não apresenta uma organização textual adequada.” (p. 105), mostrando problemas na organização de forma coerente dos textos de nota 2. Mendel (2019) também disserta sobre isso ao verificar que textos que recebem nota 2 não recuperam informações suficientes do material de insumo para que faça sentido para o interlocutor, não articulando-as de forma clara, resultando em informações confusas.

Os resultados evidenciados nesta pesquisa são relevantes não só para a distinção das notas 2 e 5, mas também para contribuir para uma descrição mais robusta dos níveis de proficiência, pois traz aspectos relevantes para se evidenciar proficiência em língua portuguesa, como o uso de paráfrases e sinônimos para realizar a recuperação de informações do material de insumo. Articular estas informações no texto de forma clara e utilizando as

paráfrases e sinônimos mostra um nível de proficiência alto nesta tarefa, ainda mais que o material de insumo contém palavras mais frequentes da língua portuguesa, facilitando o processo de parafraseamento. Estes resultados comprovam que analisar textos pela LC pode contribuir para uma melhor descrição dos níveis e das notas, dando foco para aspectos importantes para a diferenciação das notas e, conseqüentemente, dos níveis de proficiência.

6 CONSIDERAÇÕES FINAIS

O presente trabalho buscou verificar características importantes que diferenciam as notas 2 e 5, analisando aspectos relevantes para o cumprimento da tarefa a partir de subcorpora compostos por produções escritas de examinandos do Celpe-Bras. Como já comprovado em pesquisas anteriores, a Linguística de Corpus se mostra relevante para estudar de forma mais aprimorada como se constituem os diferentes níveis de proficiência no exame Celpe-Bras, contribuindo para uma descrição cada vez mais ampla e específica para cada tarefa, além de contribuir com uma visão mais abrangente dos textos, sem ser necessário visitar um por um.

Com este trabalho, foi possível perceber padrões lexicais dos textos de nota 2 e dos textos de nota 5 na Tarefa 3 da edição 2016-2 do Celpe-Bras. Os textos pertencentes ao subcorpus corcel20162t3n5 são, em média, textos mais extensos, com pouca presença de inadequações ortográficas, e que utilizam paráfrases e sinônimos para recuperar as informações do texto de insumo, não sendo necessário copiá-las de forma idêntica, mas recontextualizando as informações de forma singular e autêntica.

Os textos pertencentes ao subcorpus corcel20162t3n2 se mostram diferentes, sendo, em média, menores em extensão, mas com o TTR maior em relação aos textos de nota 5. Porém, os resultados mostram que há um número alto de inadequações ortográficas neste subcorpus, o que pode ocasionar no aumento do TTR. Também há a presença do uso de paráfrases e sinônimos nos textos deste subcorpus, mas a recuperação de informações acontece de forma mais fragmentada e há uma mistura de informações que não apresentam relação direta entre si, não articulando os elementos de forma clara e causando confusão na construção do texto.

Além disso, neste trabalho também foi possível analisar o material de insumo, que consiste no enunciado da tarefa e no texto de insumo, verificando que as palavras que compõem este material são palavras mais frequentes na língua, o que facilita o parafraseamento das informações por parte dos examinandos e o uso de sinônimos. Isso vai ao encontro das Recomendações para Elaboração da Tarefa 3 do exame, que explicitam a preferência por um vocabulário menos complexo e propósitos mais claros que o examinando deve cumprir nestas tarefas.

Uma das limitações deste trabalho é que foram analisados apenas 200 textos que receberam nota 2 e 200 textos que receberam nota 5 na Tarefa 3 da edição 2016-2. Com certeza os resultados trazidos neste trabalho já contribuem para evidenciar aspectos

importantes das produções escritas dos examinandos nesta tarefa. No entanto, ainda assim é um recorte de análises que poderia ser ampliado.

Neste estudo foi utilizado o *Sketch Engine*, um *software* de análise de corpus com várias ferramentas que trazem resultados interessantes. Porém, há diversos outros *softwares* que podem ser utilizados para análises como esta, trazendo ferramentas semelhantes ao SkE ou diferentes, dando luz a outros resultados que podem ser relevantes para a descrição dos níveis. Além disso, outra limitação que tem relação com o *software* são as ferramentas utilizadas para as análises. Neste trabalho, não foram utilizadas todas as ferramentas do SkE, mas foram escolhidas aquelas que pareciam ser mais interessantes para a pesquisa. Utilizar outras ferramentas do SkE poderia proporcionar outras análises que contribuiriam para a diferenciação dos níveis de proficiência.

Os textos dos subcorpora utilizados nesta pesquisa, como já explicado anteriormente, não tiveram a ortografia normalizada. Isso limitou a pesquisa, ao passo que as inadequações ortográficas encontradas nos textos dos examinandos ficaram muito em evidência nas ferramentas e provavelmente distorceram alguns resultados que seriam importantes para a descrição dos níveis de proficiência.

A partir destas limitações, as sugestões de trabalhos futuros seriam, então, expandir as análises, trabalhando com as outras notas recebidas nesta tarefa (nota 0, 1, 3 e 4). Uma pesquisa que contemplasse todas as notas da Tarefa 3 da edição 2016-2 poderia contribuir de forma mais completa para a descrição dos níveis de proficiência avaliados no Celpe-Bras, ao passo que não estaria apenas buscando características importantes nos textos de nota 2 e 5, mas também nas demais notas. Aumentar também o número de textos analisados poderia proporcionar resultados mais válidos, pois não seriam analisados apenas uma parte dos textos de cada nota, mas todos os textos que receberam determinada nota, fortalecendo os argumentos acerca dos resultados encontrados.

Uma segunda sugestão seria a de testar outros *softwares*, como o AntConc, o SAS, ou o Coh-Metrix, dialogando com outros trabalhos que já foram feitos (EVERS, 2013; KUNRATH, 2019). Esses são só alguns exemplos dos diversos *softwares* que existem e que podem contribuir para expandir as análises do SkE, sendo feitos estudos comparativos com os resultados de cada *software* ou então utilizar mais de um para se complementar nas análises.

Uma terceira sugestão seria usar outras ferramentas do SkE, como o *Word Sketch*, que aponta semelhanças e diferenças entre o comportamento gramatical de termos específicos, e *Thesaurus*, que indicaria quais termos são sinônimos em cada subcorpus. Ao fazer isso,

poderia-se aumentar as análises, encontrando outros aspectos que são importantes para a diferenciação das notas.

Uma quarta sugestão seria analisar como os resultados obtidos nesta pesquisa poderiam contribuir para a preparação de aulas de cursos preparatórios ao exame Celpe-Bras, refletindo sobre formas de trabalhar as características linguísticas aqui encontradas. Outra sugestão seria realizar a normalização da ortografia e repetir as análises feitas neste trabalho. Neste estudo, como também em Hanauer (2023) e Divino (2024), discutiu-se sobre como as inadequações ortográficas distorcem os dados. Por isso, seria interessante utilizar textos com esta normalização ou comparar os textos normalizados com os não normalizados, verificando as diferenças que aparecem nos resultados.

Reconhecendo as limitações deste estudo e deixando algumas sugestões para futuras pesquisas, percebe-se que a LC se mostra relevante para contribuir com uma descrição mais profunda dos níveis de proficiência, como também dos parâmetros de avaliação da parte escrita do exame. Este trabalho dialoga com os trabalhos produzidos anteriormente (DIVINO, 2021; 2024; HANAUER, 2023; SOSTRUZNIK, 2023), além de aumentar o número de pesquisas feitas sobre os textos da Tarefa 3 da edição 2016-2 (MENDEL, 2017; SCARAMUCCI ET AL, 2019; SIRIANNI, 2020), e se coloca como endereço de enunciados futuros para que o Celpe-Bras continue sendo um exame de excelência e alta relevância na vida dos examinandos.

REFERÊNCIAS BIBLIOGRÁFICAS

BAKHTIN, M. **Estética da criação verbal**. São Paulo: Martins Fontes, 2003.

BARKER, F. **How can corpora be used in language testing?** In: O'KEEFFE, A.; MCCARTHY, M. (org). *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 2010, p. 633-645.

BRASIL. **Manual do Candidato do Exame CELPE-Bras**. Ministério da Educação (MEC). Brasília, 2006. Disponível em: <https://www.ufrgs.br/acervocelpebras/manuais/>

BRASIL. **Documento-base do exame Celpe-Bras**. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020. Disponível em: <http://www.ufrgs.br/acervocelpebras/arquivos/manuais/documento-base-do-exame-celpe-bras>

CALLIES, M.; GÖTZ, S. *Learner corpora in language testing and assessment: Prospects and challenges*. In: CALLIES, M.; GÖTZ, S.. (orgs.). **Learner Corpora in Language Testing and Assessment**. Amsterdam: Benjamins, 2015.

CUSHING, S. T.. *Corpus linguistics and language testing*. In: FULCHER, G.; HARDING, L. (ed.). **The Routledge Handbook of Language Testing**. 2. ed. Londres: Routledge, 2022. Cap. 32. p. 345-360.

DELL'IOLA, R. L. P.; TOSATTI, N. M. **Movimentos retóricos em produções escritas no exame celpe-bras**. Gragoatá, v. 20, n. 38, 30 jun. 2015.

DIVINO, L. S. **Índices lexicais de análise para a caracterização dos níveis intermediário e avançado superior no Exame Celpe-Bras**. 68f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2021.

DIVINO, L. S. **Contribuições da Linguística de Corpus para a definição de níveis de proficiência escrita no exame Celpe-Bras**. Dissertação (Mestrado em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2024.

EVERS, A. **Processamento de língua natural e níveis de proficiência do português: um estudo de produções textuais do exame Celpe-Bras**. 174f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

GABLASOVA, D. *Corpora for second language assessments*. In: WINKE, P.; BRUNFAUT, B. (Orgs.). **The Routledge Handbook of Second Language Acquisition and Language Testing**. Routledge, Londres, 2020.

GERALDI, J. W. *O professor como leitor do texto do aluno*. In: MARTINS, M. H. (Orgs.). **Questões de linguagem**. São Paulo: Contexto, 1991.

GOMES, M. S. **A complexidade de tarefas de leitura e produção escrita no exame CelpeBras**. 109f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

HANAUER, I. D. **Caracterização dos níveis Intermediário e Avançado Superior do exame Celpe-Bras em produções escritas de examinandos no gênero carta/e-mail: contribuições de uma análise guiada por corpus**. 84 f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

KENNEDY, G. **Corpus Linguistics**. International Encyclopedia Of The Social & Behavioral Sciences, [S.L.], p. 2816-2820, 2001. Elsevier. <http://dx.doi.org/10.1016/b0-08-043076-7/03056-4>.

KILGARRIFF, A.; RYCHLÝ, P.; SMRŽ, P.; TUGWELL, D. *The Sketch Engine*. In: **Proceedings of the XI EURALEX International Congress**. Université de Bretagne-Sud, p. 105–116, 2004.

KILGARRIFF, A.; BAISA, V.; BUŠTA, J.; JAKUBÍČEK, M.; KOVÁR, V.; MICHELFEIT, J.; RYCHLÝ, P.; SUCHOMEL, V. *The Sketch Engine: Ten Years On*. 54 In: **Lexicography ASIALEX**, Vol. 1, p. 7–36, 2014. Disponível em: <http://link.springer.com/article/10.1007/s40607-014-0009-9>.

KUNRATH, S. P. **Os descritores gerais e a progressão dos níveis de proficiência do Exame Celpe-Bras.** 198f. Tese (Doutorado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

MENDEL, K. **A avaliação integrada de leitura e escrita no exame Celpe-Bras.** 62f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2017.

MENDEL, K. **Proficiência e autoria na avaliação integrada de leitura e escrita do exame Celpe-Bras.** 177f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

RIO GRANDE DO SUL, Secretaria de Estado da Educação, Departamento Pedagógico, **Referenciais curriculares do Estado do Rio Grande do Sul.** Porto Alegre: SE/DP, 2009.

SARDINHA, T. B. **Lingüística de Corpus: histórico e problemática.** DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, v. 16, n. 2, p. 323–367, 2000.

SCARAMUCCI, M. V. R. **O exame Celpe-Bras em contexto hispanofalante: percepções de professores e candidatos.** In: WIEDEMANN, L.; SCARAMUCCI, M. V. R. (Orgs.). *Português para falantes de espanhol: ensino e aquisição.* Campinas: Pontes, 2008. p. 175-190.

SCARAMUCCI, M. V. R.; QUITZAU, L. A. **Integração leitura/escrita na prova de redação do vestibular Unicamp.** Revista Examen, Brasília, v. 2, n. 2, p. 28, jun. 2018.

SCHLATTER, M.; ALMEIDA, A. N.; FORTES, M. S.; SCHOFFEN, J. R. *Avaliação de desempenho e os conceitos de validade, confiabilidade e efeito retroativo.* In: NASCIMENTO, V. F.; NAUJORKS, J. C.; REBELLO, L. S.; SILVA, D. S. (orgs.). **A redação no contexto do vestibular 2005: a avaliação em perspectiva.** Porto Alegre: Universidade Federal do Rio Grande do Sul, 2005. p. 11-35.

SCHLATTER, M.; SCARAMUCCI, M. V. R.; PRATI, S.; ACUÑA, L. *Celpe-Bras e Celu: impactos da construção de parâmetros comuns de avaliação de proficiência em português e em espanhol*. In: FONTANA, M. Z. (org.) **O português do Brasil como língua transnacional**. Campinas: RG Editora, 2009.

SCHLATTER, M. **Celpe-Bras: avaliação, ensino e formação de professores de português como língua adicional**. 2014. Disponível em: <http://www.ufrgs.br/acervocelpebras/um-pouco-de-historia>. Acesso em: 08 de agosto de 2024.

SCHOFFEN, J.; STUMPF, E.; AMARAL, D.; DIVINO, L.; HANAUER, I.; LISBOA, I.; RAUPP, A.; XAVIER, B. **CorCel: a Brazilian Portuguese corpus of Celpe-Bras exam written texts**. Em preparação.

SCHOFFEN, J.; STUMPF, E.; AMARAL, D.; DIVINO, L.; HANAUER, I.; LISBOA, I.; RAUPP, A.; XAVIER, B. *Compilation and tagging of a corpus with Celpe-Bras texts*. In **Proceedings of the 16th International Conference on Computational Processing of Portuguese**. Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics, p. 627–632, 2024. Disponível em: <https://aclanthology.org/2024.propor-1.72.pdf>. Acesso em: 01 maio 2024.

SCHOFFEN, J.; DIVINO, L. **Contribuição da Linguística de Corpus para a formação de professores de PLA: sugestões a partir da descrição dos níveis de proficiência avaliados em uma tarefa da Parte Escrita do Celpe-Bras**. *Letras de Hoje*, [S. l.], v. 58, n. 1, p. e44904, 2023. DOI: 10.15448/1984-7726.2023.1.44904. Disponível em: <https://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/44904>. Acesso em: 20 maio 2024.

SCHOFFEN, J. R.. **Gêneros do discurso e parâmetros de avaliação de proficiência em português como língua estrangeira no exame Celpe-Bras**. Tese (Doutorado em Linguística Aplicada) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

SCHOFFEN, J.; GOMES, M.; SCHLATTER, M. *Tarefas de leitura e produção de texto com base na noção bakhtiniana de gêneros do discurso*. In: SILVA, K. A.; TORRES, D. (orgs.). **Português como língua (inter)nacional: faces e interfaces**. Campinas: Pontes, 2013.

SCHOFFEN, J.; SCHLATTER, M.; KUNRATH, S. P.; NAGASAWA, E. Y.; SIRIANNI, G. R.; MENDEL, K.; TRUYLLIO, L. R.; DIVINO, L. **Estudo descritivo das tarefas da parte escrita do exame Celpe-Bras**. Porto Alegre: Instituto de Letras - Ufrgs, 2018. Disponível em: <https://www.ufrgs.br/acervocelpebras/wp-content/uploads/2021/12/Schoffen-et-al-2018.pdf>. Acesso em: 31 maio 2024.

SHOHAMY, E. **The Power of Tests: A Critical Perspective on the Uses of Language Tests**. 2001. Disponível em: https://www.researchgate.net/publication/240324351_The_Power_of_Tests_A_Critical_Perspective_on_the_Use_of_Language_Tests. Acesso em: 17 de abril de 2024.

SIDI, W. A. **Níveis de proficiência em leitura e escrita de falantes de espanhol no exame CELPE-Bras**. 73f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

SIMÕES, L.; RAMOS, J. W.; MARCHI, D.; FILIPOUSKI, A. M. **Leitura e autoria: planejamento em língua portuguesa e literatura**. Erechim, RS: Edelbra, 2012.

SIRIANNI, G. R. **Descrição dos níveis de proficiência em tarefa de leitura e escrita a partir de produções textuais de alunos do curso Preparatório Celpe-Bras**. 76f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

SIRIANNI, G. R. **Entre a certificação e a não certificação no Celpe-Bras: um estudo sobre os níveis de proficiência na Parte Escrita do exame**. 139f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2020.

SOSTRUZNIK, J. L. **O uso de conjunções em produções escritas no exame Celpe-Bras: um estudo baseado em corpus**. 76f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

SOUZA, G. T. **Introdução à teoria do enunciado concreto do círculo Bakhtin/Volochinov/Medvedev.** São Paulo: Numanitas/FFLCH/USP, 2002.

TOSATTI, N. M.. **O desempenho de estudantes de Países Africanos de Língua Oficial Portuguesa no Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras).** 260f. Tese (Doutorado em Linguística Aplicada) – Universidade Federal de Minas Gerais, Belo Horizonte, 2021.

VICENTINI, M. P. **As dimensões do construto compreensão oral para produção escrita no exame celpe-bras: percepções, processos, estratégias e desempenhos de examinandos.** 241f. Tese (Doutorado em Linguística Aplicada) – Universidade Estadual de Campinas, Campinas, 2022.

Anexo 1 – Tarefa 3 de 2016-2



Certificado de Proficiência em Língua Portuguesa para Estrangeiros

Tarefa 3 | Cadastro de hortas

Página 6

Você tem um *blog* sobre alimentação saudável e, depois de ler a notícia sobre o cadastro de hortas em São Paulo, decidiu implementar essa iniciativa em sua cidade. Escreva um texto em seu *blog* para pedir a participação dos leitores no cadastramento, explicando o funcionamento e a utilidade das informações coletadas.

À procura de hortas em plena São Paulo



Até esta manhã, eram 25 hortas cadastradas, sendo 20 de caráter “educativo” – ou seja, comunitárias e escolares – e 5 “comerciais” – quer dizer, voltadas à produção e venda de alimentos. O que elas têm em comum? Estão todas dentro do município de São Paulo. Várias – olhem só! – em densas áreas urbanas, como nos bairros de Vila Madalena, Vila Mariana, Pompeia, Mooca, Cambuci, Casa Verde, Sapopemba, Vila Nova Esperança, Jardim Ângela...

Já as “comerciais” inserem-se sobretudo na zona rural paulistana. Desde 2014 a cidade de São Paulo “ganhou” uma área rural, ou melhor, porções do território paulistano situadas principalmente nos extremos sul e leste da capital passaram a ser reconhecidas como “rurais”.

Esses dados vêm de uma pesquisa colaborativa em pleno andamento. Quem souber da existência de alguma horta urbana e orgânica sendo cultivada em algum “rincão”

da metrópole pode entrar na plataforma Wikiversidade e inserir os dados ali. Tem de preencher o nome da horta; se ela é educativa ou comercial; a região em que está; o bairro; o endereço e, por fim, o *link* para contato.

O cadastramento dessas hortas urbanas foi iniciado pela dupla Regiane Nigro, que trabalha no Instituto Kairós, em São Paulo, e a jornalista e uma das fundadoras do grupo Hortelões Urbanos, Claudia Visoni. Ambas têm forte atuação em duas dessas hortas comunitárias, a da Praça das Corujas, na Vila Madalena, e também a Horta do Ciclista, na Praça do Ciclista, região da Avenida Paulista. Regiane atua também, pelo Kairós, com os produtores da Associação de Agricultores da Zona Leste, que contempla bairros do extremo leste paulistano, como São Mateus, Guaianases, Cidade Tiradentes e São Miguel Paulista.

A ideia de cadastrar todas as hortas paulistanas – embora a plataforma também

permita a inclusão de cultivos urbanos em todo o país – já era antiga, conta ao blog Regiane Nigro. E foi finalmente posta em prática há cerca de 15 dias, para servir como base de uma reunião que Claudia Visoni teria com pessoas de outros países interessadas no tema hortas urbanas. “Foi este o motivador para o projeto começar a andar”, conta Regiane.

Se inicialmente, há cerca de cinco anos, quando a “onda” do resgate urbano da natureza começou aqui em São Paulo, as hortas cultivadas em plena metrópole eram raras, agora pode-se dizer que elas já têm animado bastante gente disposta a inserir mais verde na selva de pedra. Na maioria dessas hortas, dá-se preferência ao cultivo orgânico.

Regiane conta que a plataforma é importante também para estimular os produtores comerciais orgânicos a incluírem suas hortas ali e se tornarem mais conhecidos – afinal, muitos deles fornecem, em sistema de compra coletiva, alimentos para a metrópole, entre outras formas de comercialização.

Atualmente, nos “rincões” paulistanos, um dos grupos mais organizados de agricultores – e que já estão lá no cadastro – é o da Cooperapas, no extremo sul da capital, região de Parelheiros. Outra grande utilidade deste cadastro é atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade. Claudia Visoni, ao chamar, num grupo do Facebook, as pessoas a colaborarem com o cadastramento das hortas, comenta que o registro é “importante para estudos acadêmicos, para os voluntários acharem uma horta perto de casa, para possíveis apoiadores (dos agricultores profissionais) e também para quem não está no Facebook”. E continua: “Quem puder, ajude a incluir as hortas que ainda não estão na tabela. Quem quiser criar tabelas para outras cidades, é só fazer. A plataforma Wiki é totalmente colaborativa e fácil de editar”.

Anexo 2 - Lista de *keywords* do corcel20162t3mi

| Itens | Frequency (focus) | Frequency (reference) | DOCF (reference) | Itens | Frequency (focus) | Frequency (reference) | DOCF (reference) |
|---------------|-------------------|-----------------------|------------------|------------|-------------------|-----------------------|------------------|
| hortas | 12 | 1226 | 727 | servir | 1 | 17606 | 15903 |
| visoni | 3 | 6 | 3 | pedra | 1 | 17675 | 10956 |
| regiane | 5 | 691 | 426 | vila | 4 | 73585 | 38999 |
| kairós | 2 | 156 | 74 | produtores | 2 | 37253 | 22291 |
| cooperapas | 1 | 0 | 0 | comenta | 1 | 18332 | 15906 |
| wikiversidade | 1 | 0 | 0 | dessas | 3 | 57326 | 49308 |
| hortelões | 1 | 4 | 4 | praça | 2 | 39631 | 22515 |
| nigro | 2 | 842 | 397 | chamar | 1 | 19903 | 18026 |
| rincões | 1 | 145 | 139 | rurais | 1 | 20181 | 13493 |
| cadastramento | 3 | 1835 | 1353 | endereço | 1 | 20345 | 14165 |
| inserem-se | 1 | 167 | 166 | atrair | 1 | 20349 | 16695 |
| colaborarem | 1 | 228 | 223 | mariana | 1 | 20902 | 11397 |
| wiki | 1 | 238 | 162 | caráter | 1 | 21471 | 17851 |
| corujas | 1 | 239 | 144 | urbana | 1 | 22203 | 15208 |
| cambuci | 1 | 245 | 187 | andamento | 1 | 22868 | 20430 |
| incluïrem | 1 | 284 | 281 | vêm | 1 | 23252 | 21169 |
| metrópole | 3 | 2253 | 1840 | conhecidos | 1 | 23828 | 21586 |
| rincão | 1 | 324 | 186 | onda | 1 | 24696 | 18046 |
| paulistanas | 1 | 493 | 296 | incluir | 1 | 25097 | 21425 |
| guaianases | 1 | 560 | 332 | andar | 1 | 25325 | 19218 |
| sapopemba | 1 | 669 | 356 | ali | 2 | 52359 | 37616 |
| horta | 4 | 5054 | 2995 | atua | 1 | 27631 | 24516 |
| motivador | 1 | 764 | 711 | ler | 1 | 28168 | 22181 |
| fundadoras | 1 | 767 | 670 | resgate | 1 | 28209 | 18435 |
| ciclista | 2 | 2239 | 1360 | quiser | 1 | 28428 | 23382 |
| acharem | 1 | 815 | 790 | ambas | 1 | 29358 | 26402 |
| densas | 1 | 816 | 749 | link | 1 | 29802 | 22428 |
| dá-se | 1 | 856 | 828 | saudável | 1 | 30657 | 21528 |
| colaborativa | 2 | 2466 | 2057 | existência | 1 | 30833 | 25963 |
| parceiros | 1 | 951 | 542 | finalmente | 1 | 31460 | 27714 |
| cultivos | 1 | 1001 | 755 | plataforma | 4 | 129232 | 73453 |
| pompeia | 1 | 1051 | 477 | esperança | 1 | 32275 | 25750 |

| | | | | | | | |
|--------------|---|-------|-------|---------------|---|--------|--------|
| olhem | 1 | 1089 | 998 | passaram | 1 | 32439 | 29503 |
| cultivadas | 1 | 1099 | 929 | verde | 2 | 66016 | 38925 |
| mooça | 1 | 1113 | 670 | tabela | 1 | 34063 | 24755 |
| paulistano | 2 | 3197 | 2609 | jardim | 1 | 34089 | 19994 |
| souber | 1 | 1316 | 1274 | antiga | 1 | 34564 | 27361 |
| comunitárias | 2 | 3369 | 2559 | alimentos | 2 | 71263 | 35151 |
| situadas | 1 | 1397 | 1328 | trabalha | 1 | 35472 | 31282 |
| tornarem | 2 | 3527 | 3385 | inicialmente | 1 | 35740 | 33319 |
| madalena | 2 | 3540 | 2198 | registro | 1 | 35978 | 26517 |
| paulistanos | 1 | 1458 | 1221 | inclusão | 1 | 36964 | 26135 |
| voluntárias | 1 | 1512 | 1251 | paulista | 2 | 78971 | 52295 |
| escreva | 1 | 1574 | 1276 | eram | 2 | 81652 | 61438 |
| orgânico | 2 | 3915 | 2806 | alimentação | 1 | 41495 | 27025 |
| cultivada | 1 | 1649 | 1131 | coletiva | 1 | 41785 | 33836 |
| utilidade | 2 | 4124 | 3583 | pedir | 1 | 41855 | 34738 |
| claudia | 3 | 6651 | 3902 | funcionamento | 1 | 41872 | 32458 |
| ângela | 1 | 1817 | 1070 | avenida | 1 | 44425 | 24590 |
| urbanas | 3 | 7310 | 5625 | dupla | 1 | 44710 | 33014 |
| cadastradas | 1 | 2408 | 2120 | sobretudo | 1 | 46430 | 38927 |
| coletadas | 1 | 2661 | 2438 | comercial | 1 | 46638 | 33710 |
| interessadas | 2 | 6062 | 5446 | possíveis | 1 | 47528 | 41831 |
| paulistana | 1 | 2977 | 2547 | miguel | 1 | 47705 | 26708 |
| educativa | 1 | 2994 | 2444 | afinal | 1 | 49318 | 38995 |
| cultivo | 2 | 6905 | 4510 | procura | 1 | 50738 | 39489 |
| inserir | 2 | 7064 | 6057 | totalmente | 1 | 52561 | 45686 |
| Tabelas | 1 | 3206 | 2192 | formas | 1 | 53661 | 42373 |
| posta | 1 | 3365 | 3115 | zona | 2 | 108447 | 75855 |
| porções | 1 | 3404 | 2641 | decidiu | 1 | 54430 | 48372 |
| disposta | 1 | 3583 | 3380 | elas | 2 | 110706 | 81889 |
| educativo | 1 | 3676 | 2855 | dizer | 2 | 112950 | 87726 |
| reconhecidas | 1 | 3870 | 3644 | ganhou | 1 | 58262 | 49825 |
| plena | 3 | 13218 | 11746 | natureza | 1 | 62949 | 41970 |
| tiradentes | 1 | 3991 | 1986 | estudos | 1 | 63025 | 42549 |
| ajude | 1 | 4060 | 3830 | bairro | 1 | 63193 | 39839 |
| orgânicos | 1 | 4146 | 2750 | paulo | 6 | 400514 | 204916 |

| | | | | | | | |
|-----------------|---|-------|-------|----------------|---|--------|--------|
| editar | 1 | 4149 | 2912 | começar | 1 | 67388 | 56281 |
| urbanos | 2 | 9210 | 7224 | jornalista | 1 | 67424 | 46255 |
| selva | 1 | 4617 | 2999 | fácil | 1 | 69254 | 54717 |
| fornece | 1 | 4755 | 4395 | contato | 1 | 69510 | 52199 |
| organizados | 1 | 4988 | 4569 | atuação | 1 | 71158 | 54219 |
| agricultores | 3 | 16876 | 8294 | território | 1 | 72765 | 49094 |
| acadêmicos | 1 | 5279 | 4082 | comum | 1 | 73050 | 58545 |
| puder | 1 | 5855 | 5379 | notícia | 1 | 80973 | 72962 |
| cadastrar | 1 | 5979 | 4446 | capital | 2 | 164126 | 109948 |
| explicando | 1 | 6251 | 6077 | cidade | 4 | 329976 | 163465 |
| pode-se | 1 | 6253 | 5430 | venda | 1 | 83698 | 55708 |
| orgânica | 1 | 6457 | 4603 | alguma | 1 | 84455 | 69150 |
| animado | 1 | 6993 | 6255 | perto | 1 | 84881 | 68838 |
| extremo | 2 | 14738 | 11794 | iniciativa | 1 | 85885 | 64711 |
| permita | 1 | 7171 | 6721 | associação | 1 | 87904 | 57855 |
| contempla | 1 | 7656 | 7175 | entrar | 1 | 88114 | 73217 |
| blog | 3 | 24609 | 18206 | grupos | 1 | 88140 | 61103 |
| raras | 1 | 9156 | 5360 | compra | 1 | 89935 | 56090 |
| mateus | 1 | 9192 | 5689 | prática | 1 | 89961 | 68130 |
| extremos | 1 | 9221 | 6950 | algum | 1 | 90098 | 73984 |
| comerciais | 3 | 29794 | 23202 | continua | 1 | 92544 | 78113 |
| iniciado | 1 | 9715 | 9340 | deles | 1 | 92857 | 79315 |
| leste | 3 | 31249 | 22391 | têm | 3 | 283099 | 200262 |
| escolares | 1 | 10218 | 7133 | município | 1 | 94074 | 59887 |
| bairros | 2 | 21454 | 13709 | região | 3 | 289149 | 160975 |
| voltadas | 1 | 10489 | 9525 | idades | 1 | 97370 | 60343 |
| apoiadores | 1 | 10649 | 8390 | tema | 1 | 99693 | 73221 |
| cadastro | 3 | 35777 | 20063 | quem | 4 | 419222 | 260419 |
| comercialização | 1 | 11661 | 9233 | principalmente | 1 | 104373 | 86752 |
| facebook | 2 | 24940 | 17057 | manhã | 1 | 107005 | 82933 |
| preencher | 1 | 12569 | 10845 | reunião | 1 | 109099 | 69075 |
| leitores | 1 | 13170 | 8892 | instituto | 1 | 109805 | 75442 |
| estimular | 1 | 13527 | 11756 | ideia | 1 | 110172 | 85294 |
| pleno | 1 | 14495 | 12859 | teria | 1 | 111440 | 81412 |
| rural | 2 | 31350 | 19254 | forte | 1 | 112486 | 87906 |

| | | | | | | | |
|--------------|---|---------|--------|---------------|----|----------|---------|
| implementar | 1 | 15362 | 13459 | texto | 1 | 114401 | 70227 |
| urbano | 1 | 15618 | 11567 | bastante | 1 | 114995 | 87066 |
| voluntários | 1 | 15826 | 10635 | criar | 1 | 116477 | 83460 |
| preferência | 1 | 16694 | 14445 | profissionais | 1 | 118898 | 70998 |
| participação | 1 | 121374 | 91895 | sul | 2 | 242083 | 131876 |
| maioria | 1 | 122469 | 92595 | este | 1 | 438710 | 281811 |
| atualmente | 1 | 123280 | 105061 | desde | 1 | 451938 | 323129 |
| todas | 2 | 251854 | 190445 | outros | 1 | 455741 | 316596 |
| conta | 3 | 378409 | 247993 | país | 1 | 460475 | 256364 |
| cerca | 2 | 259538 | 185662 | essa | 1 | 476205 | 289763 |
| pesquisa | 1 | 132773 | 71226 | em | 18 | 8663079 | 1451529 |
| começou | 1 | 133430 | 111513 | ou | 3 | 1444935 | 545456 |
| dados | 2 | 270866 | 141196 | das | 4 | 1931555 | 805296 |
| importante | 2 | 272409 | 186071 | depois | 1 | 485951 | 325652 |
| várias | 1 | 137947 | 112906 | ela | 1 | 486432 | 225248 |
| embora | 1 | 140006 | 110647 | grande | 1 | 498971 | 315861 |
| lá | 1 | 142303 | 98554 | fazer | 1 | 521601 | 309240 |
| áreas | 1 | 143727 | 92696 | você | 1 | 575396 | 179679 |
| outras | 2 | 311316 | 231114 | sobre | 2 | 1152305 | 578332 |
| esses | 1 | 159314 | 119465 | está | 2 | 1181914 | 641958 |
| grupo | 2 | 325188 | 188067 | tem | 2 | 1205564 | 598774 |
| base | 1 | 166097 | 120200 | como | 4 | 2696011 | 921705 |
| nome | 1 | 171830 | 119995 | as | 5 | 3435010 | 1047824 |
| sendo | 2 | 353964 | 262752 | foi | 3 | 2070691 | 854192 |
| quer | 1 | 177178 | 125724 | se | 4 | 2784523 | 931725 |
| estão | 3 | 543327 | 357080 | quando | 1 | 709045 | 415251 |
| cinco | 1 | 182346 | 140176 | na | 7 | 4978628 | 1310737 |
| são | 8 | 1501716 | 639653 | é | 7 | 5022815 | 1118443 |
| outra | 1 | 191547 | 152503 | para | 11 | 7957731 | 1395011 |
| muitos | 1 | 192200 | 144834 | mais | 4 | 3087156 | 966221 |
| países | 1 | 193789 | 102553 | da | 11 | 9080492 | 1445527 |
| casa | 2 | 394151 | 223908 | e | 19 | 16101513 | 1546679 |
| produção | 1 | 198749 | 106684 | à | 2 | 1745213 | 785501 |
| num | 1 | 209164 | 149779 | no | 7 | 6118757 | 1375790 |
| só | 2 | 423239 | 275306 | ao | 3 | 2634351 | 1028685 |

| | | | | | | | |
|-------------|---|---------|--------|-------|----|----------|---------|
| dentro | 1 | 212546 | 157968 | pode | 1 | 883926 | 417533 |
| pessoas | 3 | 646445 | 328676 | dos | 3 | 2845017 | 983023 |
| gente | 1 | 230469 | 114152 | a | 21 | 20531354 | 1565129 |
| área | 1 | 233711 | 139372 | uma | 5 | 5070057 | 1224444 |
| informações | 1 | 236375 | 160210 | anos | 1 | 1022165 | 507026 |
| deste | 1 | 238191 | 193969 | entre | 1 | 1039752 | 571875 |
| sistema | 1 | 243001 | 136572 | seu | 1 | 1040937 | 520193 |
| projeto | 1 | 275165 | 133665 | até | 1 | 1051971 | 574775 |
| aqui | 1 | 275761 | 171412 | ainda | 1 | 1070811 | 623040 |
| fim | 1 | 283030 | 212233 | de | 25 | 29240183 | 1608477 |
| também | 5 | 1527531 | 766945 | os | 4 | 4899080 | 1204610 |
| já | 4 | 1263291 | 680122 | pela | 1 | 1229925 | 686609 |
| duas | 1 | 316811 | 232631 | sua | 1 | 1249099 | 592724 |
| todo | 1 | 326632 | 242142 | o | 14 | 18433508 | 1548375 |
| há | 2 | 662291 | 394506 | pelo | 1 | 1346512 | 724932 |
| melhor | 1 | 341546 | 212425 | que | 10 | 13492623 | 1474751 |
| nos | 3 | 1096947 | 594760 | do | 8 | 10826497 | 1482875 |
| seja | 1 | 380668 | 254443 | ser | 1 | 1456130 | 685950 |
| nova | 1 | 386184 | 243430 | não | 2 | 3639428 | 992126 |
| agora | 1 | 400464 | 288209 | um | 3 | 5626514 | 1259804 |
| esta | 1 | 404055 | 271522 | com | 3 | 6626196 | 1372095 |
| dias | 1 | 404308 | 264292 | por | 1 | 3670061 | 1152639 |
| era | 1 | 409497 | 237634 | suas | 1 | 429167 | 272708 |

Anexo 3 - Resultado na íntegra da ferramenta *n-grams* com 4 elementos em sequência que aparecem no mínimo em 10% no corcel20162t3n5

| N-grams encontrados no corcel20162t3n5 | DOCF | Relative DOCF (%) |
|---|------|-------------------|
| <p>1) souber da existência de alguma horta urbana e orgânica (...) pode entrar na plataforma wikiversidade e inserir os dados ali . (...) tem que preencher o nome da horta (;) (;) se ela é educativa ou comercial (;) (;) a região em que está (;) (;) o bairro (;) (;) o endereço e (;) o link para contato (52 itens)</p> | | |

| | | |
|------------------------------------|-----|------|
| o nome da horta | 104 | 52 |
| nome da horta , | 82 | 41 |
| link para contato . | 79 | 39,5 |
| é educativa ou comercial | 76 | 38 |
| o link para contato | 76 | 38 |
| se ela é educativa | 73 | 36,5 |
| ela é educativa ou | 64 | 32 |
| preencher o nome da | 55 | 27,5 |
| a região em que | 53 | 26,5 |
| educativa ou comercial , | 51 | 25,5 |
| região em que está | 47 | 23,5 |
| e inserir os dados | 47 | 23,5 |
| e o link para | 45 | 22,5 |
| da horta , se | 44 | 22 |
| na plataforma wikiversidade e | 44 | 22 |
| endereço e o link | 41 | 20,5 |
| , o bairro , | 40 | 20 |
| bairro , o endereço | 39 | 19,5 |
| o bairro , o | 39 | 19,5 |
| , se ela é | 39 | 19,5 |
| , o endereço e | 36 | 18 |
| comercial , a região | 35 | 17,5 |
| entrar na plataforma wikiversidade | 35 | 17,5 |
| horta , se ela | 35 | 17,5 |
| nome da horta ; | 33 | 16,5 |
| ou comercial , a | 30 | 15 |
| o endereço e o | 29 | 14,5 |
| plataforma wikiversidade e inserir | 29 | 14,5 |
| está , o bairro | 28 | 14 |
| educativa ou comercial ; | 28 | 14 |
| inserir os dados ali | 28 | 14 |
| bairro ; o endereço | 27 | 13,5 |
| da horta ; se | 27 | 13,5 |
| o bairro ; o | 27 | 13,5 |
| ; o endereço e | 27 | 13,5 |
| , a região em | 27 | 13,5 |
| de alguma horta urbana | 26 | 13 |
| comercial ; a região | 26 | 13 |
| horta urbana e orgânica | 26 | 13 |
| ou comercial ; a | 26 | 13 |
| wikiversidade e inserir os | 25 | 12,5 |
| existência de alguma horta | 25 | 12,5 |
| ; se ela é | 25 | 12,5 |
| horta ; se ela | 24 | 12 |
| alguma horta urbana e | 24 | 12 |
| da existência de alguma | 24 | 12 |
| que está ; o | 23 | 11,5 |

| | | |
|---|----|------|
| ; a região em | 23 | 11,5 |
| pode entrar na plataforma | 22 | 11 |
| em que está ; | 22 | 11 |
| está ; o bairro | 22 | 11 |
| que preencher o nome | 22 | 11 |
| ; o bairro ; | 21 | 10,5 |
| que está , o | 21 | 10,5 |
| , o link para | 21 | 10,5 |
| souber da existência de | 21 | 10,5 |
| em que está , | 20 | 10 |
| os dados ali . | 20 | 10 |
| tem que preencher o | 20 | 10 |
| 2) estimular os produtores comerciais orgânicos a incluírem suas hortas (...) e se tornarem mais conhecidos (14 itens) | | |
| os produtores comerciais orgânicos | 40 | 20 |
| produtores comerciais orgânicos a | 37 | 18,5 |
| se tornarem mais conhecidos | 32 | 16 |
| estimular os produtores comerciais | 30 | 15 |
| a incluírem suas hortas | 27 | 13,5 |
| e se tornarem mais | 25 | 12,5 |
| 3) atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade . (15 itens) | | |
| voluntárias no cultivo orgânico | 38 | 19 |
| pessoas interessadas em se | 38 | 19 |
| interessadas em se tornarem | 37 | 18,5 |
| em se tornarem voluntárias | 35 | 17,5 |
| atrair mais pessoas interessadas | 35 | 17,5 |
| cultivo orgânico em plena | 33 | 16,5 |
| orgânico em plena cidade | 32 | 16 |
| mais pessoas interessadas em | 32 | 16 |
| se tornarem voluntárias no | 31 | 15,5 |
| tornarem voluntárias no cultivo | 31 | 15,5 |
| no cultivo orgânico em | 30 | 15 |
| em plena cidade . | 30 | 15 |
| 4) o cadastro de hortas em São Paulo (,) (.) e (8 itens) | | |
| em são paulo , | 43 | 21,5 |
| o cadastro de hortas | 33 | 16,5 |
| cadastro de hortas em | 30 | 15 |
| hortas em são paulo | 32 | 16 |
| em são paulo . | 28 | 14 |
| de hortas em são | 24 | 12 |
| em são paulo e | 20 | 10 |
| 5) na cidade de São Paulo (,) (.) (6 itens) | | |

| | | |
|--|----|------|
| cidade de são paulo | 50 | 25 |
| na cidade de são | 30 | 15 |
| de são paulo , | 27 | 13,5 |
| de são paulo . | 21 | 10,5 |
| 6) . além disso , (4 itens) | | |
| . além disso , | 32 | 16 |
| 7) resgate urbano da natureza (...) inserir mais verde na (8 itens) | | |
| resgate urbano da natureza | 31 | 15,5 |
| inserir mais verde na | 21 | 10,5 |
| 8) uma horta perto de casa (5 itens) | | |
| uma horta perto de | 28 | 14 |
| horta perto de casa | 24 | 12 |
| 9) cadastrar todas as hortas (4 itens) | | |
| cadastrar todas as hortas | 27 | 13,5 |
| 10) produção e venda de alimentos (5 itens) | | |
| produção e venda de | 26 | 13 |
| e venda de alimentos | 24 | 12 |
| 11) o cadastramento de hortas (4 itens) | | |
| o cadastramento de hortas | 24 | 12 |
| 12) regiane nigro e claudia visoni (5 itens) | | |
| regiane nigro e claudia | 24 | 12 |
| nigro e claudia visoni | 23 | 11,5 |
| 13) em nossa cidade . (4 itens) | | |
| em nossa cidade . | 23 | 11,5 |

Anexo 4 - Resultado na íntegra da ferramenta *n-grams* com 5 elementos em sequência que aparecem no mínimo em 10% no corcel20162t3n5

| N-grams encontrados no corcel20162t3n5 | DOCF | Relative DOCF (%) |
|--|------|-------------------|
| 1) da existência de alguma horta urbana e orgânica (...) entrar na plataforma wikiversidade e inserir os dados ali (...) tem que preencher o nome da horta (:), se ela é educativa ou comercial (,) (;) a região em que está (,) (;) o bairro (,) (;) o endereço e (,) o link para contato . (50 itens) | | |

| | | |
|---------------------------------------|----|------|
| o link para contato . | 67 | 33,5 |
| o nome da horta , | 66 | 33 |
| se ela é educativa ou | 64 | 32 |
| ela é educativa ou comercial | 62 | 31 |
| preencher o nome da horta | 54 | 27 |
| a região em que está | 45 | 22,5 |
| é educativa ou comercial , | 44 | 22 |
| nome da horta , se | 43 | 21,5 |
| e o link para contato | 41 | 20,5 |
| o bairro , o endereço | 38 | 19 |
| , se ela é educativa | 37 | 18,5 |
| da horta , se ela | 35 | 17,5 |
| horta , se ela é | 34 | 17 |
| endereço e o link para | 33 | 16,5 |
| , o bairro , o | 33 | 16,5 |
| bairro , o endereço e | 32 | 16 |
| ou comercial , a região | 30 | 15 |
| entrar na plataforma wikiversidade e | 30 | 15 |
| o nome da horta ; | 30 | 15 |
| o endereço e o link | 29 | 14,5 |
| na plataforma wikiversidade e inserir | 29 | 14,5 |
| é educativa ou comercial ; | 28 | 14 |
| o bairro ; o endereço | 27 | 13,5 |
| nome da horta ; se | 27 | 13,5 |
| , a região em que | 27 | 13,5 |
| educativa ou comercial , a | 26 | 13 |
| e inserir os dados ali | 25 | 12,5 |
| ou comercial ; a região | 25 | 12,5 |
| da existência de alguma horta | 24 | 12 |
| wikiversidade e inserir os dados | 24 | 12 |
| educativa ou comercial ; a | 24 | 12 |
| horta ; se ela é | 24 | 12 |
| bairro ; o endereço e | 24 | 12 |
| da horta ; se ela | 24 | 12 |
| comercial , a região em | 24 | 12 |
| plataforma wikiversidade e inserir os | 23 | 11,5 |
| existência de alguma horta urbana | 23 | 11,5 |
| ; se ela é educativa | 23 | 11,5 |
| alguma horta urbana e orgânica | 23 | 11,5 |
| ; a região em que | 23 | 11,5 |
| que preencher o nome da | 22 | 11 |
| região em que está ; | 22 | 11 |
| de alguma horta urbana e | 22 | 11 |
| está , o bairro , | 21 | 10,5 |
| , o link para contato | 21 | 10,5 |
| ; o bairro ; o | 21 | 10,5 |
| que está ; o bairro | 21 | 10,5 |

| | | |
|--|----|------|
| em que está ; o | 20 | 10 |
| tem que preencher o nome | 20 | 10 |
| que está , o bairro | 20 | 10 |
| comercial ; a região em | 20 | 10 |
| região em que está , | 20 | 10 |
| 2) atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade . (15 itens) | | |
| pessoas interessadas em se tornarem | 35 | 17,5 |
| interessadas em se tornarem voluntárias | 33 | 16,5 |
| cultivo orgânico em plena cidade | 32 | 16 |
| se tornarem voluntárias no cultivo | 31 | 15,5 |
| mais pessoas interessadas em se | 29 | 14,5 |
| no cultivo orgânico em plena | 29 | 14,5 |
| em se tornarem voluntárias no | 29 | 14,5 |
| atrair mais pessoas interessadas em | 29 | 14,5 |
| tornarem voluntárias no cultivo orgânico | 27 | 13,5 |
| orgânico em plena cidade . | 23 | 11,5 |
| voluntárias no cultivo orgânico em | 22 | 11 |
| 3) na cidade de são paulo (5 itens) | | |
| na cidade de são paulo | 30 | 15 |
| 4) estimular os produtores comerciais orgânicos a (...) e se tornarem mais conhecidos (11 itens) | | |
| estimular os produtores comerciais orgânicos | 25 | 12,5 |
| os produtores comerciais orgânicos a | 25 | 12,5 |
| e se tornarem mais conhecidos | 23 | 11,5 |
| 5) produção e venda de alimentos (5 itens) | | |
| produção e venda de alimentos | 24 | 12 |
| 6) uma horta perto de casa (5 itens) | | |
| uma horta perto de casa | 24 | 12 |
| 7) cadastro de hortas em são paulo (6 itens) | | |
| de hortas em são paulo | 24 | 12 |
| cadastro de hortas em são | 21 | 10,5 |
| 8) regiane nigro e claudia visoni (5 itens) | | |
| regiane nigro e claudia visoni | 22 | 11 |

Anexo 5 - Resultado na íntegra da ferramenta *n-grams* com 6 elementos em sequência que aparecem no mínimo em 10% no corcel2016t3n5

| N-grams encontrados no corcel2016t3n5 | DOCF | Relative DOCF (%) |
|---------------------------------------|------|-------------------|
|---------------------------------------|------|-------------------|

| 1) da existência de alguma horta urbana e orgânica (...) entrar na plataforma wikiversidade e inserir os dados (...) tem que preencher o nome da horta (,) (;) se ela é educativa ou comercial (,) (;) a região em que está (,) (;) o bairro (,) (;) o endereço e o link para contato . (48 itens) | | |
|---|----|------|
| se ela é educativa ou comercial | 62 | 31 |
| e o link para contato . | 38 | 19 |
| o nome da horta , se | 37 | 18,5 |
| nome da horta , se ela | 34 | 17 |
| ela é educativa ou comercial , | 34 | 17 |
| da horta , se ela é | 34 | 17 |
| , se ela é educativa ou | 32 | 16 |
| , o bairro , o endereço | 32 | 16 |
| horta , se ela é educativa | 32 | 16 |
| o bairro , o endereço e | 31 | 15,5 |
| endereço e o link para contato | 30 | 15 |
| preencher o nome da horta , | 29 | 14,5 |
| educativa ou comercial , a região | 26 | 13 |
| o nome da horta ; se | 25 | 12,5 |
| é educativa ou comercial , a | 24 | 12 |
| nome da horta ; se ela | 24 | 12 |
| ela é educativa ou comercial ; | 24 | 12 |
| da horta ; se ela é | 24 | 12 |
| comercial , a região em que | 24 | 12 |
| é educativa ou comercial ; a | 24 | 12 |
| o bairro ; o endereço e | 24 | 12 |
| entrar na plataforma wikiversidade e inserir | 24 | 12 |
| educativa ou comercial ; a região | 23 | 11,5 |
| na plataforma wikiversidade e inserir os | 23 | 11,5 |
| da existência de alguma horta urbana | 22 | 11 |
| plataforma wikiversidade e inserir os dados | 22 | 11 |
| que preencher o nome da horta | 22 | 11 |
| o endereço e o link para | 22 | 11 |
| ; a região em que está | 22 | 11 |
| horta ; se ela é educativa | 22 | 11 |
| preencher o nome da horta ; | 22 | 11 |
| ou comercial , a região em | 22 | 11 |
| a região em que está ; | 21 | 10,5 |
| ; o bairro ; o endereço | 21 | 10,5 |
| ; se ela é educativa ou | 21 | 10,5 |
| de alguma horta urbana e orgânica | 21 | 10,5 |
| está , o bairro , o | 20 | 10 |
| , a região em que está | 20 | 10 |
| região em que está ; o | 20 | 10 |
| tem que preencher o nome da | 20 | 10 |
| comercial ; a região em que | 20 | 10 |
| existência de alguma horta urbana e | 20 | 10 |

| 2) atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade . (15 itens) | | |
|--|------|--|
| 31 | 15,5 | |
| 29 | 14,5 | |
| 28 | 14 | |
| 27 | 13,5 | |
| 27 | 13,5 | |
| 27 | 13,5 | |
| 26 | 13 | |
| 23 | 11,5 | |
| 21 | 10,5 | |
| 3) cadastro de hortas em são paulo (6 itens) | | |
| 21 | 10,5 | |

Anexo 6 - Resultado na íntegra da ferramenta *n-grams* com 4 elementos em sequência que aparecem no mínimo em 10% no corcel2016t3n2

| N-grams encontrados no corcel2016t3n2 | DOCF | Relative DOCF (%) |
|--|-------------|--------------------------|
| 1) é atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade . (16 itens) | | |
| 35 | 17,5 | |
| 33 | 16,5 | |
| 32 | 16 | |
| 32 | 16 | |
| 31 | 15,5 | |
| 31 | 15,5 | |
| 30 | 15 | |
| 25 | 12,5 | |
| 21 | 10,5 | |
| 20 | 10 | |
| 2) a inclusão de cultivos urbanos em todo o (8 itens) | | |
| 43 | 21,5 | |
| 41 | 20,5 | |
| 32 | 16 | |
| 30 | 15 | |
| 30 | 15 | |
| 3) existência de alguma horta urbana e (...) pode entrar na plataforma wikiversidade e inserir os dados ali (...) preencher o nome da horta , se ela é educativa ou comercial , a região em que está (...) o link para contato . (39 itens) | | |

| | | |
|--|----|------|
| o nome da horta | 39 | 19,5 |
| na plataforma wikiversidade e | 39 | 19,5 |
| é educativa ou comercial | 38 | 19 |
| e inserir os dados | 37 | 18,5 |
| plataforma wikiversidade e inserir | 36 | 18 |
| wikiversidade e inserir os | 34 | 17 |
| entrar na plataforma wikiversidade | 31 | 15,5 |
| ela é educativa ou | 30 | 15 |
| preencher o nome da | 28 | 14 |
| região em que está | 28 | 14 |
| o link para contato | 28 | 14 |
| se ela é educativa | 28 | 14 |
| pode entrar na plataforma | 27 | 13,5 |
| inserir os dados ali | 27 | 13,5 |
| nome da horta , | 26 | 13 |
| de alguma horta urbana | 26 | 13 |
| a região em que | 25 | 12,5 |
| educativa ou comercial , | 24 | 12 |
| link para contato . | 23 | 11,5 |
| existência de alguma horta | 22 | 11 |
| alguma horta urbana e | 22 | 11 |
| 4) para estimular os produtores comerciais orgânicos a (...) e se tornarem mais conhecidos (12 itens) | | |
| estimular os produtores comerciais | 32 | 16 |
| os produtores comerciais orgânicos | 30 | 15 |
| se tornarem mais conhecidos | 25 | 12,5 |
| produtores comerciais orgânicos a | 22 | 11 |
| para estimular os produtores | 22 | 11 |
| e se tornarem mais | 20 | 10 |
| 5) de cadastrar todas as hortas (5 itens) | | |
| cadastrar todas as hortas | 30 | 15 |
| de cadastrar todas as | 25 | 12,5 |
| 6) pela dupla regiane nigro (...) em são paulo , (8 itens) | | |
| pela dupla regiane nigro | 29 | 14,5 |
| em são paulo , | 25 | 12,5 |
| 7) o cadastro de hortas em são paulo (7 itens) | | |
| hortas em são paulo | 26 | 13 |
| o cadastro de hortas | 20 | 10 |
| 8) resgate urbano da natureza (...) mais verde na selva de pedra (7 itens) | | |
| resgate urbano da natureza | 25 | 12,5 |
| na selva de pedra | 21 | 10,5 |
| mais verde na selva | 20 | 10 |
| 9) produção e venda de alimentos (5 itens) | | |

| | | |
|---|----------|------------|
| produção e venda de e venda de alimentos | 22 21 | 11 10,5 |
| 10) cidade de são paulo (4 itens) | | |
| cidade de são paulo | 20 | 10 |

Anexo 7 - Resultado na íntegra da ferramenta *n-grams* com 5 elementos em sequência que aparecem no mínimo em 10% no corcel2016t3n2

| N-grams encontrados no corcel2016t3n2 | DOCF | Relative DOCF (%) |
|--|-------------|--------------------------|
| 1) a inclusão de cultivos urbanos em todo o (8 itens) | | |
| a inclusão de cultivos urbanos | 38 | 19 |
| cultivos urbanos em todo o | 30 | 15 |
| inclusão de cultivos urbanos em | 28 | 14 |
| de cultivos urbanos em todo | 27 | 13,5 |
| 2) existência de alguma horta urbana e (...) pode entrar na plataforma wikiversidade e inserir os dados ali (...) preencher o nome da horta (...) se ela é educativa ou comercial (...) a região em que está (...) o link para contato . (33 itens) | | |
| na plataforma wikiversidade e inserir | 35 | 17,5 |
| wikiversidade e inserir os dados | 31 | 15,5 |
| plataforma wikiversidade e inserir os | 31 | 15,5 |
| ela é educativa ou comercial | 29 | 14,5 |
| se ela é educativa ou | 28 | 14 |
| preencher o nome da horta | 27 | 13,5 |
| e inserir os dados ali | 26 | 13 |
| entrar na plataforma wikiversidade e | 25 | 12,5 |
| a região em que está | 23 | 11,5 |
| pode entrar na plataforma wikiversidade | 23 | 11,5 |
| o link para contato . | 22 | 11 |
| de alguma horta urbana e | 20 | 10 |
| existência de alguma horta urbana | 20 | 10 |
| 3) é atrair mais pessoas interessadas em se tornarem (...) no cultivo orgânico em plena cidade (14 itens) | | |
| pessoas interessadas em se tornarem | 31 | 15,5 |
| no cultivo orgânico em plena | 31 | 15,5 |
| cultivo orgânico em plena cidade | 30 | 15 |
| mais pessoas interessadas em se | 29 | 14,5 |
| atrair mais pessoas interessadas em | 28 | 14 |
| é atrair mais pessoas interessadas | 20 | 10 |
| 4) para estimular os produtores comerciais orgânicos (6 itens) | | |
| estimular os produtores comerciais orgânicos | 25 | 12,5 |
| para estimular os produtores comerciais | 20 | 10 |
| 5) de cadastrar todas as hortas (5 itens) | | |

| | | |
|---|----|------|
| de cadastrar todas as hortas | 25 | 12,5 |
| 6) produção e venda de alimentos (5 itens) | | |
| produção e venda de alimentos | 21 | 10,5 |

Anexo 8 - Resultado na íntegra da ferramenta *n-grams* com 6 elementos em sequência que aparecem no mínimo em 10% no corcel20162t3n2

| N-grams encontrados no corcel20162t3n2 | DOCF | Relative DOCF (%) |
|---|------|-------------------|
| 1) pode entrar na plataforma wikiversidade e inserir os dados ali (...) se ela é educativa ou comercial (16 itens) | | |
| na plataforma wikiversidade e inserir os | 31 | 15,5 |
| plataforma wikiversidade e inserir os dados | 28 | 14 |
| se ela é educativa ou comercial | 27 | 13,5 |
| entrar na plataforma wikiversidade e inserir | 23 | 11,5 |
| wikiversidade e inserir os dados ali | 22 | 11 |
| pode entrar na plataforma wikiversidade e | 20 | 10 |
| 2) atrair mais pessoas interessadas em se tornarem (...) no cultivo orgânico em plena cidade (13 itens) | | |
| no cultivo orgânico em plena cidade | 29 | 14,5 |
| mais pessoas interessadas em se tornarem | 28 | 14 |
| atrair mais pessoas interessadas em se | 26 | 13 |
| 3) a inclusão de cultivos urbanos em todo o (7 itens) | | |
| inclusão de cultivos urbanos em todo | 26 | 13 |
| a inclusão de cultivos urbanos em | 25 | 12,5 |
| de cultivos urbanos em todo o | 25 | 12,5 |

Anexo 9 - Resultado completo da ferramenta *concordance* no corcel20162t3n5

| <i>Left context</i> | Trecho de foco | <i>Right context</i> |
|---|-----------------|--|
| </s><s> É IMPORTANTE SABER QUE CADA CADASTRAMENTO APRESENTA REQUISITOS IMPORTANTES COMO: | O NOME DA HORTA | , SE TEM CARÁTER EDUCATIVO OU COMERCIAL, A REGIÃO EM QUE ESTÁ LOCALIZADA, O BAIRRO, ENDEREÇO E O LINK PARA CONTATO. |
| </s><s> É muito fácil, é só preencher | o nome da horta | , se ela é educativa ou comercial, a região em que está, o bairro, o endereço e, por fim, o link para contato. |
| </s><s> Temos que cadastrar as hortas no site wikiversidade inserindo os dados delas como | o nome da horta | , o tipo de horta, o bairro onde fica, seu endereço e o link para contato. |
| </s><s> Ali vocês podem inserir os dados como por exemplo | o nome da horta | , o caráter, isto é se ela é educativa ou comercial; a localização com o nome da região, do bairro e do endereço assim como o link para contato. |

| | | |
|--|-----------------|--|
| </s><s> O cadastramento é feito mediante a plataforma wikiversidade; ali são inseridos os dados; | o nome da horta | , se ela é educativa ou comercial; a região em que está, o bairro; o endereço e o link para contato. |
| </s><s> Tem que colocar | o nome da horta | , e de que caráter é, a região onde está, o bairro, o endereço e o link para contato. |
| </s><s> Tem que inserir | o nome da horta | , categorizar se ela é educativa ou comercial, identificar a região em que está, o bairro, o endereço, e por fim o link para contato (o seu e-mail). |
| </s><s> Quem conhecer uma horta urbana e orgânica na cidade, pode registrá-la na plataforma wikiversidade, inserindo | o nome da horta | , se é educativa ou comercial, a região e o bairro onde se encontra e o link para contato. |
| </s><s> Só tem que entrar na pagina wikiversidade e preencher dados como | o nome da horta | e a localização dela. |
| </s><s> Vocês têm que editar a plataforma wiki, preencher | o nome da horta | , indicar se é uma horta educativa ou comercial, a região em que está, o bairro, o endereço completo e por finalizar o cadastro, colocar o link para contato. |
| </s><s> Se você souber de alguém ou pode ser até você mesmo que tem em casa uma horta e ainda não cadastrou é bem fácil, só acessar na plataforma Wikiversidade e inserir os dados da sua horta; preenchendo | o nome da horta | ; se a mesma é educativa ou comercial; qual é a região em que está, endereço, e por fim, o link para contato Com o aparecimento do nosso projeto também estimulamos os produtores comerciais orgânicos a se incluírem e se tornarem mais conhecidos. |
| </s><s> Para se cadastrar, precisa-se acessar a plataforma wikiversidade e preencher | o nome da horta | ; definir se ela é educativa ou comercial; a região em que está; o bairro; o endereço e, por fim, o link para contato. |
| </s><s> Tem de preencher | o nome da horta | ; a região em que está, o bairro; o endereço e o link para contato. |
| </s><s> Por isso, se você tiver uma horta na cidade de Florianópolis, poderá inserir os seus dados na plataforma wikiversidade, preenchendo | o nome da horta | ,o bairro onde fica, seu contato, etc. |
| </s><s> Só é necessário preencher | o nome da horta | , o caráter dela, a região que está, o bairro, o endereço e o link para contato. |
| </s><s> Para participar a esse cadastro, nós temos que entrar na plataforma wikiversidade e inserir os dados: empreencher | o nome da horta | ; se ela é educativa ou comercial; a região em que está; o bairro; o endereço e por fim o link para contato. |
| </s><s> De aí virou uma rede em base wiki onde os cadastrados ganham visibilidade, podem achar apoiadores para os seus projetos, e facilitam a todos os usuarios achar uma horta perto de casa (Não só aqueles com facebook) Para o cadastramento só tem que preencher | o nome da horta | ; se ela é educativa ou comercial; a região em que está; o bairro; o endereço o link para contato. |

| | | |
|---|------------------------|---|
| <p></s><s> Hoje eu quero invitá-los apoiar minha iniciativa para cadastrar as hortas em nossa cidade, assim que quem souber da existência de hortas pode entrar na plataforma "wikihortas" criada por mim, para inserir os dados, devem ter em conta que tem que preencher</p> | <p>o nome da horta</p> | <p>, se é educativa ou comercial, região onde está ou simplesmente o link para o contato, todo este processo de cadastramento tem vantagens para estimular a que os produtores comerciais de orgânicos incluam suas hortas no programa, fornecer o sistema de compras coletiva e conhecer outras formas de comercialização, outra vantagem se enfoca em atrair pessoas voluntarias no cultivo orgânico da cidade.</p> |
| <p></s><s> Ali você so tem que preencher</p> | <p>o nome da horta</p> | <p>, onde ela está, algum número para fazer contato e você também tem que classificá-la como educativa ou comercial.</p> |
| <p></s><s> Se você tem escutado sobre a horta urbana em nossa cidade, o que vai fazer é o seguinte: Vai entrar na plataforma "wikiversidade" e preencher</p> | <p>o nome da horta</p> | <p>, se é educativo o comercial, a região o nome do bairro, o endereço e o link para contato.</p> |
| <p></s><s> A ideia é incentivar toda a população para compartilhar o que tem feito em questão de hortas nesta cidade, para trabalhar em equipe e até fornecer o trabalho voluntário, você só precisa entrar na plataforma e preencher</p> | <p>o nome da horta</p> | <p>, o tipo de horta, o bairro, o endereço e o link de contato.</p> |
| <p></s><s> Tem que preencher</p> | <p>o nome da horta</p> | <p>, se ela é educativa ou comercial, a região em que se encontra, o bairro, o endereço, e, o link para contato.</p> |
| <p></s><s> Nela a gente tem que preencher</p> | <p>o nome da horta</p> | <p>, se ela é educativa ou comercial, a região em que está, o bairro, o endereço e o link para contato.</p> |
| <p></s><s> Para fazer o cadastro é necessário entrar no site Wikiversidade, preencher</p> | <p>o nome da horta</p> | <p>, tipo (comercial ou educativa), a região o bairro, e o endereço e o link para contato.</p> |
| <p></s><s> É muito fácil: você entra neste link e preencha</p> | <p>o nome da horta</p> | <p>, especifica se ela é educativa ou comercial, a região em que está, o bairro, o endereço, etc.</p> |
| <p></s><s> São Paulo têm muitas hortas, 25 delas cadastradas na plataforma wikiversidade uma "plataforma que é colaborativa e fácil de editar", por enquanto, a ideia é criar uma tabela de nossa cidade nessa plataforma e assim cadastrar todas as hortas de Campinas; o cadastro é muito fácil, você só deve preencher</p> | <p>o nome da horta</p> | <p>; se ela é educativa ou comercial; a região em que está; o bairro, o endereço e o link para contato.</p> |

| | | |
|---|-----------------|---|
| </s><s> Isso poderia permitir a realização de estudos acadêmicos, possíveis apoiadores, maior participação de voluntários, e etc Já existe uma plataforma informática, chamada wikiversidade, onde é possível ter acesso e inserir os dados ali, como | o nome da horta | , a região, se é educativa ou comercial, e o link para contato. |
| </s><s> Existe um site chamado Wikiversidade criado pela Regiane Nigro e a Claudia Visone, elas neste site tem um cadastramento muito fácil que só precisa de preencher | o nome da horta | , colocar se ela é educativa ou comercial, em que região está, o endereço, bairro e um link de contato. |
| </s><s> Temos que entrar na plataforma wikiversidade e inserir os dados ali, tem que preencher | o nome da horta | ; se ela é educativa ou comercial; região em que está; bairro, etc. |
| </s><s> Atualmente, uma pesquisa colaborativa está permitindo cadastrar na plataforma wikiversidade os dados de hortas sendo cultivadas na metrópole em São Paulo, o funcionamento é simples, você tem que entrar na plataforma e inserir os dados, preenchendo | o nome da horta | e o tipo ao que pertence, se ela é educativa ou talvez comercial, a região, o bairro ou o endereço onde encontra-se e finalmente o 'link' para contato. |
| </s><s> Você tem que entrar no site da Wikiversidade e preencher | o nome da horta | , o tipo dela (se é comercial ou se é educativa), a região e o bairro em que se encontra e o endereço. |
| </s><s> Só com | o nome da horta | , o tipo de horta, educativa ou comercial, onde ela está, o bairro, o endereço e finalmente o "link" de contato, já colabora com o cadastramento. |
| </s><s> É necessário preencher | o nome da horta | ; se ela é educativa ou comercial; a região em que está; o endereço e o link para contato. |
| </s><s> Quem souber da existência de alguma horta urbana e orgânica sendo cultivada em algum rincão de Belo Horizonte pode entrar na plataforma wikiversidade e inserir os dados ali: | o nome da horta | , se ela é educativa ou comercial, a região em que está, o bairro, o endereço e um link para contato. |
| </s><s> Só tem que preencher | o nome da horta | ; se é educativa ou comercial; a região em que está; endereço e o link para contato. |
| </s><s> Além disso tem que preencher | o nome da horta | ; se ela é educativa ou comercial (voltada à produção e venda de alimentos); a região, o bairro, o endereço e ao final o link para contato. |
| </s><s> A ideia do mesmo, é que tenha | o nome da horta | ; se ela é educativa ou comercial; a região onde está, o bairro; o endereço e, por último, o link para contato. |

| | | |
|--|-----------------|--|
| </s><s> Tem que preencher | o nome da horta | , se ela é educativa ou comercial, a região em que fica (bairro ou endereço) e um link para contato. |
| </s><s> Para fazer o cadastro tem que preencher seu nome; | o nome da horta | ; se a horta é educativa ou comercial; bairro onde a horta está; endereço e finalmente numero para contato. |
| </s><s> Boa tarde gente, como vocês sabem eu tenho uma horta familiar-comunitária em casa; esses dias estive lendo sobre a Procura de hortas no São Paulo e gostaria muito de adotar essa iniciativa em nossa cidade, com ajuda de vocês é claro, a ideia principal é cadastrar suas hortas ou das que tenham conhecimento; as instruções para cadastrar são as seguintes; Entre na plataforma wikiversidade, preencha | o nome da horta | , se é educativa ou comercial; localização, bairro, endereço e o link para contato. |
| </s><s> Tem que preencher | o nome da horta | ; se ela é educativa ou comercial; a região em que está; o bairro; o endereço e, por fim, o link para contato. |
| </s><s> É muito facil; só precisa de preencher | o nome da horta | , o tipo (educativa ou comercial), a região onde ela se encontra, o bairro, o cadastro e o "link" para contato. |
| </s><s> Qualquer pessoa pode cadastrar uma horta nesta plataforma, somente tem que preencher | o nome da horta | ; se ela é educativa - ou seja, comunitária ou escolar - ou comercial; a região em que está; o bairro; o endereço e, por fim, o link para contato. |
| </s><s> As hortas podem ser de caráter educativo (comunitárias e escolares) ou comerciais(para produção e venda de alimentos); ambos os tipos poderão ser cadastrados entrando na plataforma Wikiversidade e inserindo os dados ali; deve ser preenchido | o nome da horta | , tipo, bairro, endereço onde está ubicada e um link para contato. |
| </s><s> Por incrível que pareça, há muitas hortas em Uberlândia que nem conhecemos, e por isso esta plataforma é uma boa possibilidade de compartilhar experiências e informações Ao entrar no site você tem que inserir dados como | o nome da horta | , a região em que se encontra, o endereço, o link de contato e especificar se ela é educativa ou comercial. |
| </s><s> Cadastrar-se é facil; somente tem que preencher | o nome da horta | , se ela é educativa ou comercial, a região em que esta, o bairro, o endereço e o link de contato. |
| </s><s> Tem de preencher | o nome da horta | ; se ela é educativa ou comercial; a região em que está; o bairro, o endereço e, por fim, o link para contato. |
| </s><s> Aqui no site disponibilizarei um formato onde os interessados podem preencher | o nome da horta | , se ela esta disponivel para fins comerciais, a região na qual esta localizada, bairro, endereço e informações |

| | | |
|--|-----------------|---|
| | | de contato. |
| </s><s> O nosso tema nesta semana será como podemos contribuir com o hábito da alimentação saudável, criando hortas não teremos baldies, iniciativa esta já aplicada e com muito sucesso na grande São Paulo, por isso quero compartilhar com vocês para implementar na nossa Vitória, é muito fácil de fazer, que se termos alguma horta ou souber da existência na sua comunidade incluindo sacolas, porque e até mesmo nas regiões rurais, pode estar na plataforma wikiversidade e inserir os dados ali, tais como | o nome da horta | , se ela é educativa ou comercial, que a diferença radica no seu objetivo que uma é criada em comunidade e escolas e a outra voltada para a produção e vivendo de alimentos, respectivamente, tem que fornecer ademais o bairro, o endereço e o link para contato, procuremos das preferências ao cultivo orgânico pelos seus benefícios à saúde, a plataforma vai ajudar também na divulgação do trabalho dos produtores comerciais orgânicos uma vez que eles possam inserir os dados devidamente, o registro terá um aporte valioso para estudos acadêmicos e até para achar |
| </s><s> Somente temos de preencher | o nome da horta | . </s><s> Se ela é educativa ou/comercial a região em que está; o bairro; o endereço e o link de contato. |
| </s><s> SE VOCÊ CULTIVA O SABE DA EXISTÊNCIA DE ALGUMA HORTA URBANA E ORGÂNICA SENDO CULTIVADA, PODE ENTRAR NA PLATAFORMA WIKIVERSIDADE E PREENCHER | O NOME DA HORTA | ; SE ELA É DE CARÁTER "EDUCATIVA OU COMERCIAL"; A REGIÃO EM QUE ESTÁ; O BAIRRO; O ENDEREÇO E O LINK PARA CONTATO. |
| </s><s> Tem que entrar na plataforma wikiversidade e inserir | o nome da horta | , se ela é de caráter educativo (ou seja para uso escolar e comunitário) ou de caráter comercial (enforcadas na produção e venda de alimentos), a sua localização e o link para contato. |
| </s><s> Só tem de preencher | o nome da horta | , se ela é educativa o comercial, a região e o bairro, e finalmente o link para contato. |
| </s><s> O cadastramento será feito pela plataforma wikiversidade, onde vamos inserir os dados ali, é muito fácil, é somente colocar | o nome da horta | ; se ela é educativa ou comercial; a região em que está; o bairro, o endereço e um link para contato. |
| </s><s> Qualquer interessado pode inserir os dados na plataforma wikiversidade, preencher | o nome da horta | e identificar, a sua característica se é educativa ou comercial, indicar a região que está, o bairro, endereço e link para contato. |
| </s><s> As informações a aportar são | o nome da horta | , se ela é educativa ou comercial, a região em que está, o bairro, o endereço e os dados de contato do produtor. |
| </s><s> Eu quero convidar a todos vocês a fazer uso do cadastro para as hortas de nossa cidade só precisam de preencher | o nome da horta | ; se ela é educativa ou comercial, a região em que está. o bairro, o endereço e o link para contato. |
| </s><s> Precisam-se simplesmente alguns dados como | o nome da horta | , o endereço, se ela tem finalidade educativa ou comercial algum jeito de contato como responsável. |

| | | |
|---|-----------------|---|
| </s><s> Para cadastrar tem que entrar na plataforma Wikiversidade e preencher os dados na tabela correspondente a nossa cidade, aí tem que preencher | o nome da horta | , se ela é educativa ou comercial, região, bairro e endereço, e o link para contato. |
| </s><s> é muito importante que vocês se cadastrem, é só preencher | o nome da horta | , se ela é educativa ou comercial; a região em que está; o bairro; o endereço e finalmente o link para o contato. |
| </s><s> Inero os dados na Wikiversidade, preenchendo | o nome da horta | , se ela é educativa ou comercial, o endereço e o link de contato. |
| </s><s> Primeiro, você preenche | o nome da horta | , descrevendo se é comercial ou educativa. |
| </s><s> Todas as pessoas que querem fazer parte desta iniciativa tem que preencher | o nome da horta | ; escrever se ela é educativa ou comercial, em que região está, o bairro, o endereço e o link para contato. |
| </s><s> Você precisará das informações seguintes para preencher o formulário: | O nome da horta | , a sua natureza (educativa ou comercial), o bairro, o endereço e no final, o link para contato. |
| </s><s> É só questão de dedicar uns minutinhos do vosso tempo para preencher | o nome da horta | , especificar se ela é educativa (quer dizer "comunitária" e "escolar") ou comercial, a região dela, o bairro, o endereço e, por fim, o link para contato. |
| </s><s> O cadastro e facil só temos que preencher | o nome da horta | ; se ela é educativa ou comercial; a região em que está; o bairro; o endereço e, por fim, o link para contato. |
| </s><s> Uma maneira de divulgar estas iniciativas seria criar uma plataforma coletiva e inserir os dados ali, a saber, | o nome da horta | , se é educativa ou comercial, o bairro e os dados de contato. |
| </s><s> Alí, vocês têm que preencher | o nome da horta | urbana, se ela é educativa, quer dizer comunitária e escolar, ou comercial, ou seja para a produção e venda de alimentos, a região, o bairro e o endereço em que esta, e o link para contato. |
| </s><s> Insira sua horta na plataforma Wikiversidade e preencha | o nome da horta | , se ela é educativa ou comercial; a região que que está; o bairro; o endereço e o link para contato. |
| </s><s> Eu estou criando uma tabela para esta cidade na plataforma Wiki e amanhã já, quem quizer, pode incluir alguma horta que conheça, é só preencher | o nome da horta | , se ela é educativa ou comercial, a região, o bairro, o endereço e o link de contato. |
| </s><s> É só preencher | o nome da horta | ; precisar se ela é educativa ou comercial; a região; o bairro; o endereço e, o link para contato. |
| </s><s> Primeiro, você tem que preencher | o nome da horta | . </s><s> Depois escrever se é educativa ou comercial, a região em que está, o bairro; o endereço e, por fim, o link para contato. |

| | | |
|--|-----------------|--|
| </s><s> Ao ler o artigo que fala sobre cadastramento numa plataforma das hortas de São Paulo, achei a ideia muito interessante e por isso convido vocês a participa desta criação, na qual fiz a mesma coisa, uma plataforma chamada wikiversidade onde vocês podem entrar e inserir os dados das hortas que conheçam dentro da cidade do Rio, é importante preenche | o nome da horta | , se é educativa o comercial, região onde se encontra e um link para contato. |
| </s><s> Você pode entrar no site e inserir dados como | o nome da horta | , objetivo da horta, região, bairro, endereço da sua horta e o link para contato. |
| </s><s> Favor entrar na plataforma wikiversidade, preencha | o nome da horta | , a classificação da horta, educativa ou comercial, a região em que está, bairro e o endereço e o link para contato. |
| </s><s> Ali, onde já existem 25 hortas cadastradas do município de São Paulo, só é preciso preencher | o nome da horta | ; se ela é educativa ou comercial, o endereço, e o link para contato. |
| </s><s> Os dados que se devem inserir são | o nome da horta | , se ela é educativa o comercial, a região em que esta e o link para contato. |
| </s><s> Quem souber da existência de alguma horta urbana e orgânica ou quem tiver uma, somente tem de entrar na plataforma que é muito amigável, e preencher | o nome da horta | , se ela é educativa ou comercial, o endereço e um link para contato. |
| </s><s> Precisamos de dados como | o nome da horta | , a sua localização, bairro, endereço e dados de contato do proprietário. |
| </s><s> Na cidade de São Paulo até agora já foram cadastradas 25 hortas, eu sei que aqui em Porto Alegre há algumas também, então quero convidar vocês a entrar na plataforma Wikiversidade e inserir os dados ali, só tem que preencher | o nome da horta | , se ela é educativa ou comercial, a região em que está, o bairro, o endereço e o link para contato. |
| </s><s> Vocês precisam das seguinte informação: | O nome da horta | A região em que está O bairro O endereço O link para contato. |
| </s><s> Os leitores devem preencher | o nome da horta | , se ela é educativa ou comercial, a região em que está, o bairro, o endereço e Por fim o link Para contato. |
| </s><s> Portanto, quem souber de alguma horta orgânica que ainda não está cadastrada ou se quiser cadastrar e criar Tabelas para outras cidades, pode entrar na plataforma wikiversidade e preencher os dados ali, como | o nome da horta | ; se ela é educativa ou comercial; a região onde está, o bairro; o endereço; e o link para contato. |
| </s><s> O funcionamento é fácil, é feito pela internet, os senhores podem entrar na plataforma | o nome da horta | , a finalidade (educativa ou comercial) e o link para contato. |

| | | |
|---|-----------------|---|
| wikiversidade, nesse site colocar | | |
| </s><s> Tem que preencher | o nome da horta | , se ela é educativa ou comercial, a região em que está incluindo endereço e finalmente um link ou e-mail para contato Qual é utilidade? |
| </s><s> Tem que preencher | o nome da horta | ; se ela é " educativa" ou "comercial", região, endereço e por fim o link de contato, o registro sera proveitoso para estudos academicos, caso tenha voluntarios estes acharem a horta que fica mais perto de casa e para possiveis apoiadores. |
| </s><s> Tem que preencher | o nome da horta | , se ela é educativa ou comercial, o endereço em que está obicada e finalmente o link para contato. |
| </s><s> A plataforma é wikiversidade e vocês podem cadastrar-se ali e preencher | o nome da horta | . </s><s> Site: saudaveles.bras.br. |
| </s><s> Tem de preencher | o nome da horta | , se ela é educativa ou comercial, a região em que ela está, o bairro, o endereço e por fim o seu contacto... Tudo muito facil. |
| </s><s> Para cadastrar a horta, é preciso preencher | o nome da horta | ; se ela é educativa ou comercial; a região em que está; o bairro;o endereço e por fim, o link para contato o cadastramento é feito todo na plataforma de wikiversidade. |
| </s><s> É só preencher | o nome da horta | ; se ela é educativa ou comercial; a região em que está; o bairro; o endereço e, por fim, um link para contato! |
| </s><s> É muito fácil: você tem que entrar na plataforma wikiversidade e inserir dados como | o nome da horta | , se ela é educativa ou comercial, a região, bairro e endereço onde ela está e um link para contato. |
| </s><s> Só precisamos entrar na plataforma wikiversidade e preencher os dados de todas as hortas que conhecemos aquela precisamos colocar | o nome da horta | , se ela é educativa ou comercial, a região em que está, o bairro, o endereço e link para contato. |
| </s><s> QUEM SOUBER DA EXISTÊNCIA DE ALGUMA HORTA URBANA E ORGÂNICA SENDO CULTIVADA EM NOSSA METRÓPOLE DEVE PREENCHER LÁ | O NOME DA HORTA | , SE ELA É COMERCIAL OU NÃO, REGIÃO, BAIRRO, ENDEREÇO E, POR FIM O LINK PARA O CONTATO. |
| </s><s> Têm que preencher | o nome da horta | ; se ela é educativa ou comercial; a região em que está; o bairro; o endereço e, por fim, o link para contato. |
| </s><s> Só basta indicar | o nome da horta | , a sua categoria, a região em que está, o bairro, o endereço e um link de contato. |
| </s><s> Para o correto cadastro, ao ingressar na plataforma deverão preencher | o nome da horta | ; se ela é educativa ou comercial; a região com que está; o bairro; o endereço e, por fim, o link para contato. |

| | | |
|---|-----------------|---|
| </s><s> É preciso colocar | o nome da horta | , se ela é educativa ou comercial, a região, o bairro, o endereço e por último o link para contato. |
| </s><s> Tem que preencher | o nome da horta | , se ela é educativa ou comercial, a região em que está, o bairro, o endereço e o link para contato. |
| </s><s> Pessoal o cadastro tem que fazer na plataforma wikiversidade e inserir os dados ali, tem de preencher | o nome da horta | , se ela é educativa ou comercial; a região as que está; o bairro; o endereço e, por fim, o link para contato. |
| </s><s> O funcionamento é muito simples, só é preciso preencher | o nome da horta | , qual é a finalidade da mesma, a localização exata e o link para acessar o contato. |
| </s><s> Tem de preencher | o nome da horta | ; se ela é educativa ou comercial, a região em que está; o bairro; o endereço e por fim, o link para contato, vamos assim aportar com São Leopoldo com a produção e vendas de alimentos, para assim ganhar uma area rural, ou melhor, porções de territorio como já aconteceu em São Paulo. |
| </s><s> Tem que preencher | o nome da horta | ; a região em que fica; o bairro; o endereço, e por fim o link para contato. |
| </s><s> Procuram em google wikiversidade, lá entram na plataforma, procuram a tabela de Bogotá e nela preenchem | o nome da horta | , escrevem se a horta é educativa ou comercial, escrevam onde ela fica e, por fim, escrevem o link para o contato. |

Anexo 10 - Resultado completo da ferramenta *concordance* no corcel20162t3n2

| <i>Left context</i> | <i>Trecho de foco</i> | <i>Right context</i> |
|---|------------------------------|---|
| </s><s> Hoje eu tenho uma noticia muito interessante para vocês; navegando na internet, encontrei uma publicação sobre o cadastro de hortas em São Paulo, e o motivo porque cada vez mais pessoas estão procurando cidades que permitão a | inclusão de cultivos urbanos | . </s><s> Nessa publicação os agricultores falaram da suas hortas e dos produtos orgânicos, uma ideia que teve começo na area rural e que pouco a pouco foi chegando as áreas urbanas, 20 delas só de caráter educativo e outras 5 comerciais, mas as vantagens de ter alimentos saudáveis perto da cidade não é mala ideia para muita gente que está disposta a inserir na selva de pedra. |
| </s><s> A utilidade que tem o cadastramento de hortas é que permite a | inclusão de cultivos urbanos | em todo o país, tambem estimula os productores comerciais orgânicos a incluir suas hortas ali e se tornem mais conhecidos. |
| </s><s> A ideia de cadastrar todas as hortas é a | inclusão de cultivos urbanos | em todo o país. |
| </s><s> A ideia de cadastrar todas as hortas, permite a | inclusão de cultivos urbanos | é importante tambem para estimular a produtos comerciais organicos pra conhecer seus hortas e se tornarm mais conhecidos. |

| | | |
|--|------------------------------|---|
| </s><s> Foi posto em marcha por duas mulheres Regiane Nigro e Claudia Visoni Esta ideia de cadastrar todas as hortas permite a | inclusão de cultivos urbanos | na maioria ao cultivo organico. |
| </s><s> Eu compartilho sempre informações sobre alimentação saudável e achei muito importante o projeto de | inclusão de cultivos urbanos | na busqueda disto. |
| </s><s> Embora tambem permite a | inclusão de cultivos urbanos | em todo o pais, dá-se preferência ao cultivo orgânico O projeto é importante para estimular os produtores a incluir suas hortas ali e se tornarem mais conhecidos. |
| </s><s> A ideia de cadastrar todas as hortas à plataforma permite a | inclusão de cultivos urbanos | e rurais no país. |
| </s><s> Vai permitir a | inclusão de cultivos urbanos | em tudo o país, estimular os produtores comerciais orgânicos a incluírem suas hortas ali e se tornarem mais conhecidas, ainda muitas delas formam um sistema de compras coletivas, alimentos para a metrópole entre outros formas de comercialização, outra grande utilidade deste cadastro é atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade, o registro é importante para Tabelas medianas, para os voluntarios acham uma horta perto de casa para possiveis apoiadores (dos agricultores profissionais) Pedem salvar da existencia e alguma horta |
| </s><s> ESTA INICIATIVA PERMITE A | INCLUSÃO DE CULTIVOS URBANOS | EM TODO O PAÍS E TAMBÉM A PARTICIPAÇÃO DE OUTROS PAÍSES INTERESSADOS NO TEMA DAS HORTAS URBANAS, E ASSIM INSERIR MAIS VERDE NA CIDADE, ALÉM DAS UTILIDADES EXPOSTAS, TAMBÉM SERVI PARA ESTIMULAR OS PRODUTORES COMERCIAIS ORGÂNICOS A INCLUÍREM SUAS HORTAS NA PLATAFORMA QUE AJUDAREM COMPRA COLETIVA, ALIMENTOS PARA A METRÓPOLE, ENTRE OUTRAS FORMAS DE COMERCIALIZAÇÃO, ASSIM COMO TAMBEM É DE GRANDE UTILIDADE O CADASTRO PARA ATRAIR MAIS PESSOAS INTERESSADAS EM SE TORNAREM VOLUNTÁRIOS NO CULTIVO ORGÂNICO EM PLENA CIDADE, UMA VEZ MAIS CONVIDO PARA SE CADASTRAR E OBTENER AS VANTAGENS DAS HORTAS |

| | | |
|---|------------------------------|---|
| </s><s> Permite | inclusão de cultivos urbanos | , animando bastante gente disposta a inserir mais verde. nas suas casas por uma vida mais saudável. |
| </s><s> Embora, é uma iniciativa para a | inclusão de cultivos urbanos | , orgânicos em todo o país, para a comercialização! |
| </s><s> A ideia do cadastro de hortas, que também permite a | inclusão de cultivos urbanos | , foi iniciado pela dupla Regiane Nigro e Claudia Visoni, há 5 anos quando a "onda" do resgate urbano da natureza começou em São Paulo. |
| </s><s> HORTAS UM JEITO PARA UMA NOVA VISÃO Precisamos de inserir mais verde na cidade de pedra: A | inclusão de cultivos urbanos | nas cidades, países e no mundo são resgate da natureza que estão ficando perdidas com a metrópole. |
| </s><s> O cadastro, de acordo ao texto, é uma horta urbana orgânica cultivada em algum lugar da metrópole, a qual entra na plataforma wikiversidade e inseri os dados ali por um lado, o cadastro de hortas permite a | inclusão de cultivos urbanos | pelo país, de esta maneira as pessoas estrangeiras interessam no tema das hortas. |
| </s><s> O funcionamento deste projeto que se apresentará no ambientes escolares, é assim, você cadastra todas as hortas paulistanas para permitir a | inclusão de cultivos urbanos | em todo o país. |
| </s><s> O projeto é muito interessante e inovador, porque e muito importante o resgate da natureza, permite também a | inclusão de cultivos urbanos | em todo o país, permite estimular os produtores comerciais orgânicos por que incluíram suas hortas ali e se tornaram mais conhecidos. |
| </s><s> Você poderia formar parte da iniciativa imagina todas as vantagens: Permite a | inclusão de cultivos urbanos | em todo o país Sistema de compra coletiva Alimentos para a cidade Atrair pessoas interessadas em se tornar voluntarias no cultivo orgânico. |
| </s><s> Em nossa cidade temos muitas hortas que não todos conhecemos por isso a ideia de cadastrar as hortas de nossa cidade e muito boa para assim permitir a | inclusão de cultivos urbanos | em todo o país dando referência aos cultivos orgânicos, para uma alimentação saudável e ter a participação de nossa comunidade e são todo para educar a nossos filhos que é bom manter uma vida saudável. |
| </s><s> A ideia de cadastrar todas as hortas paulistanas, embora a plataforma também permite a | inclusão de cultivos urbanos | em todo o país. |
| </s><s> A ideia é tomar poucas hortas de São Paulo e lograr a | inclusão de cultivos urbanos | em todo o país. |
| </s><s> Esta ideia de cadastrar todas as hortas paulistanas embora a plataforma também permite a | inclusão de cultivos urbanos | em todo o país. |
| </s><s> Esta ideia de cadastrar hortas já era antiga e foi posta em prática há 15 dias, o permite a | inclusão de cultivos urbanos | em todo o país. |

| | | |
|--|------------------------------|--|
| </s><s> Primeiro, elas desejavam trazer os cultivos e criar a | inclusão de cultivos urbanos | . </s><s> Alem de isso, eles podem ajuda as pessoas de outros países que estejam interessadas. |
| </s><s> Com isso a plataforma permite a | inclusão de cultivos urbanos | em todo o país, promovendo o resgate urbano da natureza, aumentando o cultivo orgânico, estimula os produtos comerciais orgânicos a incluir suas hortas ali e a se tornar mais conhecidas. |
| </s><s> Por isso, vamos a cadastrar-nos na plataforma para permitir a | inclusão de cultivos urbanos | em toda a cidade e estimular também aos produtores comerciais a incluir suas hortas e assim fornecer alimentos para a cidade e as pessoas que não possam cultivar, eles comprariam. |
| </s><s> Esta nova ideia iniciada pela dupla Regiane Nigro no São Paulo, tem como propósito a | inclusão de cultivos urbanos | que poderia ser implemetada neste cidade se vocês ajudassem com o cultivo de hortas nos redores que contam com regiões de áreas verdes. |
| </s><s> A ideia de cadastrar todas as hortas Paulistanas e muito importante e assim permita-se a | inclusão de cultivos urbanos | em todo Brasil pois inclusive pessoas de outros países estan interessadas no tema de hortas urbanas. |
| </s><s> Por tanto o objetivo principal é motivar a | inclusão de cultivos urbanos | em todo o país. |
| </s><s> A ideia de cadastrar todas as hortas paulistanas permite a | inclusão de cultivos urbanos | em todo o país, faz 15 dias que a posta em prática após uma reunião que a jornalista e fundadora Claudia Visoni teria com pessoas de outros países interesadas no tema horta urbana, motivo que fiz o projeto começar a andar. |
| </s><s> pido por favor a participação para nos ajudar ao surgimento e empreendimento desta ideia que o instituto Kairos em sao paulo e a jornalista foram os principais em fazer o surgimento desta maravilhosa ideia para que nossa gente permita a | inclusão de cultivos urbanos | em todo o país Espero no aguardo de sua resposta |
| </s><s> A ideia de cadastrar todas as hortas é que permite a | inclusão de cultivos urbanos | em todo o país. |
| </s><s> A ideia de cadastrar todas as hortas Cochabombinas - embora a plataforma também permita la | inclusão de cultivos urbanos | em todo o país. |
| </s><s> A iniciativa de cadastrar todas as hortas paulistas, mesmo que aceite a | inclusão de cultivos urbanos | em todo o país. |
| </s><s> Também permite a | inclusão de cultivos urbanos | em todo o país de preferencia ao cultivo orgânico outra utilidade é atrair mais pessoas interessadas em se tornarem voluntarias no cultivo orgânico em plena ciudade além de isso também fornecer alimentos saudável para nossa ciudade. |

| | | |
|--|-------------------------------------|--|
| <p></s><s> Santa Cruz 18-10-2016 Sim você tem uma horta o mais não espere mais para fazer cadastramento, poderia por as seguintes razões melhorar e ter muitos benefícios: poderia entrar na plataforma wikiversidade inserir seus dados ali e as pessoas com só um link de você poderia saber sua informação também poderia permitir</p> | <p>inclusão de cultivos urbanos</p> | <p>em todo o país.</p> |
| <p></s><s> A ideia de cadastrar todas as hortas de quito - embora a plataforma também permita a</p> | <p>inclusão de cultivos urbanos</p> | <p>na maioria do país.</p> |
| <p></s><s> Esta comunidade tem já uma plataforma a qual foi iniciada pela dupla, Regiane Nigro e Claudia Visoni, e através desta iniciativa, permite a</p> | <p>inclusão de cultivos urbanos</p> | <p>, estimulando os produtos comerciais e orgânicos.</p> |
| <p></s><s> o funcionamento das hortas urbanas vai começando por a produção de alimentos saudáveis que chava de o produtor a te o consumidor, é assim como tornase importante fazer um cadastramento das hortas urbanas dentro uma plataforma onde só tem que identificar a horta, se ela é educativa ou comercial, nesse funcionamento vai ajudar na</p> | <p>inclusão de cultivos urbanos</p> | <p>em todo o país.</p> |
| <p></s><s> Com este projeto; nos permite a</p> | <p>inclusão de cultivos urbanos</p> | <p>em todo país, assim também sendo de tão importante na estimulação dos produtores comerciais orgânicos; muitos nós fornecemos, em sistema de compra coletiva, alimentos para a metrópole; etc.</p> |
| <p></s><s> O projeto o que trará é de melhorar as condições das hortas e dar a oportunidade de fazer o mesmo nas diferentes partes do país, assim fazendo a</p> | <p>inclusão de cultivos urbanos</p> | <p>em todo o país.</p> |
| <p></s><s> Hoje estarei falando aqui no blog de alimentação saudável onde pido a participação dos leitores no cadastramento e muito importante para nossa cidade a</p> | <p>inclusão de cultivos urbanos</p> | <p>. </s><s> Hoje em horario da manhã eram 25 hortas cadastradas em São Paulo sendo 20 de caráter educativo/ comunitarias e escolares e 5 comerciais para produção de hortas de alimentos o cadastramento dessas hortas urbanas foi iniciado pela dupla Regiane Nigro que trabalha no instituto Kairos em São Paulo; e muito importante ter uma alimentação saudável e a maior forma e fazendo hortas em nossa cidade e muito boa para a saúde da população entaram muitas doenças relacionada com alimentação, ademos</p> |

| | | |
|--------------------------------|------------------------------|--|
| | | para estimular os produtores comerciais orgânicos a incluírem suas hortas. |
| </s><s> A plataforma permite a | inclusão de cultivos urbanos | em todo o país, além disso, outra grande utilidade deste cadastro é atrair mais pessoas interessadas em se tornarem voluntárias no cultivo orgânico em plena cidade. |