v. 34, n. 1, p. 422–457, 2026
ISSN: 2237-2083 · DOI: 10.17851/2237-2083.34.1.423-457
Submetido em: 17/02/2025 · Aprovado em: 06/11/2025

REVISTA DE ESTUDOS DA LINGUAGEM

Desbravando *O Continente*: a Estilística de *Corpus* aplicada à literatura regionalista brasileira

Unveiling O Continente*: Corpus Stylistics Applied to Brazilian Regionalist Literature*

Geovani Henrique Santos de Souza
Universidade Federal do Rio Grande do Sul (UFRGS) | Porto Alegre | RS | BR
geovani.henrique@ifsc.edu.br
https://orcid.org/0009-0005-8218-7159

Rozane Rodrigues Rebechi
Universidade Federal do Rio Grande do Sul (UFRGS) | Porto Alegre | RS | BR
Rozane.rebechi@ufrgs.br
https://orcid.org/0000-0002-1878-7548

Resumo: Estudos em Estilística de *Corpus* (EC), área que aplica os pressupostos teóricos e metodológicos da Linguística de *Corpus* (LC) na análise de estilo literário, ainda são incipientes nos estudos de literatura em língua portuguesa. Uma das razões que justificam a resistência em se utilizar uma metodologia já consolidada em outros campos do conhecimento para o levantamento de padrões estilísticos textuais e literários pode estar diretamente ligada às limitações das ferramentas computacionais no processamento de textos em português brasileiro, especialmente no que tange à identificação automática de domínios semânticos. Este estudo piloto testa o etiquetador semântico multilíngue PyMUSAS, com interface no Wmatrix 7, primeira versão multilíngue do software de análise textual, na investigação de *O Continente*, de Erico Veríssimo. A partir da comparação com um *corpus* de referência composto pela antologia de escritores contemporâneos ao romancista gaúcho, a ferramenta evidenciou os campos semânticos mais representativos da obra regionalista gaúcha quando comparada às dos outros autores. Os padrões estilísticos levantados automaticamente foram, então, analisados manualmente. Os resultados obtidos corroboram dados da crítica literária tradicional, além de demonstrarem outras possibilidades de análise literária. Contudo, esta pesquisa também aponta que são necessários ajustes no etiquetador, a fim de elevar sua fiabilidade e impulsionar os estudos em EC no Brasil.

Palavras-chave: Estilística de *Corpus*; *O Continente*; Erico Veríssimo; etiquetagem semântica automática; literatura regionalista brasileira.

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Abstract: Studies in *corpus* stylistics (CS), an area that applies the theoretical and methodological assumptions of *corpus* linguistics (CL) to the analysis of literary style, are still in their infancy in Portuguese literature studies. One of the reasons that explains the resistance to using a methodology already consolidated in other fields of knowledge for identifying textual and literary stylistic patterns may be directly linked to the limitations of computational tools in processing texts in Brazilian Portuguese, especially regarding the automatic identification of semantic fields. This pilot study tests the multilingual semantic tagger PyMUSAS, interfaced with Wmatrix 7, the first multilingual version of the text analysis software, to investigate *O Continente*, by Erico Veríssimo. By comparing it with a reference *corpus* composed of an anthology of writers contemporary to the southern novelist, the tool revealed the most representative semantic fields of the regionalist work when contrasted with the other authors. The stylistic patterns automatically identified were then manually analyzed. The results corroborate traditional literary criticism data and demonstrate other possibilities for literary analysis. However, this study also points out that adjustments to the tagger are necessary to increase its reliability and boost CS studies in Brazil.

Keywords: *corpus* stylistics; *O Continente*; Erico Veríssimo; automatic semantic annotation; Brazilian regionalist literature.

1 Introdução

A Linguística de *Corpus* (LC) é o campo de estudo que parte de uma coletânea de textos em formato eletrônico, organizada a partir de critérios definidos pelos objetivos de pesquisa, para levantamento de padrões textuais (Tagnin, 2013, p. 29-30). Nesse contexto, o avanço tecnológico e a informatização da ciência possibilitaram a construção de ferramentas que permitem a observação de fenômenos linguísticos textuais de forma (semi)automática. Essa revolução se deve à capacidade de compilação de imensas bases de textos e de consulta a esses bancos de dados com ferramentas computacionais, como AntConc (Anthony, 2021), WordSmith Tools (Scott, 2024) e Sketch Engine (Kilgarriff, 2014).

Embora a LC seja tradicionalmente aplicada à Lexicografia, à Terminografia, ao Ensino de Línguas, à Análise de Variação de Registro e de Discurso, à Tradução, entre outras,

essa metodologia ainda está se estabelecendo nos estudos literários. Zyngier, Carneiro e Novodvorski (2023, p. 391) defendem o uso da LC para a análise literária baseada em evidências, diminuindo, portanto, o subjetivismo das interpretações hermenêuticas. Ressalta-se, contudo, que, após apresentarem algumas possibilidades de aplicação de recursos computacionais na análise literária e estimularem seu uso como forma de questionar paradigmas tradicionais, os autores enfatizam que a metodologia não prescinde da análise crítica.

Na última década, diversos estudos aplicaram recursos computacionais na análise de estilo, fundamentando a área que passou a ser conhecida como Estilística de *Corpus* (EC). McIntyre e Walker (2019) ocuparam-se de sistematizar os conceitos e as aplicações aos quais a EC se circunscreve. No livro, os autores revisam os estudos em língua inglesa que empregam métodos e ferramentas da LC na análise estilística literária e discutem aspectos que particularizam a EC como um campo de estudos.

Algumas aplicações de EC podem ser vistas em investigações de obras literárias em língua inglesa, como em Fischer-Starcke (2009), que analisa as especificidades de *Pride and Prejudice,* de Jane Austen (1813 [2002]), quando comparada aos outros romances da autora; Maturana (2012), cuja pesquisa investiga as vozes poéticas em *The Adoption Papers*, de Jackie Kay (1991); Wynne (2012), com um trabalho de exploração estilística do livro *The Magus*, de John Fowles (1965); e Mahlberg (2013), que realiza uma pesquisa minuciosa da obra digital de Charles Dickens, enfocando técnicas de caracterização e apresentação de linguagem corporal e dialógica a partir de padrões linguísticos. Mais recentes são as publicações de Čermákóva (2018), cujo recorte explora uma abordagem de EC para o estudo da tradução literária voltada à literatura infantil; Ibrahim (2022), que avalia o uso da ferramenta Wmatrix na análise estilística de obras de Charles Dickens; e Can e Cangir (2022), que aplicam a EC na análise de poesias escritas por soldados que combateram na Primeira Guerra Mundial.

Já no contexto brasileiro, há alguns estudos que já recorrem à LC para analisar obras literárias, como Pimenta e Novodvorski (2018), Rangel et al. (2022) e Vital (2022), todavia, não foram encontrados trabalhos que mencionam ou empregam diretamente a metodologia de EC na análise de obras em português brasileiro em portais de periódico Capes e Scielo. Uma das razões da escassez de estudos em língua portuguesa que apliquem a metodologia de EC pode estar diretamente ligada às limitações das ferramentas de etiquetagem semântica dos textos em português.

Este estudo piloto tem por objetivo avaliar o desempenho da versão 7 do *software* Wmatrix[1] (Rayson, 2009), ferramenta on-line de análise de *corpus* integrada ao anotador semântico multilíngue PyMUSAS, e analisar o que a EC pode agregar à crítica literária já existente do romance *O Continente* (Veríssimo, 2013 [1949]).

Na próxima seção, será apresentada uma revisão bibliográfica de pesquisas que utilizaram a EC. Em seguida, a metodologia aplicada a esta pesquisa será explicitada, abrangendo a seleção do *corpus* de estudo e de referência e seu processamento com os recursos disponíveis na ferramenta Wmatrix. Adiante, serão discutidos os resultados obtidos a partir dos métodos selecionados, avaliando-se as vantagens e deficiências da ferramenta no processamento de textos em português. A Conclusão encerra este estudo piloto, apontando as limitações da pesquisa e perspectivas futuras.

[1] Disponível em: https://ucrel.lancs.ac.uk/wmatrix/. A utilização da ferramenta está condicionada à licença paga ou gratuita, de acordo com requisitos de uso.

2 Fundamentação teórica

Nesta seção, são apresentados conceitos e ferramentas basilares da EC: estilística e anotadores semânticos automáticos, neste caso, o *framework* de etiquetagem semântica automática UCREL Python Multilingual Semantic Analysis System (PyMUSAS), disponível na ferramenta de análise textual Wmatrix. Essa ferramenta foi utilizada em diversas pesquisas de LC e EC, como, por exemplo, McIntyre e Walker (2019), McIntyre e Archer (2010), Culpeper, Archer e Rayson (2009), Balossi (2014, 2020), Ibrahim (2022) e Vathanalaoha (2022). Além disso, será feita uma breve apresentação da obra *O Continente*, de Erico Veríssimo.

2.1 Estilística de *Corpus*

A LC é uma abordagem que compreende um conjunto de procedimentos para o estudo de textos autênticos processáveis por programas computacionais e coletados a partir de critérios bem definidos (McEnery; Hardie, 2012; Tagnin, 2013). Trata-se de uma metodologia já consolidada em áreas como Tradução (Baker, 1993; Zanettin, 2012), Lexicografia (Teubert, 2001; Krishnamurthy, 2008), Discurso (Baker, 2023) e Ensino (Biber; Reppen, 2002; O'Keeffe; McCarthy; Carter, 2007), entre outras, mas ainda incipiente na análise de estilo literário, especialmente quando tomamos por base pesquisas em língua portuguesa no Brasil.

Já a Estilística é uma ramificação da Linguística Aplicada que possibilita o estudo de textos em geral e de literatura em particular. Esse campo de estudos dispõe de um conjunto de conceitos para a descrição, interpretação e avaliação textual (Short, 1996). De acordo com o autor, uma avaliação mais objetiva pode endossar ou refutar análises realizadas pela crítica literária tradicional, por se basear em fatos linguísticos observáveis, como aspectos textuais de nível discursivo, gramatical, grafológico, morfológico, lexical, semântico e fonológico (Short, 1996, p. 5).

Da combinação da LC com a Estilística adveio a Estilística de *Corpus* (EC), abordagem que propõe a integração entre dados levantados automaticamente e sua análise manual, que começou a se expandir nos últimos anos (McIntyre; Walker, 2019). Apesar de a EC não estar restrita ao levantamento de campos semânticos, mas também abranger a análise de palavras-chave, linhas de concordância e n-gramas para os estudos de estilo literário, a pesquisa que leva em conta o significado das palavras permite que se explorem padrões temáticos e ampara com dados estatísticos a investigação dos fatos linguísticos descritos por Short (1996).

McIntyre e Walker (2019, p. 14) salientam como vantagens da EC a praticidade na sistematização dos padrões presentes em obras literárias e a diminuição da subjetividade das críticas convencionais, aumentando, portanto, o rigor das análises. Outra vantagem da combinação de abordagens, conforme os autores, é a possibilidade de se alcançar a representatividade, diferentemente do que ocorre em relação à língua geral. Isso porque a EC, conforme McIntyre e Walker (2019, p. 5), permite que se analise, por exemplo, a totalidade das obras de determinado autor, ou de determinado período literário, enquanto que a língua geral jamais poderia ser analisada em sua completude, e qualquer tentativa de se construir um *corpus* representativo e balanceado de todas as possibilidades linguísticas será frustrado (Leech, 2007).

Vale ressaltar que a análise do estilo literário de autores e obras em língua inglesa por meio de ferramentas de LC já vem sendo realizada há algumas décadas. Fischer-Starcke (2009) combina métodos quantitativos e qualitativos no campo da estilística e constitui um marco nos estudos que mais tarde seriam consolidados sob a denominação de Estilística de *Corpus*. A pesquisadora analisou as palavras-chave e frases mais frequentes do romance *Pride and Prejudice* (Austen, [1813]2002), quando comparado aos outros romances da autora, e observou a predominância de palavras relacionadas a família, relações familiares e casamento, agrupando-as em campos semânticos. A análise revelou resultados que não só corroboraram o que já havia sido discutido pela crítica literária tradicional, mas também padrões dominantes relacionados a conceitos mentais e emoções, expressões de incerteza, expressões descrevendo atos comunicativos e predominância de prosódia semântica negativa observáveis no contexto de ocorrência das palavras-chave e frases estudadas.

Čermákóva (2018) realiza um estudo em EC com os livros de Harry Potter e de Winnie the Pooh e suas traduções para o tcheco. O objetivo da autora é examinar e discutir a repetição nesses textos e suas traduções, uma vez que possuem públicos-alvo distintos, sendo a primeira coletânea voltada ao público infanto-juvenil, já alfabetizado, e a segunda a crianças ainda em processo de alfabetização. Do estudo com corpora paralelos, Čermákóva (2018) conclui que palavras-chave e combinações-chave são pontos de partida adequados para identificar redes lexicais que construam o sentido e a coesão do texto literário e que precisam ser levadas em consideração no processo tradutório. A autora observa nos corpora analisados que muitas vezes essas redes lexicais e semânticas dos romances originais acabam alteradas ou mesmo perdidas devido à tentativa de compensação lexical com o uso de sinonímia. A análise estilística é apontada pela autora como uma aliada do tradutor de obras literárias (Čermákóva, 2018, p. 130).

Braga (2020) demonstra como a EC pode auxiliar na construção de uma estratégia pré-tradutória ao se integrar a sua metodologia com o *framework* da análise de relevância tradutória de Nord (2009), a fim de se atenderem às especificidades da tradução literária. Isso porque, na tradução desse gênero, é proeminente o papel de certas funções textuais que devem ser consideradas durante o processo tradutório, como a função expressiva, que marca um texto como literário, segundo a autora. Para ilustrar seu argumento, Braga (2020) realiza um estudo estilístico do conto *Sickert at St Peter's*, de Denton Welch (1942). Para isso, construiu um *corpus* de referência com três romances do mesmo autor para verificar que itens lexicais eram salientes na narrativa selecionada, utilizando o *software* livre AntConc (Anthony, 2023) e a função *keyword* a partir da métrica estatística *log-likelihood*, que determina o índice de chavicidade de uma palavra em relação ao *corpus* de referência. Ou seja, a ferramenta salienta as palavras estatisticamente mais recorrentes no *corpus* em análise do que no de referência. Após a análise qualitativa dos dados, a pesquisadora constrói um resumo instrucional com observações de estratégias linguísticas para a tradução de narrativas, considerando o *framework* de Nord (2009) e os resultados de EC que Braga (2020) encontrou na análise do conto de Welch (1942). Deste modo, o artigo atesta a efetividade dessa combinação metodológica e aponta essa abordagem como contributiva no desenvolvimento de estratégias pré-tradutórias.

Por sua vez, a pesquisa de Vathanalaoha (2022) se baseia em um *corpus* de dramas americanos e britânicos contemporâneos escritos após a Segunda Guerra Mundial e busca responder aos seguintes questionamentos: que campos semânticos são salientes, que tipos de palavras-chave podem ser vistos nesses campos semânticos e de que modo os campos

semânticos e as palavras-chave do *corpus* se inter-relacionam. Para essa análise, o pesquisador utiliza as ferramentas LancsBox (Brezina; Platt, 2024) e Wmatrix (Rayson, 2009), que realizam o processo de etiquetagem semântica por meio do sistema USAS. Vathanalaoha (2022, p. 59) depreende da pesquisa que o uso das ferramentas auxiliou na decodificação da linguagem literária e caracterização do *corpus* em uma perspectiva holística, permitindo encontrar intersecções entre os personagens dos romances selecionados. Essa análise do *corpus* possibilitou a identificação de um padrão de ênfase no delírio e em ambições dos protagonistas, marcado especialmente por uma prosódia negativa.

Vital (2022), embora não mencione explicitamente que utiliza a abordagem de EC, emprega a metodologia da LC para analisar a obra literária poética *Ave, Palavra* (Rosa, 1985), composta por poemas atribuídos a quatro pseudônimos anagramáticos de João Guimarães Rosa. O objetivo do estudo foi identificar diferenças de estilo entre os quatro pseudônimos, combinando abordagens da LC e do Processamento de Linguagem Natural (PLN). O pesquisador analisou quantidade de palavras, riqueza lexical, número de estrofes e palavras mais frequentes para cada um dos anagramáticos. Do estudo, Vital (2022, p. 11-12) identifica traços estilísticos de cada um desses pseudônimos e apresenta, por meio de nuvens de palavras, um comparativo entre eles. A pesquisa é concluída apontando a contribuição da LC para o estudo de obras literárias e a aproximação interdisciplinar entre Ciências Exatas e Humanas que, segundo o autor, é escassa na produção científica em língua portuguesa.

Como já apontado neste trabalho, uma das possíveis razões para essa escassez encontra-se nas limitações de recursos para a etiquetagem semântica automática do português. A seguir serão apresentados dois dos principais recursos com essa finalidade.

2.2 Etiquetagem semântica automática para a língua portuguesa

A necessidade de etiquetadores semânticos para o português tem impulsionado o desenvolvimento de alguns recursos para esse fim e também o aprimoramento de ferramentas preexistentes. No contexto dos estudos de Lexicografia e de Semântica Cognitiva, Bick (2022) sugere o sistema de anotação automática Danish Framenet (Bick, 2011), centrado em papéis semânticos desempenhados sintaticamente, para aplicação em textos em língua portuguesa, obtendo resultados interessantes. A ferramenta, denominada PFN-PT, realiza essa etiquetagem por meio de uma rede que organiza o significado das palavras a partir de *frames*, que são estruturas conceituais que representam eventos, situações ou conceitos abstratos, cujo conjunto se denomina FrameNet. O objetivo do PFN-PT é auxiliar no preenchimento da lacuna existente nos anotadores automáticos baseados em regras contextuais, como o PALAVRAS (Bick, 2014), além de prover recursos para auxiliar na desambiguação do sentido de palavras (WSD, na sigla em inglês) e para tarefas de Inteligência Artificial, como tradução automática. O etiquetador adaptado por Bick (2022) alcançou um desempenho satisfatório para a desambiguação do sentido de *frames* verbais, com 92,2% de acerto.

Uma ferramenta que permite anotações semânticas e já está integrada a alguns softwares de LC é a biblioteca PyMUSAS (Rayson; Berridge; Francis, 2004), desenvolvida em Python pela Universidade de Lancaster, que combina estratégias de Aprendizagem de Máquina (AM) para atribuir uma categoria semântica para cada palavra do *corpus* a ser etiquetado. Essa ferramenta foi integrada à interface web Wmatrix (Rayson, 2009), utilizada neste estudo, e será apresentada na próxima subseção e discutida em detalhes na Metodologia.

2.3 Wmatrix na EC

Wmatrix (Rayson, 2009) é uma ferramenta que oferece recursos semelhantes aos de *softwares* comumente utilizados para pesquisas com *corpus*, como Antconc (Anthony, 2023), Wordsmith Tools (Scott, 2024) e Sketch Engine (Kilgarriff et al., 2014), mas com o diferencial de possibilitar funções relacionadas à anotação semântica. Até a versão 5, a ferramenta permitia exclusivamente a anotação semântica de textos em inglês, por meio da interface com o sistema USAS. Wmatrix 6 foi lançado como uma versão beta de testagem fechada, que passou a integrar o etiquetador semântico multilíngue PyMUSAS, que processa e etiqueta textos em outros oito idiomas, entre eles o português[2]. Atualmente, a ferramenta está em sua versão 7, que manteve a anotação semântica multilíngue, com algumas melhorias de interface.

O sistema USAS (Rayson; Berridge; Francis, 2004) foi criado a partir de um esquema de etiquetas semânticas contendo 21 Domínios Semânticos de Nível Superior (DSNS) e 232 etiquetas de subcategorias de campos semânticos específicos. Por exemplo, o DSNS "O corpo e o indivíduo" (B) contém como subcategorias o Domínio Semântico (DS) "Anatomia e fisiologia" (B1) e "Doença" (B2-), entre outras. Essa lista de DS é baseada no *Longman Lexicon of Contemporary English* (McArthur, 1981 apud Rayson; Berridge; Francis, 2004), sendo ampliada com as subcategorias, conforme mencionado. A ferramenta, desenvolvida por linguistas da Universidade de Lancaster, combina diferentes recursos para PLN, como o anotador automático de classe gramatical (POS tagger) CLAWS, dicionários semânticos, listas de modelos e regras contextuais para a desambiguação de sentido.

Esse sistema vem sendo utilizado em diversos estudos das ciências da linguagem e foi aprimorado para uma versão de suporte multilíngue (Piao *et al.*, 2015), o PyMUSAS. Assim como o FrameNet, o PyMUSAS também possui suas limitações em relação à necessidade de bancos de dados robustos para treinamento de AM, resultando em palavras não categorizadas ou etiquetadas erroneamente.

Cabe salientar que os desenvolvedores optaram por criar categorias semânticas nas quais são agrupados itens gramaticais, lexicais e discursivos. Portanto, nem todo DS apresenta necessariamente uma temática, como pode-se observar em subcategorias como "Frequência", que agrupa advérbios que desempenham essa função (às vezes, frequentemente), "Velocidade: Rápido" (imediatamente, subitamente, de repente) e "Números" (numerais em geral)[3]. A relação de DS de nível superior completa pode ser observada no Quadro 1, com as categorias traduzidas para o português:

Quadro 1 – USAS - Domínios semânticos de nível superior

A	Termos Abstratos e Gerais	N	Números e Medidas
B	Corpo e o Indivíduo	O	Substâncias, Materiais, Objetos e Equipamentos
C	Artes e Ofícios	P	Educação
E	Ações Emocionais, Estados e Processos	Q	Ações Verbais, Estados e Processos
F	Alimentação e Agropecuária	S	Ações Sociais, Estados e Processos

[2] Para detalhes sobre o sistema PyMUSAS, ver https://pypi.org/project/pymusas/.

[3] Para o detalhamento dos domínios semânticos do Wmatrix, recomendamos a leitura de Archer, Wilson e Rayson (2002).

G	Governo e Domínio Público	T	Tempo
H	Arquitetura, Construção, Casas e o Lar	W	Mundo e Nosso Ambiente
I	Dinheiro e Comércio	X	Ações Psicológicas, Estados e Processos
K	Entretenimento, Esportes e Jogos	Y	Ciência e Tecnologia
L	Vida e Seres Vivos	Z	Nomes e Palavras Gramaticais
M	Movimento, Localidades, Viagem e Transporte		

Fonte: Rayson et al. (2004) [tradução própria].

Por fim, outro aspecto da EC que McIntyre e Walker (2019) destacam é a abordagem definida como *corpus-informed stylistics* (estilística informada por *corpus*, em tradução livre), que consiste em usar grandes corpora, como o British National *Corpus* (BNC), para apoiar a análise estilística de textos ou trechos textuais. Segundo os precursores desse campo de estudos, esses corpora atuam como bancos de dados de linguagem, oferecendo a possibilidade de confirmar, com base em evidências empíricas, uma percepção ou interpretação intuitiva sobre as características estilísticas de um texto ou conjunto de textos, ou de questionar, incentivando uma reavaliação dessa intuição inicial (McIntyre; Walker, 2019, p. 26).

Para que a investigação do *corpus* seja eficaz, mediante os objetivos da pesquisa, é preciso selecionar a ferramenta adequada, em meio a tantas disponíveis. Caso, por exemplo, o pesquisador tenha como objetivo contrapor uma lista de palavras de determinado *corpus* de estudo com a lista extraída de um *corpus* de referência, ele terá à disposição diversas ferramentas, inclusive de acesso gratuito, que cumprem com esse propósito. Contudo, se a intenção for partir de um levantamento de campos semânticos de obras em português, como é o caso deste estudo, o pesquisador não encontrará a mesma disponibilidade, conforme apontado anteriormente.

A ferramenta Wmatrix, bem como o etiquetador USAS, já foram utilizados em diversas pesquisas em EC. Como exemplo, aponta-se o estudo de McIntyre e Walker (2010), que utiliza o Wmatrix para a análise comparativa de duas obras de William Blake: *Songs of Innocence* (Blake, 1789) e *Songs of Experience* (Blake, 1794). Por meio da investigação das diferenças lexicais e semânticas manifestadas em ambas as obras do mesmo autor, o objetivo dos pesquisadores foi compreender de que modo elas diferem.

No mesmo artigo, McIntyre e Walker (2010) também demonstram a aplicabilidade da pesquisa baseada em *corpus* para o estudo de roteiros de filmes *blockbusters*, utilizando metodologicamente a EC para validar algumas observações da crítica especializada audiovisual sobre os papéis de gênero nesses filmes, concluindo que a EC pode cumprir com êxito o objetivo de amparar diferentes tipos de análises em obras literárias e ficcionais a partir de suas ferramentas informatizadas.

Ibrahim (2022) realizou uma pesquisa a fim de explorar os recursos da EC, particularmente o uso da ferramenta Wmatrix (Rayson, 2009), para analisar DS em um *corpus* da obra de Charles Dickens, relatando alguns dos limites e desafios da EC, incluindo a dificuldade de se obterem cópias digitais de textos literários ou o direito de usar os textos disponíveis. Da análise dos resultados, o autor também conclui que a sinergia entre as abordagens quantitativa e qualitativa possibilitadas pela metodologia empregada evidenciam a efetividade das ferramentas de análise de *corpus* para serem incorporadas à estilística tradicional.

A partir do exposto, pode-se verificar a eficácia da ferramenta Wmatrix, bem como de seu etiquetador USAS, para a análise de textos em inglês. Neste estudo, testamos as funcionalidades da ferramenta, bem como do etiquetador semântico PyMUSAS, em um *corpus* literário composto pela obra *O Continente*, de Erico Veríssimo.

2.4 Objeto de estudo: *O Continente* de Erico Veríssimo

Erico Veríssimo (1905-1975) é um escritor consagrado da literatura brasileira e sua obra possui fundamental importância para a cultura gaúcha, especialmente considerando-se o sucesso da trilogia *O Tempo e o Vento* (Veríssimo, 2004), composta por três romances: *O Continente, O Retrato* e *O arquipélago*, lançados ao longo da carreira do escritor. Erico Veríssimo não fora bem acolhido pela crítica com seus romances anteriores à trilogia, mas a receptividade do público e a maturidade literária do escritor ficaram nítidas com o lançamento do primeiro livro da saga, *O Continente* (Veríssimo, [1949]2013). concedendo ao autor um grande sucesso e fazendo abrir os olhos dos críticos literários que, até então, consideravam Erico Veríssimo um escritor de literatura inferior, como bem revisa Santos (2019):

> Na atualidade, Erico Veríssimo é considerado um dos mais importantes escritores da Literatura Brasileira, tido como um dos grandes nomes do Modernismo. Suas principais obras são continuamente reeditadas e algumas delas foram adaptadas nos últimos 30 anos para a televisão e para o cinema [...]. Mas, se em relação ao público o autor sempre teve muito prestígio, em relação à crítica literária nem sempre foi assim. No início de sua carreira literária, nas décadas de 30 e 40, o escritor sofreu avaliações muito depreciativas de suas obras por parte dos críticos literários de sua época. Essas apreciações críticas negativas do passado deixaram sequelas que por muito tempo acompanharam o romancista (Santos, 2019, p. 100).

O romance *O Continente*, objeto desta pesquisa, desenrola-se como uma vasta saga multigeneracional, concentrando-se na formação e nos destinos entrelaçados das famílias Terra e Cambará na fronteira do Rio Grande do Sul com o atual Uruguai. O início da narrativa remonta aos tempos coloniais, apresentando os jesuítas e a sua missão de catequização, personificada na figura do Padre Alonzo e do índio Pedro, e o choque de culturas entre colonizadores e povos nativos. Em paralelo, a história mergulha na vida árdua da família Terra, a aparição de Pedro Missioneiro na vida da resiliente Ana Terra, cujo romance é brutalmente interrompido quando o pai descobre que sua filha está grávida de Pedro. Além das adversidades do ambiente inóspito, Ana Terra se vê em meio aos conflitos territoriais entre portugueses e castelhanos, acaba sendo estuprada por castelhanos que invadem as terras da família, assassinando todos os homens adultos da casa.

O enredo ganha novos rumos com a mudança de Ana Terra, sua cunhada, sobrinha e o filho para Santa Fé, onde as vidas dos Terra e dos Cambará se cruzam com a chegada na cidade do capitão Rodrigo Cambará, que desposa Bibiana Terra, neta de Ana, após um duelo de facas com Bento Amaral, filho do coronel e neto do fundador da cidade. Os capítulos do romance se desenvolvem com a alternância entre a narrativa do passado de Ana Terra e Pedro Missioneiro, a vida conjugal de Rodrigo Cambará e Bibiana Terra na cidade de Santa Fé, a ida dele à Revolução Farroupilha, sua morte após regressar e atacar o casarão da família Amaral

e o cenário de guerra da revolução federalista já em 1895, que marca o tempo presente da narrativa e em que há o fechamento do primeiro romance da trilogia.

Esta pesquisa elegeu *O Continente* como objeto de estudo devido ao seu papel fundamental no reconhecimento do escritor pela crítica literária da época, ao seu amplo alcance junto ao público, resultando, inclusive, em duas adaptações para a televisão. Por essa razão, muitas vezes o romance primogênito da trilogia é confundido pelo público leigo com o todo da obra. Esse impacto de *O Continente* sobre a crítica também influenciou a quantidade de estudos direcionados a esse livro, possibilitando uma revisão mais ampla sobre as interpretações literárias e sua correlação com os resultados encontrados com a EC.

Nesse contexto, destaca-se, aqui, a análise crítica empenhada por Regina Zilberman (2004a), que escrutina a trilogia supracitada, a partir da qual tece análises contundentes do caráter estilístico da obra, esmiuçando sua estrutura narrativa e revisitando, inclusive, declarações do autor sobre o processo de escrita dos volumes que a compõem. Dessa análise, salientam-se alguns pontos em relação ao primeiro volume, *O Continente* (Veríssimo, 2013 [1949]). Primeiramente, seu caráter de narrativa fechada, mas que, posteriormente, obteve continuidade com a publicação de *O Retrato* (Veríssimo, 1951) e *O Arquipélago* (Veríssimo, 1961). Em segundo lugar, a circularidade narrativa, construída após a finalização da trilogia, que revela o narrador, até então anônimo, conectando a última página de *O Arquipélago* à primeira de *O Continente*. Por fim, um dos recursos estilísticos que permeia a narrativa dos três livros, conforme destacado por Zilberman (2004a, p. 31), é o papel que os objetos desempenham no preenchimento das lacunas entre os episódios da trama, como a tesoura de Ana Terra, que acompanha as gerações de mulheres parteiras, que a utilizam para cortar o cordão umbilical dos recém-nascidos, e o punhal de Pedro Missioneiro, passado de geração para geração aos homens da família Terra-Cambará.

Entre as características da obra de Veríssimo, destaca-se seu caráter histórico e o planejamento do romance pelo autor, que buscou em fontes históricas primárias – notícias de jornais, depoimentos pessoais etc. – a matéria prima de sua feitura (Bordini, 2004, p. 76). A pesquisadora também aponta o dualismo da obra, que marca as contradições da trajetória da família Terra-Cambará e a presença da família Caré como representativa da miséria camponesa e da invisibilidade social. A trajetória de decadência das personagens femininas, conforme os anos avançam no romance, também é observada por Bordini (2004), tanto quanto a tipificação das personagens, marcada pela debilidade e hesitação conforme as gerações avançam em *O Continente* e o livro se encaminha para a conclusão.

Bordini (2004) afirma que o espaço de *O Continente* vem pouco caracterizado em termos físicos e muito mais em termos daqueles que o habitam: "As personagens são descritas com poucos traços, os mais significativos para suscitar as ações que praticam, do mesmo modo que o espaço diegético é apenas sugerido, acentuando seu potencial simbólico (Bordini, 2004, p. 82)

Santos (2005) analisa como *O Continente* desenvolve a temática da revolução federalista e de que modo o autor recorre ao espaço diegético do sobrado para retratar as implicações da guerra civil numa perspectiva familiar. O pesquisador traça um paralelo entre a ação das forças legalistas, vitoriosas do confronto, que fazem o cerco do sobrado da família Cambará, delineando o cenário de guerra, e o papel do patriarca e chefe político Licurgo, que, resistindo à derrota iminente, vê sua família em condições precárias e precisa lidar com os conflitos internos e a insatisfação dos entes aquartelados.

Já Santos (2009) revisita a fortuna crítica literária de *O Tempo e o Vento* para analisar o tratamento dado à natureza da obra. Parte dos críticos pesquisados pelo autor considera *O Continente* um romance histórico, enquanto se refere às sequências (*O Retrato* e *O Arquipélago*) como romances majoritariamente políticos. Por outro lado, também há críticos que se opõem a essa separação e apontam a totalidade dos textos como um romance histórico, independentemente de Erico Veríssimo ter vivido durante a época retratada na continuidade do primeiro volume. Em resumo, após uma revisão da crítica literária favorável e contrária à natureza histórica do romance analisado, o artigo de Santos (2009, p. 103) considera a trilogia *O Tempo e o Vento* em sua unidade um romance histórico, mas, se analisada individualmente, a condição de romance histórico seria atribuída ao primeiro livro e de romance político aos outros dois.

Contextualiza-se, a seguir, um ponto complexo da obra de Erico Veríssimo: a temática do machismo, que, inclusive, levou o autor a precisar se defender na década de 1970 de acusações de que sua obra glorificaria o comportamento que promove a superioridade do homem.

Garcia e Lisboa Filho (2013) realizaram um estudo da obra *O Continente* a fim de identificar as intersecções entre identidade gaúcha e violência no romance. As autoras constataram que a figura masculina na obra, especialmente em personagens como o protagonista Capitão Rodrigo Cambará, é forjada em valores como virilidade, honra e belicosidade, paradoxalmente coexistindo com comportamentos de machismo e irresponsabilidade. Paralelamente, elas destacam a força e a resiliência das personagens femininas, como Ana Terra e Bibiana, que, embora apareçam frequentemente em contextos de submissão, demonstram sua capacidade de superação diante das adversidades. O artigo também sublinha a naturalização e a transmissão transgeracional da violência – inclusive de abusos sexuais, agressões físicas e psicológicas e violência conjugal pelas quais as personagens são afligidas no enredo.

Bisol e Porto (2015), similarmente, fazem uma análise de como a violência é representada no romance *O Continente*, especulam se existe um posicionamento crítico da obra em relação a esse tema e discutem a relação da obra com a construção de uma memória da história do Rio Grande do Sul. Baseando-se em Ginzburg (2013) e Halbwachs (2006), as autoras inferem que a violência e as guerras são realidades naturalizadas entre as personagens. Embora estas, em alguns momentos, reflitam sobre tal condição – especialmente as mulheres e os personagens religiosos do romance –, não há, contudo, uma ação concreta em direção ao enfrentamento ou à mudança desse paradigma. Bisol e Porto (2015, p. 154-155) explicam que há uma visão positiva das personagens, especialmente masculinas, em relação à violência, e a narrativa enfatiza as consequências trágicas das decisões dessas personagens e seu impacto sobre as famílias retratadas no romance, bem como a aceitação da guerra e da violência. Por fim, concluem que o romance contribui para uma reflexão crítica do papel da violência na construção da formação do Rio Grande do Sul, recompondo uma memória coletiva da história que *O Continente* representa.

Alves (2005), por sua vez, analisa em sua dissertação a presença de passagens de *O Continente* em que se destacam os aspectos religiosos do contexto da obra. Nela, um capítulo inteiro é dedicado a contar a história dos Sete Povos das Missões, de onde surge um personagem chave para a trama, Pedro Missioneiro, em torno do qual o misticismo e a religiosidade se configuram no romance.

Rosa (2019) se vale da análise do discurso para avaliar as temáticas salientes do romance a partir da escolha de três temáticas: "o telurismo (a influência da natureza/solo/local no caráter, costumes ou tradição de determinado povo), as relações étnico-raciais (ques-

tões indígenas e negras apresentadas ou não na obra) e de gênero (o papel das mulheres no enredo)" (Rosa, 2019, p. 81). O pesquisador coleta manualmente as passagens da obra que abordam essas temáticas, e, a partir dessa coletânea de excertos relacionados aos temas apontados, realiza sua análise. De seu estudo, conclui-se que *O Continente* oferece vários caminhos de análise, como a representatividade dos negros, o racismo, a questão das mulheres e seu silenciamento no romance.

Um trabalho recente sobre a obra de Erico Veríssimo é o de Jacobi (2023), que aborda a presença da melancolia em *O Continente*, apontando as guerras apresentadas durante o romance como precursora de uma visão fatalista de mundo pelas personagens. O autor aponta que a perspectiva das personagens femininas no romance permite a compreensão da guerra como destruidora e fonte de perdas e desilusões, bem como a resignação das mulheres ante a sede dos homens por conflitos e poder.

Dando continuidade ao estudo sobre o impacto dessa obra, e em particular de seu primeiro volume (*O Continente*), busca-se demonstrar como a EC pode contribuir para a análise do estilo presente, avaliando em que medida essa abordagem pode confirmar ou não as interpretações da crítica especializada, além de oferecer novas perspectivas.

3 Metodologia

Este estudo piloto se baseia na pesquisa realizada por Ibrahim (2022), que utilizou a versão 5 da ferramenta Wmatrix, bem como seu etiquetador USAS, para avaliar a eficácia da ferramenta na detecção de temas predominantes em quatro romances de Charles Dickens. O autor conclui que a significância estatística dos domínios semânticos (DS) identificados corroborou os temas relacionados às obras de Dickens pela crítica literária.

Conforme salientado anteriormente, o etiquetador desenvolvido para a língua portuguesa, o PyMUSAS, está disponível apenas a partir da versão 6, que foi beta e já descontinuada e, por fim, a partir de 2025 com a versão definitiva, Wmatrix 7. Para este estudo foi realizado um teste preliminar com um *corpus* de referência (CR) parcial no Wmatrix 6 e a pesquisa foi concluída com os corpora definitivos processados no Wmatrix 7. Além de contribuir com a avaliação das capacidades do Wmatrix para a EC, esta pesquisa também aponta algumas limitações da interface em relação ao processamento de textos em português brasileiro. Abaixo, elencam-se os procedimentos metodológicos utilizados neste estudo.

3.1 O *corpus* de estudo e o *corpus* de referência

O *corpus* de estudo (CE) é a obra que passará por análise de EC. Nesta pesquisa ele é composto pela íntegra da 4ª edição da obra *O Continente* (Veríssimo, [1949]2013). Para possibilitar o processamento automático por meio de software de análise linguística, o texto passou pelos processos de digitalização e limpeza - excluindo-se cabeçalhos, elementos pré-textuais e pós-textuais –, preservando-se apenas o conteúdo literário - e foi convertido para o formato Texto sem Formatação (TXT), codificação UTF-8, mais adequada para a análise de textos em línguas abundantes em diacríticos, como é o caso do português. O CE totaliza 247.326 palavras (*tokens*).

Vale enfatizar que, ao selecionar como *corpus* de estudo apenas *O Continente*, não temos como objetivo traçar o estilo autoral de Erico Veríssimo, uma vez que um único livro não permitiria generalizações dessa natureza (Cf. Evert, 2006). O foco da pesquisa é avaliar o potencial da ferramenta Wmatrix para a análise estilística a partir de campos semânticos, investigando características específicas dessa obra, possibilitando, assim, corroborar ou refutar interpretações oriundas da estilística tradicional.

A comparabilidade entre corpora é um princípio metodológico essencial, e está diretamente ligada à representatividade e à adequação da construção dos corpora aos objetivos específicos da pesquisa (Kilgarriff, 2001). Para salientar o que é característico do CE, é necessário contrastá-lo com um CR. De acordo com Tagnin (2013), o CR, também chamado *corpus* de comparação ou *corpus* de contraste, possibilita o levantamento de uma lista de palavras-chave, ou seja, das palavras estatisticamente mais frequentes no CE do que no CR.

Berber Sardinha (2004) recomenda que o *corpus* de referência (CR) tenha de três a cinco vezes o tamanho do *corpus* de estudo (CE), e que a composição dos textos varie de acordo com o tipo de pesquisa a ser realizada. Contudo, a relação de tamanho entre CE e CR permanece uma questão controversa. Pojanapunya e Todd (2018) ressaltam que o resultado de uma análise de *keywords* depende não apenas da proporção entre os corpora, mas também da escolha do índice estatístico utilizado para calcular a chavicidade.

No caso da interface Wmatrix, o CR evidencia, a partir da etiquetagem com o PyMUSAS, quais DS são salientes no CE em comparação com o CR. Para esta pesquisa, elegeu-se como CR uma compilação de escritores contemporâneos de Erico Veríssimo, pertencentes à mesma época de produção literária. Como critérios de seleção dos autores que comporiam o CR foi considerada sua menção no levantamento sugerido por Fischer (2008, p. 113): escritores do período histórico descrito como "República Pós-30 e modernização econômica; o romance neo-realista dos anos 1930-1950".

Outro critério adotado foi a busca por representantes de diferentes regiões do país e com publicações acessíveis em meio digital, uma vez que a necessidade de digitalização de todas obras poderia inviabilizar a construção do CR. O CR coletado possui um total de 117 romances e coletâneas, que passaram pelo mesmo procedimento de limpeza e armazenamento descrito acima para o CE e totaliza 7.146.884 palavras (*tokens*). No Quadro 2 é possível ver a relação completa dos autores e obras que compõem o CR:

Quadro 2 – Relação de obras do *corpus* de referência

Autor	Obras	*Tokens*
Bernardo Élis	*Veranico de Janeiro, Apenas um Violão, Os melhores Contos de Bernardo Elis, O Tronco*	237.280
Cornélio Pena	*Fronteira, Repouso, A Menina Morta*	305.662
Cyro dos Anjos	*Romances: Abdias, Montanha, O Amanuense Belmiro.*	196.165
Dyonellio Machado	*Os Ratos*	40.088
Graciliano Ramos	*A Terra dos Meninos Pelados, Angústia, Histórias de Alexandre, Infância, Memórias do Cárcere, Os Filhos da Coruja, Vidas Secas*	355.917

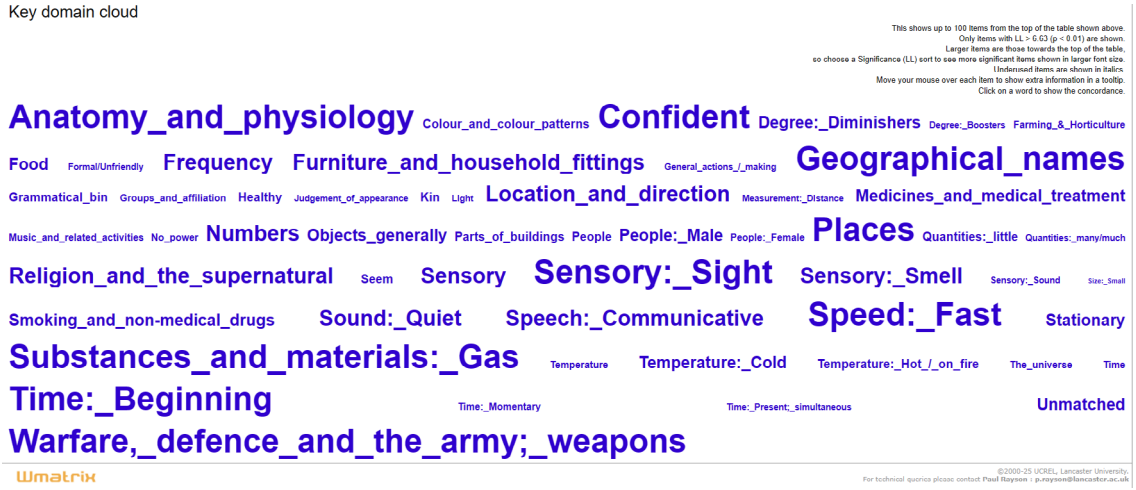
João Guimarães Rosa	*Ficção completa: Sagarana, Manuelzão e Miguelim, No Urubuquaquá, no Pinhém, Noites do Sertão, Grande Sertão: Veredas, Primeiras Estórias, Tutameia, Estas Estórias, Ave, palavra*	808.796
Jorge Amado	*Obra Completa: A Descoberta da América Pelos Turcos, A Morte e a Morte de Quincas Berro D'Água, Cacau, Capitães de Areia, Farda Fardão Camisola de Dormir, Gabriela Cravo e Canela, Jubiabá, Mar Morto, Navegação de Cabotagem, O Capitão de Longo Curso, O Menino Grapiúna, O País do Carnaval, A Luz no Túnel, Agonia da Noite, Os Ásperos Tempos, Seara Vermelha, Suor, Tenda dos Milagres, Tereza Batista Cansada de Guerra, Terras do Sem-fim, Tieta do Agreste, Tocaia Grande*	2.024.959
José Américo de Almeida	*Obra Completa: A Bagaceira, O Boqueirão*	54.876
José Lins do Rego	*Obra Completa: Água Mãe, Caminho de Pedras, Cangaceiros, Menino de Engenho, Doidinho, Banguê, Usina, Fogo Morto, Meus Verdes Anos, Pedra Bonita, Pureza, Riacho Doce*	920.542
Mário Palmério	*Vila dos Confins, Seleta, Chapadão do Bugre*	278.218
Murilo Rubião	*Obra Completa: O pirotécnico Zacarias, O ex-mágico da Taberna Minhota, Bárbara, A cidade, Ofélia, meu cachimbo e o mar, A flor de vidro, Os dragões, Teleco, o coelhinho, O edifício, O lodo, A fila, A Casa do Girassol Vermelho, Alfredo, Marina, a Intangível, Os Três Nomes de Godofredo, Memórias do Contabilista Pedro Inácio, Bruma (a estrela vermelha), D. José Não Era, A Lua, A Armadilha,O Bloqueio, A Diáspora, O Homem do Boné Cinzento, Mariazinha, Elisa, A Noiva da Casa Azul, O Bom Amigo Batista, Epidólia. Petúnia, Aglaia, O Convidado, Botão-de-rosa, Os Comensais*	53.943
Octávio de Faria	*Mundos Mortos, O Anjo de Pedra, O Lodo das Ruas, O Retrato da Morte, O Senhor do Mundo, Os Caminhos da Vida, Os Loucos, Os Renegados*	1.420.062
Rachel de Queiroz	*Obra Completa: As três Marias, Dôra Doralina, João Miguel, Memorial de Maria Moura, O Galo de Ouro, O Quinze*	450.376

Fonte: Elaborado pelos autores.

3.2 Processamento do *corpus* pela ferramenta

Como mencionado anteriormente, a ferramenta Wmatrix 7 (Rayson, 2009) permite a geração de nuvens de DS cujas etiquetas circunscrevem-se na lista de domínios do etiquetador PyMUSAS, apresentada na revisão de literatura. Esse utilitário permite que se ajuste o tamanho da fonte de acordo com a sua significância no CE. Além disso, o usuário pode utilizar um filtro que permite ajustar a visualização da nuvem e da tabela de acordo com seus objetivos. Na Imagem 1, observa-se a nuvem de palavras com o filtro de visualização "incluir itens usados em excesso" e "excluir itens subutilizados". Esses filtros são selecionados para que a interface não exiba a lista completa de todos os campos semânticos, mas apenas os que são salientes no CE, o que facilita a visualização para a análise de dados do CE:

Imagem 1 – Nuvem de DSs do CE na interface Wmatrix

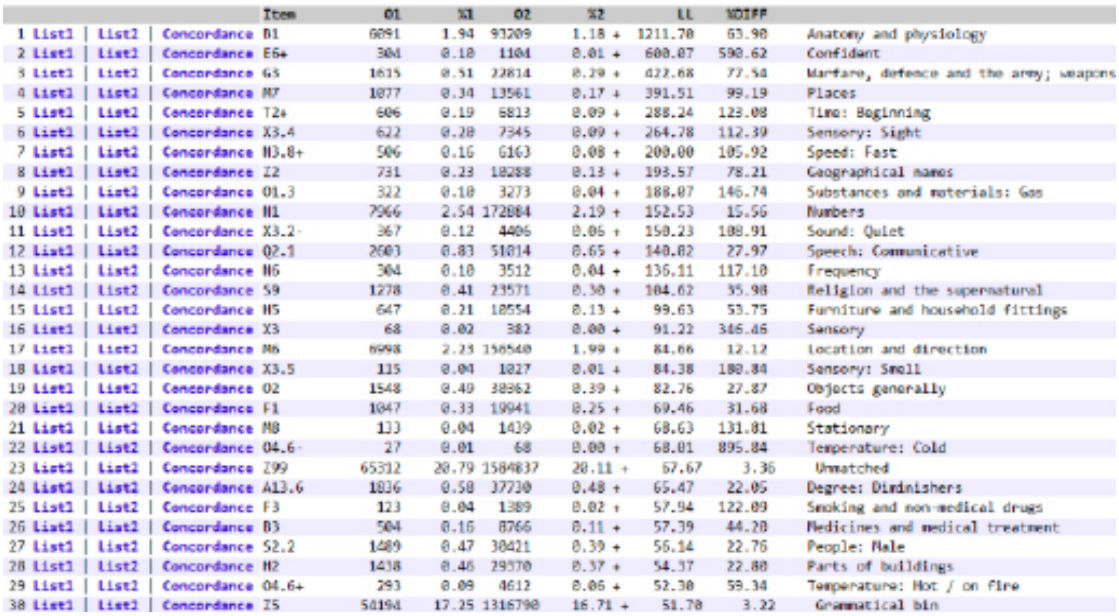


Fonte: Elaborado pelos autores com dados extraídos do Wmatrix 7 (Rayson, 2009).

Juntamente a essa nuvem, é apresentada uma tabela que lista a chavicidade, isto é, o grau de significância estatística (SE) de cada DS no CE em relação ao CR, com os dados estatísticos medidos a partir do índice *log likelihood* (LL). A tabela também disponibiliza os hiperlinks para o acesso às listas de palavras do CE (List1) e do CR (List2), bem como para as linhas de concordância dos itens (Concordance), o código da etiqueta semântica atribuída ao item (Item), a frequência observada no CE (O1), a frequência observada no CR (O2), a frequência relativa do DS em CE e CR (%1 e %2, respectivamente), o valor de SE - *log likelihood* (LL) - e o tamanho de efeito (TE) (%DIFF). De acordo com Gabrielatos (2018), a SE testa a confiabilidade da diferença de frequência entre os corpora comparados, enquanto o TE permite organizar os resultados de acordo com o tamanho da diferença de frequência entre CE e CR.

Na configuração padrão do Wmatrix, são apresentados todos os campos semânticos proeminentes tanto no CE quanto no CR, com a indicação de saliência ou pouco uso nesse contraste por meio dos sinais + e - precedendo o valor de SE (LL) na tabela, informando que determinado DS é mais ou menos saliente no CE do que no CR. A interface permite diversos ajustes, entre eles a possibilidade de ocultar os DS que não são salientes no CE. Esses elementos podem ser observados na Imagem 2, que apresenta os 10 campos semânticos mais representativos do CE, em ordem decrescente de chavicidade (LL):

Imagem 2 – Captura de tela da tabela de DS do CE na interface Wmatrix



			Item	01	%1	02	%2	LL	%DIFF	
1	List1	List2	Concordance	B1	6091	1.94	93209	1.18 +	1211.70	65.90	Anatomy and physiology
2	List1	List2	Concordance	E6+	304	0.10	1104	0.01 +	608.07	590.62	Confident
3	List1	List2	Concordance	G5	1615	0.51	22814	0.29 +	422.68	77.54	Warfare, defence and the army; weapons
4	List1	List2	Concordance	M7	1077	0.34	13561	0.17 +	391.51	99.19	Places
5	List1	List2	Concordance	T2+	606	0.19	6813	0.09 +	288.24	123.08	Time: Beginning
6	List1	List2	Concordance	X3.4	622	0.20	7345	0.09 +	264.78	112.39	Sensory: Sight
7	List1	List2	Concordance	N3.8+	506	0.16	6163	0.08 +	200.00	105.92	Speed: Fast
8	List1	List2	Concordance	Z2	731	0.23	18288	0.13 +	193.57	78.21	Geographical names
9	List1	List2	Concordance	O1.3	322	0.10	3273	0.04 +	188.07	146.74	Substances and materials: Gas
10	List1	List2	Concordance	N1	7066	2.54	172884	2.19 +	152.53	15.56	Numbers
11	List1	List2	Concordance	X3.2-	367	0.12	4496	0.06 +	150.23	108.91	Sound: Quiet
12	List1	List2	Concordance	Q2.1	2603	0.83	51814	0.65 +	140.02	27.97	Speech: Communicative
13	List1	List2	Concordance	N6	304	0.10	3512	0.04 +	136.11	117.18	Frequency
14	List1	List2	Concordance	S9	1278	0.41	23571	0.30 +	104.62	35.98	Religion and the supernatural
15	List1	List2	Concordance	H5	647	0.21	10554	0.13 +	99.63	53.75	Furniture and household fittings
16	List1	List2	Concordance	X3	68	0.02	382	0.00 +	91.22	346.46	Sensory
17	List1	List2	Concordance	M6	6998	2.23	158540	1.99 +	84.66	12.12	Location and direction
18	List1	List2	Concordance	X3.5	115	0.04	1827	0.01 +	84.38	180.84	Sensory: Smell
19	List1	List2	Concordance	O2	1548	0.49	38962	0.39 +	82.76	27.87	Objects generally
20	List1	List2	Concordance	F1	1047	0.33	19941	0.25 +	69.46	31.68	Food
21	List1	List2	Concordance	M8	133	0.04	1439	0.02 +	68.63	131.81	Stationary
22	List1	List2	Concordance	O4.6-	27	0.01	68	0.00 +	68.01	895.84	Temperature: Cold
23	List1	List2	Concordance	Z99	65312	20.79	1584837	20.11 +	67.67	3.36	Unmatched
24	List1	List2	Concordance	A13.6	1836	0.58	37730	0.48 +	65.47	22.05	Degree: Diminishers
25	List1	List2	Concordance	F3	123	0.04	1389	0.02 +	57.94	122.09	Smoking and non-medical drugs
26	List1	List2	Concordance	B3	504	0.16	8766	0.11 +	57.39	44.20	Medicines and medical treatment
27	List1	List2	Concordance	S2.2	1489	0.47	30421	0.39 +	56.14	22.76	People: Male
28	List1	List2	Concordance	H2	1418	0.46	29370	0.37 +	54.37	22.80	Parts of buildings
29	List1	List2	Concordance	O4.6+	293	0.09	4612	0.06 +	52.30	59.34	Temperature: Hot / on fire
30	List1	List2	Concordance	Z5	50194	17.25	1316790	16.71 +	51.70	3.22	Grammatical bin

Fonte: Elaborado pelos autores com dados extraídos do Wmatrix 7 (Rayson, 2009).[4]

Durante a análise qualitativa apresentada nos resultados, avaliamos quais campos semânticos são salientes no CE e, de modo geral, representam o romance *O Continente* do ponto de vista temático. Também discutimos se esses resultados corroboram ou não as observações já feitas pela crítica literária. Em seguida, analisamos alguns dos DS encontrados na etapa anterior quanto às palavras-chave que os representam no romance, discutindo possíveis discrepâncias de classificação e interpretando os dados com base no que a crítica literária já identificou ou não em relação ao tema. Por fim, fazemos observações sobre as possíveis limitações da ferramenta, a partir dos dados analisados durante as etapas anteriores.

4 Resultados

4.1 Sobre o que, afinal, é *O Continente*?

Antes de proceder à análise dos dados obtidos com o Wmatrix, reitera-se, como explicado na fundamentação teórica deste estudo, que o etiquetador PyMUSAS atribui DS para itens lexicais e gramaticais. Portanto, nem todas as categorias se constituem como temáticas do CE. A depender dos objetivos da análise estilística, essas categorias podem ter maior ou menor relevância durante a etapa qualitativa da pesquisa. Em nosso estudo piloto, pretendemos nos ater aos DS que apresentam temáticas salientes do romance, quando comparados a outros de autores contemporâneos a Erico Veríssimo. Para gerar os resultados, foi utilizado, como linha de corte, o valor de LL 10.83, que corresponde a um nível de significância de 0.001 (0,1%). Rayson, Berridge e Francis (2004) esclarecem que esse limite mínimo já emite resultados precisos e confiáveis em

[4] O quadro integral com todos os DS está disponível no Apêndice A deste artigo.

análises estatísticas que utilizam a métrica LL, indicando uma menor probabilidade de que os resultados se devam ao acaso[5]. A seguir, apresentam-se no Quadro 3 os DS salientes do CE comparado ao CR, em tradução literal para facilitar a compreensão dos resultados:

Quadro 3 – dados exportados de subcategorias de DS salientes no CE

Class.	Etiqueta	LL	Subcategorias de DS
1	B1	1211.70	Anatomia e fisiologia
2	E6+	600.07	Confiante
3	G3	422.68	Guerra, defesa e o exército; armas
4	M7	391.51	Lugares
5	T2+	288.24	Tempo: Início
6	X3.4	264.78	Sensorial: Visão
7	N3.8+	200.00	Velocidade: Rápido
8	Z2	193.57	Nomes Geográficos
9	O1.3	188.07	Substâncias e materiais: Gás
10	N1	152.53	Números

Fonte: Elaborado pelos autores com dados extraídos do Wmatrix 7 (Rayson, 2009).

A partir dos resultados apresentados acima, foram removidos os DS referentes à categoria A (Termos Abstratos e Gerais), que contém, principalmente, palavras gramaticais, como advérbios e preposições, à categoria N, que contém números e medidas, à categoria Q - Ações Verbais, Estados e Processos, que concentra a maior parte dos verbos de elocução, à categoria Z (itens não classificados), e os falsos positivos de outras categorias (que serão discutidos na seção 4.3), de modo a facilitar a apresentação das temáticas salientes no romance analisado. O Quadro 4 apresenta uma compilação dos itens mantidos, cujos DS mostram indícios da temática que constrói a narrativa de *O Continente*:

Quadro 4 – Panorama dos DS temáticos do CE

Domínios Semânticos analisados					
1	B1	Anatomia e fisiologia	17	F3	Fumo e drogas não medicinais
2	G3	Guerra, defesa e o exército; armas	18	B3	Medicina e tratamentos medicinais
3	M7	Lugares	19	S2.2	Pessoas: Masculino
4	T2+	Tempo: início	20	H2	Partes de construções
5	X3.4	Sensorial: Visão	21	O4.6+	Temperatura: quente / em chamas
6	Z2	Nomes Geográficos	22	S2	Pessoas
7	O1.3	Substâncias e materiais: gás	23	O4.3	Cor e padrões de cores
8	X3.2-	Som: Silencioso	24	S4	Parentesco
9	Q2.1	Fala: Comunicativo	25	B2+	Saudável
10	S9	Religião e o Sobrenatural	26	O4.6	Temperatura

5 Para maiores detalhes sobre os limites de corte e sua efetividade nos resultados, recomendamos a leitura de Rayson, Berridge e Francis (2004).

11	H5	Mobília e utensílios domésticos	27	K2	Música e atividades relacionadas
12	X3	Sensorial	28	T1.2	Tempo: Momentâneo
13	X3.5	Sensorial: Olfato	29	S2.1	Pessoas: Feminino
14	O2	Objetos em geral	30	T1	Tempo
15	F1	Comida	31	S5	Grupos e Afiliações
16	O4.6-	Temperatura: frio	32	F4	Agricultura e Horticultura

Fonte: Elaborado pelos autores com dados extraídos do Wmatrix 7 (Rayson, 2009).

4.2 Domínios Semânticos de *O Continente*: análise

Para realizar uma análise detalhada dos resultados, optou-se por agrupar os DS por suas categorias de nível superior (DSNS), conforme apresentado no Quadro 1 desta pesquisa e cuja organização pode ser visualizada no Apêndice B. Para cada DSNS, foi realizada uma análise dos dados encontrados e suas implicações no entendimento das temáticas de *O Continente*. Os resultados com discrepâncias devido a erros de etiquetagem serão comentados na seção que trata das limitações da ferramenta.

Reitera-se que não foram analisados os DSNS "A - Termos Gerais e Abstratos", "N - Números e Medidas", "Q - Ações Verbais, Estados e Processos" e "Z - Nomes e Palavras Gramaticais". Deste último, a exceção é para a subcategoria "Z2 - Nomes Geográficos", dada sua relevância para a compreensão da perspectiva histórica do romance, que será discutida junto à categoria "M - Movimento, Localidades, Viagem e Transporte". As categorias excluídas podem ser investigadas futuramente em um estudo que vise às estratégias estilísticas de natureza gramatical e discursiva, que não são o foco da presente pesquisa.

4.2.1 DSNS: B - O Corpo e o Indivíduo

"Anatomia e fisiologia" (B1) aponta dados interessantes sobre o romance e tem destaque como o DS mais saliente de todo o *corpus*, com LL de 1211.70. "Olhos" ocupa a primeira posição, com 654 ocorrências, seguido por "cabeça" (486), "rosto" (266), "mão" (246), "mãos" (205) e "corpo" (199). Observando-se essas palavras em contexto, é possível constatar que a comunicação gestual e interação física das personagens de Erico Veríssimo é intensa e, de modo geral, elas estão constantemente direcionando sua atenção ou respondendo aos interlocutores com gestos enfáticos e reagindo emocionalmente por meio desses elementos, como se observa nas linhas de concordância "Alice sacode a cabeça dum lado para outro [...]", "Alonzo meteu a cabeça no regaço materno e desatou o pranto [...]" e "Desdobrou-o com a mão trêmula e apresentou-o ao dono da casa [...]".

Ainda sobre a categoria "Anatomia e fisiologia", cabe uma ponderação em relação ao romance de (Veríssimo, 2013 [1949]).): há um total de 261 termos com uma frequência relativa, isto é, a proporção de ocorrências de uma palavra em relação ao total de palavras de um *corpus*, de 1,94%, e observa-se uma diversidade de palavras maior do que aquelas relacionadas ao espaço físico. Isso fica perceptível ao se analisarem as categorias M7 - Lugares (77; 0,34%), H2 - Partes de construções (66; 0,46%), H5 - Móveis e utensílios domésticos (46; 0,21%), corroborando a observação de Bordini e Zilberman (2004) de que Veríssimo trabalha pouco o espaço

diegético na trama, centrando seu estilo literário no desenvolvimento das personagens. Essa estratégia estilística, porém, é pouco explorada diretamente na crítica literária sobre o autor.

A segunda subcategoria mais saliente do grupo é "Medicina e tratamentos medicinais" (B3) e possui na primeira posição a palavra "doutor", com 187 ocorrências, seguida por "médico" (157). A proeminência desse DS se justifica pela presença de um personagem secundário relevante para a trama, o médico alemão Dr. Winter, sendo que a maior parte das ocorrências das palavras coocorre com seu nome ou a ele se refere. A palavra "remédio" (30) ocorre na sequência, seguida por "cinta" (24) e "hospital" (10), que surge no início do romance nas reduções jesuíticas e na passagem pelo dr. Winter por Hamburgo, antes de sua chegada em Santa Fé, onde se tornou médico familiar e amigo dos Terra Cambará, como se observa em "Faço as minhas visitas quase diárias, como médico que sou da casa" e "Estou aqui não só como seu amigo, mas principalmente como médico da casa". Ao se analisarem as coocorrências de "cinta" (24), observa-se um erro de classificação, pois todas as linhas de concordância estão relacionadas a um acessório em que os homens prendem suas armas, como se observa na Imagem 3:

Imagem 3 – Captura de tela das primeiras dez linhas de concordância do termo "cinta"

Line	Left context	Key	Right context	File
1	– Sim , eu o levava à	cinta	. O cura deu uma palmada na	CTN1-1
2	, experimentava-lhe a ponta , punha-o na	cinta	e imaginava-se um guerreiro como o corregedor	CTN1-1
3	penacho azul e amarelo , espadim à	cinta	e pés descalços . Os famosos Dragões	CTN1-1
4	de couro na cabeça , facão na	cinta	, veem os açorianos suando ao sol	CTN1-1
5	Deus que me mandou . Tira da	cinta	a faca , aproxima-a do pescoço de	CTN1-1
6	com um penacho , e trazia à	cinta	um espadagão e duas pistolas . E	CTN1-2
7	Examinou-a por alguns instantes , pô-la à	cinta	, ergueu-se e , sem dizer palavra	CTN1-2
8	tiracolo e o facão de mato à	cinta	. Decerto ele acampara ali numa noite	CTN1-2
9	o punhal de prata que trazia à	cinta	. Ana voltou para casa com a	CTN1-2
10	, segurando o punhal que trazia à	cinta	. Eulália perdera a fala , e	CTN1-2

Fonte: Elaborado pelos autores com dados extraídos do Wmatrix 7 (Rayson, 2009).

4.2.2 DSNS: E - Ações Emocionais, Estados e Processos

A subcategoria "Confiante" (E6+) constitui-se como falso positivo, com a palavra "fé" ocorrendo 300 vezes, seguida apenas por "confidências" (2), "mente" (1) e "crenças" (1). Trata-se, todavia, de um erro de etiquetagem, pois a cidade fictícia em que se passa boa parte da trama chama-se Santa Fé, e, pelo fato de o anotador semântico não reconhecer nomes compostos, tratou todas as ocorrências da palavra como E6+.

4.2.3 DSNS: F - Alimentação e Agropecuária

A subcategoria "Comida" (F1) reflete uma diversidade de alimentos, frutas e pratos típicos do Rio Grande do Sul, como "churrasco", "guisado", "massa", "bergamota", "compota", "bolacha" e "salame", além de nomes de frutas, verduras e outros termos relacionados à alimentação. "Fumar e drogas não medicinais" (F3) revela um comportamento habitual das personagens masculinas no romance: o tabagismo. O termo "cigarro", que figura em primeiro lugar, traz 80 ocorrências, todas envolvendo personagens masculinas. Há apenas uma ocorrência de "fumar" (14) e de "fumava" (8) em que é uma mulher o sujeito da ação : "— gritou para a Paraguaia, que fumava a um canto seu cachimbo de barro" e "A Paraguaia continuou a fumar, ouvindo agora os ruídos que vinham [...]". Por fim, a categoria "Agricultura e horticultura"

(F4) apresenta o aspecto rural do romance, sendo o termo de maior ocorrência "terra" (213), seguido por "campos" (73), "campo" (71), "gado" (62), "lavoura" (37) e "lavouras" (18).

4.2.4 DSNS: G - Governo e Domínio Público

A saliência de "Guerra, defesa e o exército; armas" (G3) corrobora a perspectiva da crítica literária sobre o papel das guerras e seu impacto no romance, bem como sua responsabilidade na caracterização de *O Continente* como um romance histórico (Santos, 2009). Destacam-se aqui também as considerações de Santos (2005) sobre essa temática, que retrata as condições precárias causadas pelo cerco militar ao sobrado dos Cambará, ocasionando uma mistura entre espaço doméstico e palco de guerra. O DS conta com 111 termos e uma significância de 0.51%, ocupando a terceira posição na classificação geral dos DS salientes do CE. A palavra "guerra" lidera a categoria com 259 ocorrências, seguida de termos que designam as patentes das personagens - "capitão" (201), "coronel" (82) e "major" (76) -, seguidos de uma grande diversidade de termos que remetem ao confronto bélico, como "soldados", "tropas", "fortes" e palavras em geral que designam armamentos e agentes do conflito.

4.2.5 DSNS: H - Arquitetura, Construção, Casas e o Lar

Abarcam termos relacionados à construção do cenário as categorias "Móveis e utensílios domésticos" (H5), com 46 termos e "Partes de Edifícios" (H2), com 67. Observa-se em H5 o termo "cama" (172) na primeira posição, seguido de "mesa" (151) e "cadeira" (110). Analogamente, em H2 aparecem os principais ambientes onde se passa a trama: "quarto" (151), "sala" (122) e "janela(s)" (166). Considerando que a maior parte da trama se desenvolve no sobrado em que as personagens estão confinadas, a limitação do espaço diegético da trama por esses ambientes é compreensível.

4.2.6 DSNS: K - Entretenimento, Esportes e Jogos

"Música e atividades relacionadas" (K2) é a subcategoria saliente deste DS, com LL de 28.18 e aponta o papel que a música desempenha na vida das personagens de *O Continente*, especialmente na primeira parte do romance, que narra os tempos das reduções jesuíticas, cujo uso da música se mostra uma indispensável ponte entre o povo indígena e os colonizadores no romance. Isso pode ser observado nas linhas de concordância de "música" (84 ocorrências), palavra mais saliente do DS: "Aqueles índios amavam a música[...]", "No princípio a música fora a linguagem por meio da qual padres e índios se entendiam." e "Por meio da música os jesuítas induziam os índios ao estudo [...]". Também é pelo uso da música que Pedro Missioneiro se aproxima de Ana Terra e aos poucos a conquista: "De repente Ana Terra descobriu que aquela música estava exprimindo toda a tristeza que lhe vinha" e "Aquela música saía do corpo de Pedro e entrava no corpo dela... Oh!".

Assim como mencionado no encontro de Pedro Missioneiro com Ana Terra, observa-se que a música é um recurso estilístico recorrente de Erico Veríssimo para conduzir o estado emocional das personagens. Analisando todas as ocorrências de "música" também é possível observar que a sua presença estabelece um elo geracional entre vários personagens da história dos Terra Cambará, pois é com um violão nas costas que chega à Santa Fé o capitão Rodrigo e

com o instrumento ele se aproxima dos citadinos: "E aquela noite as gentes de Santa Fé ouviram música de violão na casa de Nicolau", também é tocando cítara que Luzia inquire seu filho Licurgo quanto ao tamanho de seu amor por ela: "A música doce envolvia Licurgo, que se imaginava no Angico olhando o pôr-do-sol." e a banda marcial criada na cidade após os acontecimentos da revolução farroupilha também exerce um poder quase hipnótico sobre Licurgo: "Aos compassos vibrantes da música Curgo teve, mau grado seu, um estremecimento de entusiasmo".

Além do exposto, a música também é utilizada para caracterizar os homens gaúchos, seu estilo de vida e relação com a arte: "[...] olhavam com repugnada desconfiança para os que se preocupavam com poesia, pintura ou certo tipo de música que não fossem as toadas monótonas de seus gaiteiros e violeiros.", "Foi também na guerra que Chiru pela primeira vez na vida ouviu uma banda de música.", "A fala deles tem música e é doce como laranja madura e meio parecida com a nossa." e "As gaitas enchiam os salões com sua música rasgada e chorona.".

4.2.7 DSNS: M - Movimento, Localidades, Viagem e Transporte e Subcategoria: Z2 - Nomes Geográficos

A subcategoria "Lugares" (M7) ocupa a quarta posição na classificação geral dos DS e possui 77 termos, tendo como palavra de maior impacto "praça" (130), local de grande importância para o desenvolvimento da trama de *O Continente*, uma vez que fica de frente para o sobrado e é onde se encontram as personagens antagonistas da família cercada. Na sequência aparecem os termos "província" (121), "vila" (106), "cidade" (81), "cemitério" (80), "rancho" (63) e "povoado" (61). A subcategoria "Nomes geográficos" (Z2) tem grande saliência no *corpus* devido à alta frequência das palavras "índios" (92), "índio" (85), "alemão" (43) e "Brasil" (40). Essas duas categorias são adequadas para demonstrar como se constrói o cenário em que a trama se passa e reflete cada um dos aspectos históricos do romance, que narra desde as missões jesuíticas, a vida no campo e os conflitos da cidade marcada pelo confronto político e a chegada de imigrantes n*O Continente* de São Pedro. M6 (Localização) tem impacto nos resultados devido à alta ocorrência da preposição "a" (3007), "sem" (439), "onde" (368) e "lá" (322), cujo análise foge ao escopo desta pesquisa por sua característica morfossintática.

Já "Estacionário (M8)" tem como palavras mais salientes "sentado" (57), "imóvel" (30) e "parado" (10). A palavra "sentado" é utilizada como recurso cênico em diversos momentos da narrativa a fim de indicar o estado emocional das personagens, como se observa nas linhas de concordância a seguir: "Sentado no seu canto, o velho Florêncio Terra está imóvel, de cabeça baixa", "[...] Pedro, que continuava sentado, imóvel [...]", "Ao entrar encontrou-o sentado, encurvado sobre a mesa, com a cabeça metida nos braços, soluçando como uma criança", "Os homens estavam sentados em silêncio" e "Sentado numa cadeira, pitando tranquilamente seu cigarro de palha [...]".

4.2.8 DSNS: O - Substâncias, Materiais, Objetos e Equipamentos

"Substâncias e materiais: Gás" (O1.3) é o DS de maior saliência do grupo e ocupa a nona posição geral dos DS, tendo "ar" (236) como a palavra mais recorrente, seguida de "fumo" (40) e "fumaça" (19). Há uma diversidade polissêmica notável no uso de "ar", que não aparece apenas em seu sentido literal de "gás" ou "atmosfera". A palavra em questão é utilizada como elemento descritivo para cenários, estados emocionais e também ações. Um dos usos mais

proeminentes está na descrição do ambiente físico (23 ocorrências), que surge relacionado à temperatura ("[...] o ar estava frio e úmido", "[...] ar morno ao sol e fresco à sombra") e à qualidade ambiental ("[...] ar límpido", "[...] penumbra leitosa azulava o ar"). Além disso, também está presente em muitas descrições sensoriais (50) ("[...] o ar cheirava a sereno", "[...] o ar cheirava a incenso", "[...] um ar enfumaçado") e sonoras ("o ar enchia-se de sinos e das vozes", "[...] sinos subiam no ar"). Também constam nas linhas de concordância o uso da palavra em ações dinâmicas (79) ("[...] ergueu-o dramaticamente no ar [...]", "[...] moscas voavam no ar pesado.") e no sentido figurado (23) ("[...] palavras morreram no ar [...]" e "[...] de pernas pro ar."). Por fim, é importante observar também o uso de "ar" para expressão de caráter e comportamento (53 ocorrências) das personagens, que se observa na incidência recorrente da construção sintática "com ar + adjetivo/substantivo", delineando uma gama de emoções humanas: "ar palerma", "ar céptico", "ar indeciso", "ar fúnebre", "ar sentencioso", "ar cansado", "ar desamparado", "ar agressivo", "ar sonolento", "ar trocista" etc.

"Temperatura" (O4.6) e "Temperatura: frio" (O4.6-) surgem entre os DS salientes com pouca diversidade de palavras, sendo as mais relevantes "morno" (22) para o primeiro e "geada" (20) para o segundo. "Temperatura: quente / em chamas" (O4.6+) tem "fogo" (109), seguido por "quente" (48) e "calor" (31). O uso de "fogo" traz a maioria das ocorrências no sentido literal: "[...] melhor é ir para junto do fogo [...]." e "[...] acender o fogo para aquentar a água do chimarrão." são exemplos desse uso. Também há o uso como metáfora/ sentido figurado, observável em "aquecido pelo fogo duma raiva intensa.", "[...] orelhas em fogo [...]" e "Um fogo ardia no peito de Rodrigo.". As demais ocorrências da palavra surgem relacionadas a "arma de fogo/ disparo" e seriam classificadas de forma mais adequada como G3 (Guerra, defesa e o exército; armas): "[...] com a arma de fogo [...]", "[...] faça fogo sem piedade." e "Bolívar fez fogo e Dentinho de Ouro tombou de joelhos.".

A subcategoria "Cor e padrões de cores" (O4.3) é um domínio semântico sobre o qual cabem algumas observações relevantes. O termo "Pardo" (94 ocorrências), de maior frequência da categoria, na maioria das entradas se refere a nome próprio, que deveria constar na categoria Z1. Na sequência temos o termo "vermelho" (51), extremamente significativo no contexto histórico do romance, uma vez que representa uma das facções políticas (os maragatos) envolvidas no cerne do conflito da trama e coocorre principalmente com a palavra "lenço" ou aparece em contextos associados a esse acessório - na cultura dos gaúchos, o lenço é parte integrante de sua indumentária, e, no contexto da revolução federalista, sua cor reflete a ideologia política do usuário.

Outro ponto relevante nessa mesma categoria é em relação às palavras "branca" (47) e "branco" (47), que são pouco utilizadas para a descrição étnica das personagens. A primeira vem, predominantemente, acompanhada de "bandeira", "arma" ou "marmelada" e, das 47 ocorrências de "branco", em apenas três a palavra se refere à etnia. Já os termos "negro(s)", "negra(s)" e "preto(s)" surgem em grande maioria para designar personagens secundários. O termo "negro", por exemplo, designa etnia em 75 das 91 ocorrências. Na maior parte das situações, se refere a personagens negras anônimas e que passam por situações ora de violência, ora de discriminação, ora de objetificação. Isso também é corroborado pela saliência do DS "Sem poder", cujas palavras mais proeminentes são "escravos" (64) e "escravo" (31). O estudo desses DS corrobora a análise do discurso que Rosa (2019) realiza em sua dissertação sobre presença da temática do racismo no romance.

4.2.9 DSNS: S - Ações Sociais, Estados e Processos

"Religião e o sobrenatural" (S9) é a subcategoria com maior proeminência no grupo S e revela, em seus 109 termos, aspectos do romance *O Continente* relacionados ao seu contexto histórico religioso, objeto de estudo da crítica literária de Alves (2005). Alguns dos capítulos do romance ocorrem no período das missões jesuíticas no Rio Grande do Sul, e indícios dessa temática podem ser inferidos pelos termos "deus" (291), "igreja" (135), "céu" (100), "diabo" (76) e "alma" (65), que ocupam as cinco primeiras posições de S9. Os resultados dessa subcategoria também indicam a prevalência do cristianismo na obra.

"Objetos em geral" (O2) ocupa a segunda posição dentro da categoria; porém, levanta dúvidas sobre a prevalência de seus três primeiros termos: "coisas" (280), "amigo" (136) e "amigos" (69), isso porque a primeira palavra possui apenas 26 ocorrências relacionadas a objetos em geral, tal qual em "[...] as coisas do rancho [...]" e "[...] as coisas de comer[...]", todas as demais são usos abstratos: "[...] aquelas coisas tristes [...] e "[...] as coisas que sentiam antes". A presença de "amigo" em "Objetos em geral" também é incoerente, já que existe o DS "Grupos e afiliações" (S5), que, a propósito, também é saliente no *corpus*. Porém, apresenta inconsistências de classificação e só contém as palavras contíguo (5) e contígua (4), ambas designando cômodos de uma casa: "quarto contíguo" e "sala contígua".

A subcategoria Pessoas (S2) possui como palavra com maior frequência "filho" (317). Todavia, as demais palavras relacionadas à família foram etiquetadas na categoria Parentesco (S4) ("filha" e "filhos", por exemplo, estão nessa subcategoria). Não foi possível compreender por que o anotador separou a palavra "filho" das demais que se referem a laços familiares com etiqueta S4.

A subcategoria "Pessoas: masculino" (S2.2) demonstra a prevalência das personagens masculinas na obra, com "homem" em primeiro lugar (579) e "homens" (261) na sequência. Comparativamente, observa-se uma menor quantidade de ocorrências das palavras "mulher" (328) e "mulheres" (165), ambas da subcategoria "Pessoas: feminino" (S2.1), que também é saliente no *corpus*. Além disso, uma análise das linhas de concordância de S2.2 corrobora a impressão da crítica especializada em relação à caracterização das personagens masculinas, como a de Santos (2005) e Rosa (2019), que comentam a temática do machismo que perpassa os capítulos de *O Continente*.

A obra traz para o primeiro plano a imposição da sociedade sobre o comportamento das personagens masculinas e da idealização do homem gaúcho, destemido, bruto e violento. Isso pode ser observado nas linhas de concordância dos termos de S2.2, como em "Um homem bem macho não chora nunca [...]".

Ao se analisarem as linhas de concordância dos termos "menino" (143) e "rapaz" (95), é possível notar que esses não surgem em contextos de afirmação da masculinidade, como observado nos exemplos acima. O comportamento das personagens referidas por esses termos demonstra traços como pureza, educação e imaturidade. Como se observa em "[...] o rosto do menino tinha uma pureza de imagem.", "Ele era um menino bom que não fazia mal pra ninguém.", "Talvez o rapaz houvesse falsificado a firma do pai [...]" e "O rapaz respondeu com um sorriso meio constrangido".

Por fim, é importante observar que o termo "macho" (35) revela nas linhas de concordância seu uso para afirmação da masculinidade das personagens e idealização do que é "ser

homem", como se observa na Imagem 4. Esses resultados da ferramenta já dão indícios sobre a construção de um cenário de masculinidades exacerbadas:

Imagem 4 – captura de tela das linhas de concordância de "macho"

Line	Left context	Key	Right context	File
1	o último momento . " Lírio é	macho	" , murmurou Liroca para si mesmo	CTN1-1
2	para si mesmo . " Lírio é	macho	. " Sempre que ia entrar num	CTN1-1
3	repetia estas palavras : " Lírio é	macho	" . Levantou-se devagarinho , apertando a	CTN1-1
4	passos . Cinco segundos . Lírio é	macho	, Lírio é macho . José Lírio	CTN1-1
5	. Lírio é macho , Lírio é	macho	. José Lírio continuava imóvel , olhando	CTN1-1
6	. Pontaria , Liroca . Lírio é	macho	. Vamos . É ÁMete bala .	CTN1-1
7	é negócio de mulher . É de	macho	. Maria Valéria abranda um pouco a	CTN1-1
8	reprimir as lágrimas . Um homem bem	macho	não chora nunca , haja o que	CTN1-1
9	vergonha ! O que ela queria era	macho	. E pensava em Pedro só porque	CTN1-2
10	ou para negro . Um homem bem	macho	devia saber manejar a espada , a	CTN1-2
11	a nuca , a bela cabeça de	macho	altivamente erguida , e aquele seu olhar	CTN1-2
12	– Por aqui hai também muito homem	macho	. Houve um silêncio desconfiado . Juvenal	CTN1-2
13	podia jurar que nunca vira cara de	macho	mais insinuante . Os cabelos do capitão	CTN1-2
14	. Há um ditado : " Cambará	macho	não morre na cama" . E ao	CTN1-2
15	lhe digo . – Nunca nenhum Cambará	macho	conseguiu passar dos cinquenta anos . Para	CTN1-2

Fonte: Elaborado pelos autores com dados extraídos do Wmatrix 7 (Rayson, 2009).

A subcategoria Parentesco (S4) revela um traço característico da própria concepção de *O Continente* e da trilogia "O Tempo e o Vento": a importância da família, já que toda a trama do romance se circunscreve em torno das linhagens dos Terra e dos Cambará e de seus descendentes e ascendentes, conflitos e dramas familiares em meio ao contexto histórico, algo bem observado na crítica literária de Bordini e Zilberman (2004) e Santos (2005). Ressalta-se que essa subcategoria teria um tamanho de efeito muito maior se as ocorrências do termo "filho" também tivessem a etiqueta S4, devido à alta incidência no romance.

4.2.10 T - Tempo

O DSNS T apresenta palavras relacionadas à passagem do tempo de maneira geral e à ideia abstrata de tempo como conceito. Em *O Continente*, são predominantes as subcategorias "Tempo" (T1), "Tempo: início" (T2+), "Tempo: momentâneo" (T1.2) e "Tempo: presente; simultâneo" (T1.1.2). A presença das subcategorias T1.1.2, T1.2 e T2+ entre os DS mais salientes está relacionada à marcação e passagem temporal na narrativa por meio de advérbios, locuções adverbiais e verbos, não sendo objeto do presente estudo, que é a análise temática dos campos semânticos.

A subcategoria "Tempo" (T1) possui entre o maior número de ocorrências os advérbios "nunca" (316) e "vezes" (158), ambos também utilizados para marcação temporal no romance. O maior número de ocorrências dessa subcategoria, porém, vai para o substantivo "tempo" (358). Dessas ocorrências destacam-se algumas considerações: ordenando as ocorrências na ferramenta pela palavra imediatamente à esquerda de "tempo" constatou-se que a locução adverbial "(ao) mesmo tempo" é o uso predominante da palavra com 73 recorrências expressando simultaneidade, seguido pelo uso nominal "o tempo" (58) e "algum tempo" (45).

Enquanto "mesmo tempo" e "algum tempo" desempenham exclusivamente a função acessória de adjunto adverbial nas linhas de concordância, como em "[...] cansada da corrida e ao mesmo tempo surpreendida de ter vindo" e em "Por algum tempo avistou as ruínas do rancho, [...]" é possível observar nas ocorrências nominais de "o tempo" um uso bastante diversificado do ponto de vista semântico, que será explorado a seguir.

O uso predominante de "o tempo" é o de sua principal acepção dicionarizada: "Aquilo que é medido em horas, dias, meses ou anos; período; duração" (Aulete, 2024), totalizando 47 ocorrências, como em "Como o tempo custa a passar quando a gente espera [...]", sendo 12 desses resultados relacionados ao instante em que algo ocorre, como em "[...] durante todo o tempo das danças [...]" e "ficava o tempo todo como que sobre brasas [...]". Cabe ressaltar também que, dentre o total mencionado, há 10 linhas de concordância que chamam a atenção por conterem prosopopeia, um recurso da função poética da linguagem comumente utilizado na literatura em que um objeto ou conceito abstrato é personificado. São exemplos: "Parece que o vento maneia o tempo"; "ele tinha a impressão de ver o tempo parado sobre os telhados [...]" e "E o tempo continuava a andar num tranco lento [...]" (nossa ênfase).

Há três registros como "Época, lapso de tempo futuro ou passado" (Aulete, 2024): "Recordou o tempo em que ela era menina [...]". Além desses, um dos usos tem a acepção de "Estação" – "[...] pois o inverno era o tempo que mais custava [...]" e mais quatro como "Estado meteorológico": "[...] outras maneiras de um campeiro prever o tempo sem precisar olhar naquelas geringonças [...]". Por fim, há três ocorrências em que é utilizado o sentido figurado: "[...] puxou a adaga. 'Fechou o tempo', contava Fandango" e "como o tempo é remédio que cura tudo".

A saliência desse DSNS ilustra a relevância que o tempo possui na narrativa, pois nesses resultados é possível visualizar de que modo esse conceito é explorado pelo romancista, como ele é utilizado linguisticamente e também analisar como as personagens da trama são impactadas por esse tema: percebido ora veloz e incontrolável - "Como o tempo voa – refletiu Rodrigo" -, ora irremediavelmente lento - "Na venda o tempo se arrastava como lesma"; em alguns momentos fonte de angústias - "Mas por que será que o tempo custa tanto a passar quando há guerra"; em outros, a solução de todos os males - "Mas, como o tempo é remédio que cura tudo [...]".

4.2.11 DSNS: X - Ações Psicológicas, Estados e Processos

Esse DSNS demonstra algumas das estratégias estilísticas do autor para o desenvolvimento das personagens, uma vez que é a partir desse domínio que são expressas as percepções sensoriais e ações mentais. Agrupam-se nesse domínio as subcategorias a seguir, com seus respectivos resultados: "Som: Silêncio" (X3.2-) - com o termo "silêncio" (219) como o de maior ocorrência, seguido por "calado" (33); "Sensorial: Visão" (X3.4), com "olhando" (140) em primeiro lugar, seguido de "olhou" (131), "olhava" (103) e "olhar" (100); em "Objeto mental: Objeto conceitual" (X4.1), há o termo "ideia" (78), seguido por "pensamentos" (57) e "pensamento" (47); "Sensorial" (X3) tem apenas "sensação" (67) e "sensações" (1) como termos; por fim, "Sensorial: Olfato" (X3.5) traz "cheiro" (104) como termo mais saliente e "Sensorial: sonoro" (X3.2) apresenta "som" (29), "sons (13) e "batidas" (11), enquanto a subcategoria "Ações psicológicas, estados e processos" (X1) apresenta "mente" (34), seguida por instinto (4) e mentes (1).

Conclui-se, a partir de uma visão geral a respeito da utilização desse DSNS, que é relevante para a caracterização das personagens do romance a apresentação das suas percepções sensoriais, impressões, pensamentos e ideias. Uma análise aprofundada das linhas de concordância de alguns dos termos sinalizados acima pode também corroborar as considerações de Jacobi (2023) a respeito da melancolia que permeia a obra, uma vez que, na categoria X3, por exemplo, observa-se a prevalência de uma prosódia semântica negativa em torno do termo "sensação", como em "[...] uma sensação de desalento gelado a invadiu [...]".

5 Discussão

Uma análise dos resultados permite inferir algumas características do estilo literário de Erico Veríssimo: pode-se observar o uso dos sentidos (olfato, visão, audição, tato) como recursos estilísticos que realçam diferentes percepções das personagens e também estabelecem o tom e a ambientação da narrativa, como se conclui pela saliência desses domínios semânticos na obra. Esse resultado apresenta uma nova perspectiva sobre o estilo literário presente no romance, uma vez que esses elementos não foram destacados na crítica de *O Continente*.

Outro fator que não teve proeminência na crítica literária é o papel não casual desempenhado pela música em diferentes momentos do romance, como observado ao se analisar o DS K2 "Música e atividades relacionadas" e suas linhas de concordância. O uso polissêmico como estratégia estilística que o autor faz das palavras, como observado nas análises de "ar" e "fogo" também se revela um aspecto pouco explorado nos estudos de *O Continente*.

A conformação do romance do ponto de vista temático, histórico e regional está presente na saliência dos DS "Tempo", "Lugares", "Objetos em geral", "Nomes geográficos", "Religião e o sobrenatural", "Cores e padrões de cor", "Móveis e acessórios domésticos", além, é claro, da proeminência dos DS "Guerra, defesa e o exército; armas" e "Formal/não amigável" e endossa os estudos do romance como Zilberman (2004b). A análise dos DS relativos aos conflitos históricos retratados na trama, além do estudo das palavras caracterizadoras da violência, confrontamento e inimizades em *O Continente* endossam a visão de Garcia e Lisboa Filho (2013) e de Bisol e Porto (2015) sobre a naturalização da violência no romance.

Além disso, a prevalência de "Ações, estados e processos psicológicos" e "Ações, estados e processos emocionais gerais" corrobora a percepção de Bordini (2004) de que o romance enfoca a caracterização psicológica das personagens em detrimento de suas descrições físicas e do espaço diegético.

Há destaques, além disso, para o DS "Pessoas: Masculino", indicando uma prevalência desse gênero na trama, e "Estacionário", que demonstra a postura e linguagem corporal dos personagens ao longo das cenas, bem como "Fumar e drogas não medicinais" revela o comportamento predominante desses personagens gaúchos, enquanto os DS "Pessoas" e "Parentesco" sinalizam as relações familiares do romance.

Os resultados encontrados com o Wmatrix permitem uma exploração detalhada de cada um dos DS indicados como salientes, uma vez que é possível verificar que palavras pertencem a cada DS e também analisá-las em seu contexto por meio das linhas de concordância e conferir falsos positivos e possíveis inconsistências dos resultados. Na seção a seguir, pretende-se explorar alguns dos DS em relação às suas palavras-chave e seu papel na construção do estilo literário de *O Continente* e também tecer possíveis correlações com a crítica estabelecida da obra.

5.1 Wmatrix: Limitações da ferramenta

O primeiro fato a se ponderar, e esse diz respeito a uma das atuais limitações do sistema PyMUSAS na anotação semântica do português, é que ele depende do léxico (anotado manualmente) com o qual é treinado para uma etiquetagem efetiva. No caso do léxico para o portu-

guês, que foi feito a partir de tradução automática do léxico da versão inglesa, foi verificado na fonte do etiquetador que há anotações inconsistentes, que não correspondem ao modo como a língua portuguesa funciona, especialmente quanto às anotações morfossintáticas, que fazem parte da etapa automatizada, necessária para que o código do anotador analise o sentido das palavras e determine sua classificação. Esse fato causa dificuldades ao anotador para desambiguar palavras polissêmicas, lidar com nomes próprios e palavras compostas, por exemplo.

Observou-se, nesta pesquisa, que o anotador falhou com a desambiguação de algumas palavras, classificando-as equivocadamente, desconsiderando seus contextos de uso. Como exemplo, é possível observar a subcategoria "Saudável" (B2+), que revelou um falso positivo decorrente do erro de desambiguação da palavra "cura", que ocorre 45 vezes, sendo apenas 3 com relação à saúde. Em 40 contextos, a palavra é utilizada pelo narrador para se referir a uma personagem religiosa nos primeiros capítulos do romance, conforme demonstrado nas linhas de concordância apresentadas na amostra da Imagem 5:

Imagem 5 – Captura de tela das primeiras dez linhas de concordância do termo "cura"

Line	Left context	Key	Right context	File
1	acesas . Preciso contar meu sonho ao	cura	– decidiu ele . E entrou no	CTN1-1
2	tempo em meditação . Por fim o	cura	ergueu-se , e Alonzo fez o mesmo	CTN1-1
3	velas e das lamparinas o rosto do	cura	tinha um tom alaranjado . Era uma	CTN1-1
4	nos sentar ali ... Sentaram-se . O	cura	respirava fundo . Era um homem corpulento	CTN1-1
5	à missão para servir de companheiro ao	cura	, que pouco sabia de sua vida	CTN1-1
6	instante . – Vamos – disse o	cura	– , conta tudo . – Nesse	CTN1-1
7	não ouso mencionar neste templo . O	cura	fez com a cabeça um grave sinal	CTN1-1
8	– É só ? – perguntou o	cura	. – É só . Pelo menos	CTN1-1
9	mais tarde , quando se confessasse ao	cura	. Mas era preciso contar agora .	CTN1-1
10	de minha vida . De novo o	cura	estava de cabeça baixa e olhos cerrados	CTN1-1

Fonte: Elaborado pelos autores com dados extraídos do Wmatrix 7 (Rayson, 2009).

Também é possível notar inconsistências de anotação dentro dos DS, tal como na subcategoria "Doença" (B2-), que apresenta variações de anotação ilutradas aqui com algumas linhas de concordância da palavra frio (92 ocorrências): "[...] envolvido pelo ar frio da noite. [...]", "[...] tocam um objeto frio. [...]" e "[...] Estava exausta, com um frio de morte no corpo [...]". Essa discrepância se dá porque há um DS específico para a palavra em questão: "Temperatura: frio" (O4.6-). Outros termos com discrepâncias também classificados como B2- são "mal" (25), em diversas das ocorrências utilizado como advérbio de modo, por exemplo, "[...] a luz da lamparina mal alumia. [...]" e "raiva" (30), relacionada à emoção sentida e não à doença, esse último caso observável na Imagem 6.

Imagem 6 – Captura de tela das linhas de concordância do termo "raiva"

Line	Left context	Key	Right context	File
1	esperando , subitamente aquecido pelo fogo duma	raiva	nascente . " Dou-lhe um pontapé no	CTN1-1
2	em artes mágicas , ficaram loucos de	raiva	quando foram vencidos nas Cruzadas pelos cristãos	CTN1-2
3	coxas como uma grande aranha-caranguejeira . Numa	raiva	Anaagarrou com fúria os cabelos de Pedro	CTN1-2
4	mundo . Mas queria viver também de	raiva	, de birra . A sorte andava	CTN1-2
5	em cima da mesa num gesto de	raiva	e ao mesmo tempo de alegria .	CTN1-2

Fonte: Elaborado pelos autores com dados extraídos do Wmatrix 7 (Rayson, 2009).

Justamente pelo exposto acima, um dos primeiros passos da análise com a ferramenta diz respeito à análise manual criteriosa das listas de palavras indicadas do campo semântico que se quer analisar, a fim de se observarem possíveis discrepâncias, antes de chegar a conclusões baseadas apenas no levantamento automático. Também sugere-se uma análise atenta das

palavras classificadas na categoria Z99 - "Não Classificado" e que carecem de anotação manual a fim de serem incluídas no léxico do sistema PyMUSAS para aumentar sua efetividade. Isso porque o anotador não reconheceu palavras de variedades diatópicas, como gaúchos, fandango, chimarrão, bombachas, nem diacrônicas, como vosmecê e vossuncê, e até em alguns casos, palavras simples como filha, lamparina e mamãe, etiquetando todas como Z99. Além disso, o sistema ainda não lida com unidades multi-palavras, classificando-as individualmente.

Outra limitação está nos equívocos de etiquetagem relacionados à sensibilidade de maiúsculas e minúsculas, como observado na etiquetagem de "fé", que recebeu a etiqueta E6+ - "Confiante", inclusive nos casos em que se referia ao nome de cidade - Santa Fé -, em vez de ser incluído na categoria Z, destinada a nomes próprios em geral. O mesmo ocorre com diversos nomes, como Ana Terra, Maneco Terra e outras pessoas dessa descendência, cujo sobrenome foi classificado como "Ações Gerais/ fazer" (A1.1.1), quando deveriam ser classificados como Z99.

6 Considerações finais

A EC é um campo de estudos promissor gestado no encontro de duas grandes metodologias de pesquisa: de um lado, as técnicas e recursos de PLN de que dispõe a LC, que já possui diversas ferramentas bem estabelecidas e estáveis à disposição dos estudos linguísticos e, de outro, um campo já consagrado no âmbito da análise literária, que é a Estilística tradicional. Com essa combinação, se torna mais fácil escrutinar uma ou mais obras com alguns cliques e revelar padrões e proeminências que enriqueçam uma crítica literária de forma quantitativa e qualitativa. Não se pretendeu aqui, todavia, esgotar as possibilidades de análise de cada DS relevante em *O Continente*, uma vez que esta pesquisa visou, sobretudo, avaliar a eficácia da versão 7 do Wmatrix (Rayson, 2009) e, complementarmente, investigar e demonstrar de que maneira a EC pode enriquecer e contribuir à crítica já existente da obra.

Considera-se que este trabalho alcançou seus objetivos de modo satisfatório, uma vez que foi possível vislumbrar aplicações da LC para a análise de uma obra literária, aproximando-se do que se espera ser a abordagem metodológica da EC. Ter a oportunidade de avaliar uma ferramenta promissora como o Wmatrix 7 (Rayson, 2009), que está implementando em sua interface o sistema PyMUSAS com uma série de novas funcionalidades, é uma experiência enriquecedora. Constatar, também, suas limitações e desafios em relação à classificação e desambiguação de palavras na língua portuguesa aponta caminhos possíveis de se contribuir nos estudos em EC, uma vez que os resultados podem ser aproveitados no aprimoramento da ferramenta.

Reitera-se, também, o caráter experimental desta pesquisa com o romance *O Continente*, uma vez que resultados mais precisos da análise de domínios semânticos é influenciado pela eficácia que a ferramenta testada pode entregar. Avalia-se que alguns dos resultados podem ser alterados significativamente quando a interface for capaz de classificar de forma apropriada os termos que recaem inadequadamente na categoria Z99, por exemplo, junto a outras palavras que o etiquetador PyMUSAS ainda não é capaz de reconhecer.

Por fim, esta pesquisa abre um leque de possibilidades futuras, uma vez que estudos preliminares com a trilogia integral de *O Tempo e o Vento* apontaram diferenças nos resultados obtidos, como o surgimento de outros DS, como "Política" e "Roupas e pertences pessoais". Pesquisas futuras com essa ferramenta podem contribuir tanto em relação a estudos da obra

consagrada de Erico Veríssimo, assim como em relação à ampliação e ao refinamento de um léxico para PLN que proveja recursos mais robustos para que os etiquetadores semânticos do PyMUSAS sejam mais efetivos.

Referências

ALVES, M. de B. *Tratado das gentes d'O Continente por uma definição da identidade gaúcha*. 2005. 123f. Dissertação (Mestrado em Literatura Comparada) - Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2005. Disponível em: https://lume.ufrgs.br/handle/10183/4432 Acesso em: 04 mai. 2023

ANTHONY, L. *AntConc* (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University, 2021. Disponível em: https://www.laurenceanthony.net/software/antconc/ Acesso em: 21 out. 2023

ARCHER, D.; WILSON, A.; RAYSON, P. *Introduction to the USAS Category System*. Outubro 2002. Disponível em: https://ucrel.lancs.ac.uk/usas/usas_guide.pdf. Acesso em: 04 mai. 2023

AULETE, C. *Aulete Digital – Dicionário contemporâneo da língua portuguesa*: Dicionário Caldas Aulete, vs online. Disponível em: https://www.aulete.com.br/ Acesso em 14 jul. 2024.

AUSTEN, J. *Pride and prejudice*. [S.l.]: [1813] 2002. Disponível em: https://www.gutenberg.org/ebooks/1342 . Acesso em: 23 fev. 2024.

BAKER, M. *Corpus* linguistics and translation studies: Implications and applications. In: BAKER, M.; FRANCIS, G.; TOGNINI-BONELLI, E. (eds.) *Text and Technology*: In Honour of John Sinclair. Amsterdam/Philadelphia: John Benjamins, 1993. p. 233-250.

BAKER, P. *Using Corpora in Discourse Analysis*. Bloomsbury Publishing, 2023.

BALOSSI, G. *A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves*. John Benjamins, 2014.

BALOSSI, G. Key Pronouns through Wmatrix in a Novel of Formation: Conrad's The Shadow-Line. *Umanistica Digitale*, Bologna, v. 5, n. 9, 79-96, 2020. DOI: https://doi.org/10.6092/issn.2532-8816/10542

BERBER SARDINHA, T. *Lingüística de Corpus*. Barueri: Manole, 2004.

BIBER, D.; REPPEN, R. What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition*, Cambridge, v.24, n.2, p. 199-208, 2002. DOI: https://doi.org/10.1017/S0272263102002048

BICK, E. A FrameNet for Danish. In: NORDIC CONFERENCE OF COMPUTATIONAL LINGUISTICS (NODALIDA 2011), 18, Northern European Association for Language Technology (NEALT), 2011, Tartu. *Proceedings Series*, Tartu: Tartu University Library, 2011. p. 34-41.

BICK, E. PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. *In*: BERBER SARDINHA, T.; FERREIRA, T. de L. S. B. (ed.). *Working with Portuguese Corpora*. London/New York: Bloomsbury Academic, 2014. p. 279-302.

BICK, E. PFN-PT: Anotação semântica automática: um novo Framenet para o português. *Domínios de Linguagem*, Uberlândia, v.16, n.4, p. 1401-1435, 2022. DOI: https://doi.org/10.14393/DL52-v16n4a2022-7

BISOL, L. V.; PORTO, L. T. Violência e memória: uma leitura do romance *O Continente*, de Erico Verissimo. *Navegações*, [S. l.], v. 8, n. 2, p. 146–155, 2015. DOI: 10.15448/1983-4276.2015.2.20532.

BLAKE, W. *Songs of Innocence*.Londres: Edição do Autor, 1789.

BLAKE, W. *Songs of Innocence and of Experience:* shewing the two contrary states of the human soul. Londres: Edição do Autor, 1794. Disponível em: https://www.blakearchive.org/copy/songsie.b?descId=songsie.b.illbk.01 . Acesso em: 8 ago. 2024.

BORDINI, M. D. G. *O Continente*: um romance de formação? Pós-colonialismo e identidade política. In: BORDINI, M. D. G.; ZILBERMAN, R. (ed.) *O tempo e o vento: história, invenção e metamorfose*. Porto Alegre: Edipucrs, 2004. p.65-86.

BORDINI, M. D. G.; ZILBERMAN, R. (ed.) *O tempo e o vento: história, invenção e metamorfose*. Porto Alegre: Edipucrs, 2004.

BRAGA, G. da S. *Corpus* stylistics in translation-oriented text analysis: approaching the work of Denton Welch from a functionalist perspective. *Diacrítica*, Coimbra, v. 32, n. 3, p. 227-248, 2020. DOI: https://doi.org/10.21814/diacritica.5163

BREZINA, V. & PLATT, W. *#LancsBox X* [software], Lancaster University, 2024. Disponível em: http://lancsbox.lancs.ac.uk.

CAN, T.; CANGIR, H. A warring style: A *corpus* stylistic analysis of the First World War poetry. *Digital Scholarship in the Humanities*, Cidade, v. 37, n. 3, p. 660-680, 2022.

ČERMÁKOVÁ, A. Traduzindo literatura infantil: algumas ideias de estilística de *corpus*. *Ilha do Desterro*, Florianópolis, v. 71, n. 1, p. 117-134, jan/abr 2018. DOI: https://doi.org/10.5007/2175-8026.2018v71n1p117

CULPEPER, J.; ARCHER, D.; RAYSON, P. Love - 'A familiar of a devil'? An exploration of key domains in Shakespeare's comedies and tragedies. In: ARCHER, D. (Ed.), *What's in a word-list?* Investigating word frequency and keyword extraction. Digital research in the Arts and Humanities. Ashgate, Farnham, 2009, pp. 136-157. Disponível em: https://eprints.lancs.ac.uk/id/eprint/12671/1/ACR_MethNet06.pdf Acesso em 01 out. 2024

EVERT, S. How Random is a *Corpus*? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, Berlin, v. 54, n. 2, 2006, pp. 177-190. DOI: 10.1515/zaa-2006-0208

FISCHER, L. A. Sobre a vigência do Regionalismo no Brasil. *Terceira Margem*. Rio de Janeiro, v. 12, n. 19, p. 103-117, ago./dez. 2008. DOI: https://doi.org/10.55702/3m.v12i19.11058

FISCHER-STARCKE, B. Keywords and frequent phrases of Jane Austen's Pride and Prejudice: A *corpus*-stylistic analysis. *International Journal of Corpus Linguistics*, Amsterdam, v. 14, n. 4, p. 492-523, 2009. DOI: https://doi.org/10.1075/ijcl.14.4.03fis

FOWLES, J. *The Magus*. 1 ed. London: Jonathan Cape, 1965.

GABRIELATOS, C. Keyness analysis: Nature, metrics and techniques. In: TAYLOR, C.; MARCHI, A. (eds.).*Corpus approaches to discourse*: A critical review. New York: Routledge, 2018. p. 225–258.

GARCIA, C.; LISBOA FILHO, F. Representações da identidade do gaúcho e da violência: uma leitura de "O Tempo e o Vento". *Violência intrafamiliar: discutindo facetas e possibilidades*. Jundiaí: Paco Editorial, 2013. p. 171-182, Disponível em: https://www.ufsm.br/app/uploads/sites/513/2020/08/REPRESENTA%C3%87%C3%95ES-DA-IDENTIDADE-DO-GA%C3%9ACHO-E-DA-VIOL%C3%8ANCIA-UMA-LEITURA-DE-O-TEMPO-E-O-VENTO.pdf. Acesso em: 22 Ago. 2025

GINZBURG, J. *Literatura, violência e melancolia*. Campinas, SP: Autores Associados, 2013.

HALBWACHS, M. *A memória coletiva*. Trad. Beatriz Sidou. São Paulo: Centauro, 2006.

IBRAHIM, W. M. A. Utilizing *corpus* stylistics to facilitate literary analysis: An assessment of the effectiveness of semantic domains in identifying major literary themes in a selection of Charles Dickens novels. *AJELP: Asian Journal of English Language and Pedagogy*, Kedah, v. 10, n. 1, p. 114-138, 2022. DOI: https://doi.org/10.37134/ajelp.vol10.1.9.2022

JACOBI, L. dos S. *A construção da melancolia em O continente, de Erico Verissimo: a dor renitente do povo gaúcho*. 2023. 95 f. Dissertação (Mestrado em Letras) — Escola de Humanidades, Pontifícia Universidade Católica do Rio Grande do Sul, 2023. Disponível em: https://tede2.pucrs.br/tede2/handle/tede/11068 Acesso em 22 Ago. 2025

KAY, J. *The adoption papers*. 1 ed. Newcastle: Bloodaxe Books, 1991.

KILGARRIFF, A. Comparing corpora. *International Journal of Corpus Linguistics*, Amsterdam, v. 6, n. 1, 97-133, 2001. DOI: https://doi.org/10.1075/ijcl.6.1.05kil

KILGARRIFF, A. et al. The Sketch Engine: ten years on. *Lexicography*, Toronto, v. 1, p. 7-36, 2014. DOI: https://doi.org/10.1007/s40607-014-0009-9

KRISHNAMURTHY, R. *Corpus*-driven lexicography. *International Journal of Lexicography*, Oxford, v. 21, n. 3, 231-242, 2008. DOI: https://doi.org/10.1093/ijl/ecn028

LEECH, G. New resources, or just better old ones? The Holy Grail of representativeness. *In*: Hundt, M.; Nesselhauf, N.; Biewer, C. (eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 2007. p. 133-149. DOI: https://doi.org/10.1163/9789401203791_009

MAHLBERG, M., SMITH, C. & PRESTON, S. Phrases in literary contexts: patterns and distributions of suspensions in Dickens's novels. *International Journal of Corpus Linguistics*, Amsterdam, v. 18, n. 1, p. 35-56, 2013. DOI: https://doi.org/10.1075/ijcl.18.1.05mah

MATURANA, M. del C. C. *Maternidad y voces poéticas en 'The Adoption Papers' de Jackie Kay*. 2012. Tese (Doutorado em Literatura e Linguística Inglesas) - Universidad de Granada, 2012. Disponível em: https://digibug.ugr.es/bitstream/handle/10481/70114/Tesis_Coral%20Calvo%20Maturana_16Oct.pdf?sequence=4&isAllowed=y Acesso em: 22 Ago. 2024

MCARTHUR, Tom. *Longman lexicon of contemporary English*. Harlow: Longman, 1981.

McENERY, T.; HARDIE, A. *Corpus Linguistics*. Cambridge: Cambridge University Press, 2012.

McINTYRE, D.; ARCHER, D. A *corpus*-based approach to mind style. *Journal of Literary Semantics*, Berlin, v. 39, n. 2, 167-182, 2010. DOI: https://doi.org/10.1515/jls-2021-2045

McINTYRE, D.; WALKER, B. How can corpora be used to explore the language of poetry and drama? In: O'KEEFFE, A.; MCCARTHY, M. (Eds.). *The Routledge Handbook of Corpus Linguistics*. 1 ed. London: Routledge, 2010. p. 516-530.

McINTYRE, D.; WALKER, B. *Corpus Stylistics*: Theory and Practice. Edinburgh: Edinburgh University Press, 2019.

NORD, C. *Textanalyse und Übersetzen*. 4 ed. Tübingen: Julius Groos, 2009.

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

PIAO, S.; BIANCHI, F.; DAYRELL, C.; D'EGÍDIO, A.; RAYSON, P. Development of the Multilingual Semantic Annotation System. In: *Proceedings of the 2015 Conference of The North American Chapter of the Association for Computational Linguistics:* Human Language Technologies, 2015, Denver, Colorado. Association for Computational Linguistics, 2015. p. 1268-1274. Disponível em: https://aclanthology.org/N15-1137.pdf Acesso em 25 Ago. 2025

PIMENTA, A.; NOVODVORSKI, A. Linguística de *Corpus* e Fraseologia: um estudo das colocações com "Feito" em Grande Sertão: Veredas. *Linguagem*: *Estudos e Pesquisas*, Goiânia, v. 21, n. 1, 2018. DOI: https://doi.org/10.5216/lep.v21i1.52224.

POJANAPUNYA, P.; TODD, R. W. Log-likelihood and odds ratio: keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, Berlin, v. 14, n. 1, p. 133-167, 2018. DOI: https://doi.org/10.1515/cllt-2015-0030

RANGEL, M. A. *Proposta metodológica para análises de obras literárias através da Linguística de Corpus:* o caso de Paulo Coelho. Dissertação (Mestrado em Estudos Linguísticos) - Universidade Federal de Uberlândia, 2022. DOI: http://doi.org/10.14393/ufu.di.2022.148

RAYSON, P. *Wmatrix: a web-based corpus processing environment*. Computing Department, Lancaster University, 2009.

RAYSON, P.; ARCHER, D. E.; PIAO, S.; MCENERY, T. THE UCREL SEMANTIC ANALYSIS SYSTEM. In: FOURTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 2004, Lisbon. *Proceedings...* Paris: European Language Resources Association (ELRA), 2004. p. 1813-1816. Disponível em: http://www.lancaster.ac.uk/staff/rayson/publications/usas_lrec04ws.pdf Acesso em: 12 Set. 2025

RAYSON, P.; BERRIDGE, D.; FRANCIS, B. Extending the Cochran rule for the comparison of word frequencies between corpora. In: INTERNATIONAL CONFERENCE ON STATISTICAL ANALYSIS OF TEXTUAL DATA (JADT), 7, 2004, *Proceedings...*, v. 2, Louvain-la-Neuve, Belgium, March 10-12, 2004, Presses universitaires de Louvain, pp. 926 - 936. ISBN 2-930344-50-4. Disponível em: http://ucrel.lancs.ac.uk/people/paul/publications/rbf04_jadt.pdf

RAYSON, P.; GARSIDE, R. Comparing corpora using frequency profiling. In: WORKSHOP ON COMPARING CORPORA, 2000, Hong Kong. *Proceedings...* Hong Kong: [s.n.], 2000. p. 1-6. Disponível em: https://aclanthology.org/W00-0901.pdf Acesso em 12 set. 2025

ROSA, J. G. *Ave, Palavra*. Rio de Janeiro, RJ: Nova Fronteira, 1985.

ROSA, L. M. da. *Literatura de ficção e Educação Geográfica:* Tradição inventada, modelação social e discurso na obra *O Continente*, de Erico Verissimo. 2019. 148f. Dissertação (Mestrado em Geografia) – Programa de Pós-Graduação em Geografia, Instituto de Ciências Humanas, Universidade Federal de Pelotas, Pelotas, 2019. Disponível em: https://guaiaca.ufpel.edu.br/handle/prefix/4733 Acesso em: 12 set. 2025

SANTOS, P. B. Aspectos do romance histórico em Erico Verissimo. *O Eixo e a Roda: Revista de Literatura Brasileira*, [S.l.], v. 11, p. 53-59, dez. 2005. DOI: https://doi.org/10.17851/2358-9787.11..53-59

SANTOS, D. A. dos. O tempo e o vento: romance histórico e romance político. *Cadernos do IL*, Porto Alegre, n. 39, p. 96-104, 2009. DOI: https://doi.org/10.22456/2236-6385.25205

SANTOS, D. A. dos. A crítica literária dos anos 40 e a autocrítica de Erico Verissimo. *Literatura em Debate*, Frederico Westphalen, v. 13, n. 24, p. 100-115, 2019.

SANTOS, P. B. Aspectos do romance histórico em Erico Verissimo. *O Eixo e a Roda: Revista de Literatura Brasileira*, [S.l.], v. 11, p. 53-59, dez. 2005. DOI: https://doi.org/10.17851/2358-9787.11..53-59

SCOTT, M., *WordSmith Tools version 9* (64 bit version) Stroud: Lexical Analysis, 2024.

SHORT, M. *Exploring the language of poems, plays and prose*. London: Longman, 1996.

TAGNIN, S. E. O. *O jeito que a gente diz*: combinações consagradas em inglês e português. Barueri: DISAL, 2013.

TEUBERT, W. *Corpus* Linguistics and Lexicography. *International Journal of Corpus Linguistics*, Amsterdam, v. 6, n. 1, p. 125-153, December 2001. DOI: https://doi.org/10.1075/IJCL.6.SI.11TEU

VATHANALAOHA, K. *Corpus* Stylistics in Contemporary English Dramas: Keywords and Semantic Fields of Delusions. *GEMA Online Journal of Language Studies*, Bangi, v. 22, n. 2, p. 43-62, 2022. DOI: https://doi.org/10.17576/gema-2022-2202-03

VERÍSSIMO, E. *O Continente.* Porto Alegre: Globo, 2013 [1949].

VERÍSSIMO, E. *O Retrato.* Porto Alegre: Globo, 1951.

VERÍSSIMO, E. *O Arquipélago.* Porto Alegre: Globo, 1961.

VERÍSSIMO, E. *O Tempo e o Vento*. 5 v. São Paulo: Companhia das Letras, 2004.

VITAL, Á. A. S. Do contato entre a Literatura, a Linguística de *Corpus* e o Processamento de Língua Natural: o caso dos anagramáticos de Guimarães Rosa. *Texto Livre*, Belo Horizonte, v. 15, p. e39316, 2022. DOI: https://doi.org/10.35699/1983-3652.2022.39316

WELCH, D. *Sickert at St. Peter's*. Horizon, 1942.

WYNNE, M. *Corpus* Stylistics in Principles and Practice. A Stylistic Exploration of John Fowles' The Magus. Yufang Ho. *Literary and Linguistic Computing*, Oxford, v. 27, n. 4, p. 474–476, December 2012. DOI: https://doi.org/10.1093/llc/fqs023

ZANETTIN, F. *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. New York: Routledge, 2012.

ZILBERMAN, R. O Tempo e Vento: História, mito, literatura. In: BORDINI, M. D. G.; ZILBERMAN, R. (ed.) *O tempo e o vento: história, invenção e metamorfose*. Porto Alegre: Edipucrs, 2004a. p. 21-48

ZILBERMAN, R. Luiza Silva Cambará – revendo a tradição do mito. *In:* BORDINI, M. D. G.; ZILBERMAN, R. (ed.) *O tempo e o vento: história, invenção e metamorfose*. Porto Alegre: Edipucrs, 2004b. p. 87-102.

ZYNGIER, S.; CARNEIRO, R. M. O.; NOVODVORSKI, A. Reflexões sobre a estilística e o ensino de literatura: uma entrevista com Sonia Zyngier. *Trabalhos em Linguística Aplicada*, Campinas, v. 62, n. 2, 388-399, 2023. DOI: https://doi.org/10.1590/01031813v62220238667192

Apêndice A

Versão integral do Quadro 2: dados exportados de DS salientes no CE

Class.	Etiqueta	LL	Domínio Semântico
1	B1	1211.70	Anatomia e fisiologia
2	E6+	600.07	Confiante
3	G3	422.68	Guerra, defesa e o exército; armas
4	M7	391.51	Lugares
5	T2+	288.24	Tempo: Início
6	X3.4	264.78	Sensorial: Visão
7	N3.8+	200.00	Velocidade: Rápido
8	Z2	193.57	Nomes Geográficos
9	O1.3	188.07	Substâncias e materiais: Gás
10	N1	152.53	Números
11	X3.2-	150.23	Som: Silencioso
12	Q2.1	140.02	Fala: Comunicativo
13	N6	136.11	Frequência
14	S9	104.62	Religião e o sobrenatural
15	H5	99.63	Mobília e utensílios domésticos
16	X3	91.22	Sensorial
17	M6	84.66	Localização
18	X3.5	84.38	Ssensorial: Olfato
19	O2	82.76	Objetos em geral
20	F1	69.46	Comida
21	M8	68.63	Estacionário
22	O4.6-	68.01	Temperatura: frio
23	Z99	67.67	Não Classificável
24	A13.6	65.47	Grau: Diminutivos
25	F3	57.94	Fumo e drogas não medicinais
26	B3	57.39	Medicina e tratamentos medicinais
27	S2.2	56.14	Pessoas: Masculino
28	H2	54.37	Partes de construções
29	O4.6+	52.30	Temperatura: quente / em chamas
30	Z5	51.70	Caixa Gramatical
31	A8	50.92	Parecer
32	S2	45.35	Pessoas
33	N5-	44.61	Quantidades: pouco
34	O4.3	42.79	Cor e padrões de cores
35	S4	39.36	Família
36	B2+	38.93	Saudável
37	O4.6	37.75	Temperatura: quente / em chamas
38	K2	28.18	Música e atividades relacionadas

39	T1.2	26.45	Tempo: Momentâneo
40	S2.1	25.75	Pessoas: Feminino
41	T1	23.79	Tempo
42	S5	21.91	Grupos e afiliações
43	F4	20.72	Agricultura e horticultura
44	W1	19.66	O Universo
45	S7.1-	18.61	Sem Poder
46	A1.1.1	17.78	Ações em geral / fazer
47	A13.3	17.62	Grau: Aumentativo
48	S1.2.1-	16.81	Formal/ Não Amigável
49	X3.2	15.99	Sensorial: Som
50	W2	14.79	Luz
51	O4.2	14.26	Julgamento de aparência
52	N5++	13.07	Quantidades: bastante/muito
53	N3.3	13.01	Medidas: Distância
54	T1.1.2	12.76	Tempo: Presente, simultâneo
55	N3.2---	10.99	Tamanho: Pequeno

Fonte: próprios autores [tradução própria]

Apêndice B

Quadro de Domínios Semânticos agrupado por Categoria Geral

Domínio Semântico - Categoria geral	Etiqueta	Log Likelihood	DS - Subcategoria
A - Termos abstratos e gerais	A13.6	65.47	Grau: Diminutivos
	A8	50.92	Parecer
	A1.1.1	17.78	Ações em geral / fazer
	A13.3	17.62	Grau: Aumentativo
B - O Corpo e o Indivíduo	B1	1211.70	Anatomia e fisiologia
	B3	57.39	Medicina e tratamentos medicinais
	B2+	38.93	Saudável
E - Ações Emocionais, Estados e Processos	E6+	600.07	Confiante
F - Alimentação e Agropecuária	F1	69.46	Comida
	F3	57.94	Fumo e drogas não medicinais
	F4	20.72	Agricultura e horticultura
G - Governo e Domínio Público	G3	422.68	Guerra, defesa e o exército; armas
H - Arquitetura, Construção, Casas e o Lar	H5	99.63	Mobília e utensílios domésticos
	H2	54.37	Partes de construções
K – Entretenimento, Esportes e Jogos	K2	28.18	Música e atividades relacionadas

M - Movimento, Localidades, Viagem e Transporte	M7	391.51	Lugares
	M6	84.66	Localização
	M8	68.63	Estacionário
N - Números e Medidas	N3.8+	200.00	Velocidade: Rápido
	N1	152.53	Números
	N6	136.11	Frequência
	N5-	44.61	Quantidades: pouco
	N5++	13.07	Quantidades: bastante/muito
	N3.3	13.01	Medidas: Distância
	N3.2---	10.99	Tamanho: Pequeno
O - Substâncias, Materiais, Objetos e Equipamentos	O1.3	188.07	Substâncias e materiais: Gás
	O2	82.76	Objetos em geral
	O4.6-	68.01	Temperatura: frio
	O4.6+	52.30	Temperatura: quente / em chamas
	O4.3	42.79	Cor e padrões de cores
	O4.6	37.75	Temperatura
	O4.2	14.26	Julgamento de aparência
Q - Ações Verbais, Estados e Processos	Q2.1	140.02	Fala: Comunicativo
S - Ações Sociais, Estados e Processos	S9	104.62	Religião e o sobrenatural
	S2.2	56.14	Pessoas: Masculino
	S2	45.35	Pessoas
	S4	39.36	Parentesco
	S2.1	25.75	Pessoas: Feminino
	S5	21.91	Grupos e afiliações
	S7.1-	18.61	Sem Poder
	S1.2.1-	16.81	Formal/ Não Amigável
T - Tempo	T2+	288.24	Tempo: Início
	T1.2	26.45	Tempo: Momentâneo
	T1	23.79	Tempo
	T1.1.2	12.76	Tempo: Presente, simultâneo
W - O Mundo e Nosso Ambiente	W1	19.66	O Universo
	W2	14.79	Luz
X - Ações Psicológicas, Estados e Processos	X3.4	264.78	Sensorial: Visão
	X3.2-	150.23	Som: Silencioso
	X3	91.22	Sensorial
	X3.5	84.38	Ssensorial: Olfato
	X3.2	15.99	Sensorial: Som
Z - Nomes e Palavras Gramaticais	Z2	193.57	Nomes Geográficos
	Z99	67.67	Não Classificável
	Z5	51.70	Caixa Gramatical

Fonte: próprios autores [tradução própria]