Pontifícia Universidade Católica de São Paulo PUC-SP

Stephan Arthur Solomon Hughes

Caracterização multidimensional da linguagem de postagens em bate-papos educacionais no Twitter

Doutorado em Linguística Aplicada e Estudos da Linguagem

São Paulo
2025

Pontifícia Universidade Católica de São Paulo
Faculdade de Filosofia, Comunicação, Letras e Artes
Programa de Pós-Graduação em Linguística Aplicada e Estudos da Linguagem

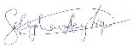
Stephan Arthur Solomon Hughes

Caracterização multidimensional da linguagem de postagens em bate-papos educacionais no Twitter

Tese apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo como parte dos requisitos para obtenção do título de Doutor em Linguística Aplicada e Estudos da Linguagem, sob orientação do Prof. Dr. Tony Berber Sardinha.

São Paulo
2025

Autorizo, exclusivamente para fins acadêmicos e científicos, a reprodução total ou parcial desta tese de doutorado por processos fotocopiadores ou eletrônicos, desde que devidamente citada.

Assinatura: 

Data: 01 de julho 2025

E-mail: stephan.hughes@gmail.com

Currículo Lattes: http://lattes.cnpq.br/7607359869436137

Orcid: https://orcid.org/0000-0002-6335-1591

HUGHES, S. A. S.
Caracterização multidimensional da linguagem de postagens em bate-papos educacionais no Twitter – São Paulo: 2025. xvi + 175 f.
Orientador: Professor Doutor Antonio Paulo Berber Sardinha
Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo. Programa de Pós-Graduação em Linguística Aplicada e Estudos da Linguagem, 2025.
Área de Concentração: Linguística Aplicada e Estudos da Linguagem

1. Educação. 2. Linguística de Corpus. 3. Análise Multidimensional. 4. Redes sociais.

Stephan Arthur Solomon Hughes
Caracterização multidimensional da linguagem de postagens em bate-papos educacionais no Twitter

Aprovada em: 14/8/2025.

Tese apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo como exigência parcial para obtenção de título de Doutor em Linguística Aplicada e Estudos da Linguagem, sob a orientação do Prof. Dr. Antonio Paulo Berber Sardinha.

Banca Examinadora

_____
Prof. Dr. Tony Berber Sardinha (Orientador)

_____
Profa. Dra. Sandra Madureira

_____
Prof. Dr. Carlos Henrique Kauffmann

_____
Profa. Dra. Maria Claudia Nunes Delfino

_____
Prof. Dr. Rafael Fonseca de Araújo

_____
Profa. Dra. Simone Vieira Resende

_____
Profa. Dra. Cristina Gil Borges

Dedico esta tese à minha esposa, Kátia, e à minha filha, Yasmin, por seu apoio incondicional ao longo de toda esta jornada.

Sem o amor e a força silenciosa de vocês, este trabalho não teria sido possível. Sou nada sem vocês.

.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq Processo 142304/2019-0

Agradecimentos

Ao professor Tony Berber Sardinha, meu orientador, meu sincero agradecimento. Esta tese simplesmente não teria sido possível sem sua orientação firme, generosa e instigante. Obrigado por compartilhar comigo seu rigor acadêmico, sua paixão pela pesquisa e sua confiança em meu trabalho desde o início.

À Maria Claudia Delfino, à Simone Vieira Resende, ao Rafael Fonseca de Araujo e ao Carlos Henrique Kauffman, minha profunda gratidão. Vocês foram muito mais do que avaliadores e membros da banca de qualificação — foram leitores atentos, críticos sensíveis e verdadeiros colaboradores nesta caminhada. Suas contribuições foram fundamentais para o amadurecimento deste trabalho.

Ao GELC – Grupo de Estudos de Linguística de Corpus, minha admiração e reconhecimento. Vocês personificam o que é fazer pesquisa de qualidade: um trabalho coletivo, plural, ético e comprometido com a construção colaborativa do conhecimento. Foi uma honra compartilhar esse espaço com vocês.

À Maria Lúcia dos Reis, um agradecimento especial. Você foi — e continua sendo — muito mais do que uma secretária para os pós-graduandos. Sua escuta atenta, sua disposição em ajudar e sua gentileza cotidiana foram faróis durante este percurso.

A cada um de vocês, o meu mais sincero muito obrigado.

"Sem dados, você é só mais uma pessoa com um palpite."
William Edwards Deming

HUGHES, Stephan Arthur Solomon. Caracterização multidimensional da linguagem de postagens em bate-papos educacionais no Twitter. 2025. 175 f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) – Programa de Linguística Aplicada e Estudos da Linguagem, Pontifícia Universidade Católica de São Paulo, São Paulo, 2025.

RESUMO

Esta tese tem como objetivo analisar a variação lexical-discursiva nos bate-papos educacionais realizados via Twitter (atualmente X), com base na metodologia da Análise Multidimensional Lexical (AMDL), conforme desenvolvida por Berber Sardinha e Fitzsimmons-Doolan (2025). Esses bate-papos configuram encontros síncronos mediados por redes sociais, nos quais educadores e demais participantes interagem por meio de tuítes em tempo real, discutindo temas ligados à educação, formação docente e inovação pedagógica. O corpus da pesquisa é composto por 127.418 textos (tuítes) extraídos de 1682 bate-papos distintos, totalizando 3.142.870 palavras. Os bate-papos aconteceram entre 2009 e 2022. A etiquetagem do corpus foi realizada com o programa TreeTagger. Foram identificadas sete dimensões lexicais, as quais foram interpretadas como discursos subjacentes aos tuítes. O trabalho contribui para os estudos de variação linguística em contextos digitais, ampliando o escopo da Linguística de Corpus aplicada à análise de registros mediados por tecnologia, e oferecendo subsídios para a compreensão da formação continuada docente em ambientes colaborativos online.

Palavras-chave: bate-papo educacional; Análise Multidimensional; Linguística de Corpus; Twitter; formação docente.

ABSTRACT

This thesis aims to analyze lexical-discursive variation in educational chats conducted via Twitter (currently X), using the methodology of Lexical Multidimensional Analysis (LMD Analysis) as developed by Berber Sardinha and Fitzsimmons-Doolan (2025). These chats are synchronous meetings mediated by social networks, in which educators and other participants interact through real-time tweets, discussing issues related to education, teacher training, and pedagogical innovation. The research corpus consists of 127,418 texts (tweets) drawn from 1,682 distinct chats, totaling 3,142,870 words. The chats took place between 2009 and 2022. The corpus was tagged using the TreeTagger program. Seven lexical dimensions were identified, which were interpreted as discourses underlying the tweets.The study contributes to research on linguistic variation in digital contexts by expanding the scope of Corpus Linguistics applied to the analysis of technology-mediated registers and by providing insights into the understanding of continuing teacher education in online collaborative environments.

Keywords: educational chat; Multidimensional Analysis; Corpus Linguistics; Twitter; teacher education.

Sumário

1. INTRODUÇÃO

Desde o fim do século passado, as novas tecnologias de informação e comunicação impulsionaram a sala de aula e o processo de ensino de aprendizagem. O professor se viu cada vez mais compelido a adotar recursos digitais para tornar aulas mais dinâmicas, credenciando-o para falar "o idioma" dos nativos digitais. As quatro habilidades do então novo século - a comunicação, a colaboração, a criatividade e o pensamento crítico se estabeleceram como mantra pedagógico, tendo nas mídias digitais, entre elas as redes sociais, possibilidades reais de um ensino centrado no aluno, visto que a época da informação exige novas alfabetizações tal como manejar devida fluência tecnológica, em especial a autoria ou a criação de conteúdo digital com fins pedagógicos. Implementar abordagens e conceitos como as metodologias ativas, o ensino baseado em projetos ou problemas, a sala de aula invertida, o ensino semipresencial, a aprendizagem assistida por computador, o ensino remoto e híbrido, aprendizagem móvel, para citar alguns, passou a ser viável uma vez tendo acesso a uma conexão a internet e a um aparelho digital (computador, tablet, celular).

As redes sociais emergiram como espaço potencializador de interação e troca de informações, o que representa para o docente uma forma de se relacionar profissionalmente com um número exponencial de seus pares por meio da tela digital. As redes sociais pressupõem a partilha de informações, conhecimentos, desejos e interesses". No universo educativo, as plataformas como Facebook, Instagram, Threads e Twitter despontam como espaços para a criação e desenvolvimento de comunidades de prática ou de aprendizagem movidas por uma intencionalidade educativa explícita, pois o ensino nestes ambientes virtuais de aprendizagem posiciona os discentes e docentes como coconstrutores de conhecimento mediados pelo mundo.

Os bate-papos educacionais no Twitter, objeto do presente estudo, satisfazem esta e outras condições, tais como a representabilidade de ambientes intelectuais, culturais e sociais, a facilitação e sustentação da aprendizagem e a promoção da interação, colaboração e do desenvolvimento de um sentimento de pertencimento dos seus membros. Desde 2009, os chamados bate-papos educacionais, são

encontros virtuais em tempo real que passaram a existir como importante alternativa ou complemento à aprendizagem proporcionada nos fóruns de discussão virtual, instrumento mais recorrido institucionalmente para promover o pensamento crítico, construção do conhecimento, treinamento profissional e formação continuada dos aprendizes.

Nesse contexto, o objetivo desta pesquisa é investigar o universo discursivo dos bate-papos educacionais na plataforma Twitter (atual X), a partir de um referencial teórico-metodológico que compreende a Linguística de Corpus (Berber Sardinha, 2004), além da Análise do Discurso assistida por Corpus (Gillings, Mautner, & Baker, 2023). O uso da Análise Multidimensional Lexical (Berber Sardinha & Fitzsimmons-Doolan, 2025), abordagem metodológica eleita para este estudo, possibilitou o exame de discursos em corpora de larga escala.

A Análise Multidimensional Lexical representa um desdobramento em relação à Análise Multidimensional (Biber, 1988), ao deslocar o foco da caracterização situacional para a identificação de discursos e ideologias por meio de agrupamentos lexicais. Essa abordagem mostra-se particularmente eficaz para o estudo de corpora digitais massivos e diversos, como os bate-papos educacionais no Twitter, ao permitir a detecção de padrões de coocorrência léxica que revelam práticas discursivas recorrentes e significativas. Por essa razão, a obra dos autores Berber Sardinha e Ftizsimmons-Doolan (2025) constitui a base teórico-metodológica central desta tese.

Essa abordagem é capaz de detectar grupos de itens lexicais que coocorrem em textos, de modo a revelar discursos recorrentes na linguagem. Para a análise, 127.418 tuítes foram submetidos à etiquetagem e lematização, o que é uma técnica que reduz as variantes de palavras à sua forma básica, seguidas de análise fatorial. Neste estudo, seguimos os procedimentos descritos por Berber Sardinha e Fitzsimmons-Doolan (2025), aplicando análise fatorial exploratória sobre variáveis léxicas selecionadas a partir do corpus de bate-papos educacionais, com o objetivo de identificar dimensões latentes de organização discursiva.

1.1. JUSTIFICATIVA

Compreender os usos discursivos das redes sociais em contextos de formação continuada docente por meio de uma análise multidimensional lexical conforme

proposta por Berber Sardinha e Fitzsimmons-Doolan (2025) revela-se metodologicamente relevante como contribuição para um estudo da variação léxico-discursiva dos bate-papos educacionais no Twitter, que constituíram-se entre 2009 e 2022 como espaços significativos de aprendizagem colaborativa e desenvolvimento profissional. Ao integrar a perspectiva da linguística de corpus com uma análise discursiva, esta pesquisa busca jogar luz sobre os discursos que permeiam e estruturam as conversas nesses bate-papos, considerando sua inserção na trajetória mais ampla das redes sociais aplicadas à educação e seu papel na formação continuada do professor.

1.2. REDES SOCIAIS VOLTADAS PARA PROFESSORES

Entender a aplicabilidade das redes sociais para fins educacionais constitui uma justificativa amplamente fundamentada para investigar a variação discursiva dos bate-papos no Twitter. Das pesquisas que estudam a relação rede social-ensino e aprendizagem, estas merecem destaque.

Partindo do pressuposto do que as pesquisas empíricas sobre o uso de redes sociais online por acadêmicos na literatura sobre tecnologia educacional são insignificantes, o autor buscou compreender as práticas naturalísticas de docentes nas redes sociais em geral e no Twitter em particular a partir de uma análise qualitativa de postagens de 45 sujeitos. Os resultados indicaram sete temas dominantes, sendo estes a partilha de informações, recursos e mídias relacionadas à prática profissional; a troca de informações sobre a sala de aula e os discentes, solicitação e oferta de ajuda profissional mútua, interação social, gestão da identidade digital e presença social, busca por e estabelecimento de conexões com outros usuários, e participação em redes online além do Twitter. Ao passo que essas descobertas desse autor ajudam o campo a compreender a prática emergente da participação de docentes em redes online com base nas postagens de um universo restrito de usuários, a pesquisa atual visa a aprofundar esses insights ao identificar os posicionamentos ideológicos e representações sociais que sustentam as falas dos participantes de 1082 bate-papos educacionais.

Um estudo com objetivos e propostas metodológicas semelhantes à pesquisa referenciada acima é o de Greenhalgh, Rosenberg e Wolf (2017) abordam a

percepção dos educadores sobre a eficácia dos Twitter chats como ferramenta de aprendizado contínuo. Os autores analisaram cerca de 9.300 tuítes relacionados ao programa de mestrado em Tecnologias Educacionais (MAET) da Michigan State University, coletados a partir de 12 hashtags institucionais e submetidos a amostragem e codificação temática. A metodologia combinou coleta automatizada de dados, análise qualitativa com codificação temática e refinamento de categorias até a saturação, apontando como resultado seis propósitos principais do uso do Twitter: contribuir e engajar-se em conversas disciplinares, construir comunidade, conectar-se a outras comunidades, pedir e oferecer apoio, além de usos incertos ou irrelevantes. Os autores concluíram que o Twitter atua como uma tecnologia fundacional no programa, por sustentar múltiplos contextos formais e informais de aprendizagem e favorecer uma ecologia de desenvolvimento profissional contínuo.

A pesquisa desses autores difere deste trabalho no que tange o arcabouço teórico-metodológico, os objetivos e seus resultados. Caracterizar os discursos que embasam e permeiam em dimensões lexicais permite ultrapassar o campo da semântica e semiótica e apurar os posicionamentos ideológicos dos usuários participantes dos bate-papos educacionais.

Até o momento, a literatura especializada sobre bate-papos educacionais no Twitter, tem se concentrado predominantemente em três vertentes: na categorização e análise da participação do usuário em comunidades de prática digitais; na investigação dos tópicos mais frequentemente abordados nas conversas entre usuários; e na avaliação dos estudos de caso e entrevistas realizadas com usuários participantes dos bate-papos educacionais.

Britt e Paulus (2016) apresentam um estudo de caso qualitativo sobre o #Edchat — um bate-papo semanal de educadores no X/Twitter — para investigar o desenvolvimento profissional informal à luz da teoria das comunidades de prática. Metodologicamente, as autoras combinaram três fontes de dados: observações sistemáticas das sessões do chat, entrevistas com participantes e análise de documentos arquivados (materiais e registros associados ao #Edchat), orientando a interpretação pelo enquadramento de comunidades de prática; essa estratégia permitiu examinar tanto dinâmicas interacionais quanto percepções dos membros. Os resultados indicam que o #Edchat exibe múltiplos indicadores de comunidade de

prática: convergência na descrição de quem pertence ao grupo, relações mútuas sustentadas, ausência de preâmbulos nas conversas (os participantes “entram falando”), formulação rápida do problema a discutir e fluxo acelerado de informação.

A partir disso, as autoras concluem que o #Edchat funciona como espaço fértil de aprendizagem profissional docente — com sobreposições a boas práticas de formação — e recomendam que gestores educacionais observem e aprendam com o que ocorre nesses ambientes de mídia social para melhor apoiar a aprendizagem de professores, recomendações que vão ao encontro do que se conclui a partir dos resultados desta tese: os discursos apontam a valorização da relação dialógica, do pertencimento e da prática docente colaborativa, assim como a importância em adquirir competências técnicas, praticar a inovação pedagógica, e realizar uma curadoria de recursos digitais. Ao analisar exemplos reais de língua em uso de mais de 1.082 hashtags (inclusive a usada no estudo de Britt e Paulus), a tese se apoia em dados quantitativos e qualitativos para descobrir que os participantes dos bate-papos educacionais não se limitam a reproduzir temas curriculares ou conteúdos programáticos, mas constroem coletivamente representações sobre o ser professor, o ensinar e o aprender em contextos digitais conectados, um dado que poderá constar no design de programas de formação continuada docente, tornando-os mais voltados para os interesses e preferências dos seus formandos.

Xing & Gao (2018) enfocaram a formação de comunidades de prática através de hashtags educacionais, utilizando análise de redes sociais para mapear interconexões. Analisam por que alguns educadores permanecem engajados em comunidades de aprendizagem profissional no Twitter (hashtag #edchat) enquanto outros desistem ao longo do tempo. A pesquisa coletou mais de 600 mil tweets entre 2009 e 2015 e aplicou técnicas de mineração de texto para identificar três dimensões do discurso online — cognitiva, interativa e social — utilizando modelos de classificação automática. A precisão mais alta foi obtida com regressão logística usando expressões regulares como recurso (precisão: ~69,8 %; recall: ~60,3 %). A partir dos dados classificados, foi realizada uma análise de sobrevivência (survival analysis), determinando como cada dimensão de discurso afeta o tempo de permanência de um usuário na comunidade. A exposição a tuítes nas dimensões cognitiva e interativa reduziu significativamente o risco de abandono — usuários

expostos a uma unidade de desvio padrão a mais dessas dimensões ficaram cerca de 46 % mais propensos a permanecer. Por outro lado, exposição a tuítes na dimensão social aumentou em aproximadamente 13,6 % a chance de abandono.

Podemos afirmar, portanto, que a lacuna crítica na literatura consiste na ausência de estudos que caracterizem sistematicamente as dimensões lexicogramaticais e discursivas dos bate-papos educacionais no Twitter.

A análise multidimensional, embora bem estabelecida para outros gêneros (Berber Sardinha & Veirano Pinto, 2019), não hava sido adequadamente aplicada a corpora de bate-papos educacionais pelo Twitter. Esta pesquisa visa a contribuir para a área de pesquisa ao utilizar o modelo de Berber Sardinha e Fitzsimmons (2025) para revelar dimensões discursivas dos bate-papos educacionais no Twitter. Espera-se obter insights sobre os posicionamentos ideológicos que permeiam as conversas entre os participantes, caracterizando os bate-papos educacionais como gênero discursivo digital.

A literatura atual, embora valiosa em seu foco em redes, temas e percepções individuais, não oferece uma caracterização sistemática das dimensões discursivas e variações linguísticas em bate-papos educacionais no Twitter. Meu estudo visa a preencher esta lacuna crítica ao aplicar metodologia baseada em corpus – especificamente a análise multidimensional lexical – para descrever o que sustenta os temas destes bate-papos.

Esta abordagem não apenas documenta um fenômeno comunicativo contemporâneo, mas também contribui metodologicamente para a intersecção entre Linguística de Corpus, análise do discurso digital e educação mediada por tecnologia, oferecendo insights que as abordagens dominantes não conseguem capturar.

1.3. POR QUE ESTUDAR TWEETS EDUCACIONAIS?

Apesar do declínio da participação de professores nos bate-papos educacionais do Twitter, esse objeto continua relevante do ponto de vista analítico-discursivo. Os bate-papos educacionais, estruturados por hashtags e realizados em tempo real, constituíram práticas significativas de construção de redes pessoais de

aprendizagem, circulação de saberes e negociação de sentidos sobre educação (Britt & Paulus, 2016).

Embora muitos bate-papos tenham desaparecido do Twitter, o formato permanece vivo em ambientes como Bluesky e Mastodon, onde educadores retomam a lógica de perguntas e respostas mediadas por hashtag. A migração revela a resiliência do gênero: mais do que uma prática ligada a uma plataforma específica, trata-se de um modo de interação que promove aprendizagem profissional e construção coletiva de conhecimento. Com a proposta metodológica desta pesquisa, espera-se através de uma análise multidimensional lexical da linguagem dos bate-papos educacionais que ocrreram entre 2009 e 2022, encontrar discursos que ratifiquem as características identificadas por Greenhalgh e Rosenberg e assim atestar a viabilidade de ambientes virtuais de aprendizagem para professores conforme os moldes dos bate-papos educacionais no Twitter.

Assim, o estudo desses bate-papos é pertinente não por sua centralidade atual no Twitter, mas por sua contribuição histórica e discursiva para compreender como educadores se organizam, constroem comunidades e adaptam gêneros digitais em contextos mutáveis (Williams, 2025). Williams destaca o crescente vazio deixado pela retração do uso acadêmico/pedagógico do Twitter ao avaliar os impactos dessa fragmentação sobre a comunidade acadêmica, sobretudo para educadores em início de carreira. Ressalta o papel quase que insubstituível e irreplicável que a rede social teve em estabelecer e mobilizar comunidades de pesquisa, em ofereccer um canal crucial para a internacionalização da academia e a visibilidade global de pesquisadores e ideias e por fim promover encontros espontâneos e enriquececom com educadores de outras áreas. O autor reconhece a função do Twitter como ambiente de desenvolvimento profissional onde os bate-papos educacionais figuram como espaços regulares que promovem conexões, apoio emocional, troca de práticas e interações autênticas entre educadores. Esta tese não focará, porém, no papel do Twitter como rede social nem nos bate-papos educacionais em si como espaços de construção colaborativa e troca, mas sim nas motivações pessoais, políticas, filosóficas e sociais das escolhas linguísticas dos usuários ao interagirem com outros participantes desses bate-papos educacionais por meio de um arcabouço teórico-

metodológico formado pela Linguística de Corpus e a Análise Multidimensional Lexical.

1.4. BATE-PAPO EDUCACIONAL NO TWITTER

Este estudo objetiva investigar a variação discursiva da linguagem gerada nos bate-papos educacionais pelo Twitter. A arquitetura desta rede social privilegia a partilha, interação e discussão de ideias e temas de interesse comum. De modo geral, Twitter é visto como plataforma de intenso ativismo social e militância política, além da sua funcionalidade inicial, que é de relatos pessoais em tempo real sobre assuntos corriqueiros (o usuário é incentivado a postar respondendo à pergunta: "o que está fazendo?"). A despeito de suas características favoráveis ao diálogo e à troca de informações, usar o microblog para debater assuntos ligados à educação em tempo real com usuários separados geograficamente pareceu improvável desde o primeiro bate-papo educacional no Twitter em 2009. Afinal, quando o primeiro tuíte foi enviado, em 2006, era inconcebível a importância e o alcance que essa plataforma teria sobre o mundo, assim como era inimaginável as muitas formas com que o Twitter poderia ser usado, inclusive no espaço educacional.

Lançado em março de 2006 como forma de socialização, o Twitter permite que uma pessoa envie uma mensagem instantânea, chamada tuíte, com no máximo 280 caracteres, para seus seguidores, ou seja, as pessoas que escolheram ficar em contato. Essas mensagens podem ser enviadas por meio de um computador ou smartphone. O Twitter também é considerado como um microblog, por possuir características de um Weblog, pois nele é possível publicar postagens diárias, que ficam armazenadas em ordem cronológica. Trata-se de uma realidade hiperflexível por poder assumir vários significados: desde mensagens instantâneas até um verdadeiro instrumento de rede social como forma peculiar de blog coletivo, que permite criar, trocar e integrar ideias, notícias e conceitos; em resumo, um verdadeiro e próprio laboratório de micro comunicação em ebulição.

Os bate-papos educacionais[1], são conversas online entre educadores que acontecem no Twitter, usando uma hashtag específica, para discutir temas de interesse na educação. Eles funcionam da seguinte forma: um moderador ou organizador propõe um tema e algumas perguntas para guiar a discussão, e os participantes respondem usando a hashtag do chat. Participam professores, gestores, estudantes, pesquisadores e outros profissionais ligados à educação. O propósito deles é criar uma comunidade de aprendizagem, trocar experiências, compartilhar recursos e ampliar a rede de contatos. A relevância deles para a educação é que eles permitem que os educadores se atualizem sobre as tendências, as práticas e as inovações na área, além de se inspirarem e se apoiarem mutuamente em um espaço de expressão e sociabilização, possibilitadas pela formação, muitas vezes, espontânea de redes sociais.

Desde seu surgimento, no ano de 1996, as redes sociais vêm transformando como as pessoas se relacionam umas com as outras, tornando-se a principal fonte de informação e comunicação para metade dos quase 8 bilhões de habitantes do nosso planeta. A base de usuários das plataformas digitais na última década disparou de 970 milhões em 2010 para uma notável marca que ultrapassou os 4,95 bilhões de usuários em outubro de 2023. O aumento exponencial do uso e da influência das mídias sociais aponta para a necessidade de compreender como as teorias linguísticas interpretam esses ambientes dinâmicos que seguem em constante evolução.

Nota-se que, os modelos de aprendizagem profissional que antecederam as plataformas da web e outros artefatos digitais, em particular as redes sociais se mostram incapazes de ofertar um aprendizado participativo, colaborativo e autodeterminado. Uma das primeiras opções foi a dos fóruns de discussão virtual, ferramenta pedagógica que fomenta o compartilhamento e a reflexão de crenças e práticas do professor de línguas.

A análise das representações veiculadas por discursos educacionais em redes sociais digitais, particularmente em chats educacionais no X, permite observar a

[1] Em inglês, educational Twitter chats

complexidade ideológica em torno da figura docente. Tais representações são construídas e veiculadas por meio de recursos linguísticos recorrentes que podem ser identificados, descritos e interpretados com o apoio da Linguística de Corpus, a qual possibilita identificar padrões lexicais e gramaticais que, por sua frequência e distribuição, revelam as regularidades ideológicas dos discursos sociais. Foram selecionados alguns estudos que se propuseram a investigar as crenças e proposições que sustentam as crenças dos envolvidos direta e indiretamente com a educação.

1.5. REDES SOCIAIS NA FORMAÇÃO DO PROFESSOR

A formação continuada, em especial aquela que leva ao desenvolvimento profissional autodirecionado tem seus pilares na utilização intencional das redes sociais. As redes sociais favorecem o encontro com outras culturas e vivências de sala de aula e figuram como espaço onde os usuários suscitam dúvidas, possibilidades, constroem significados, compartilham seus questionamentos, resultando na criação de novos saberes advindos da reciprocidade coletiva. Os avanços nos campos de Tecnologias de Informação e Comunicação, e mais recentemente, a Inteligência Artificial Gerativa fazem surgir novos formatos de prática pedagógica autônoma, bem como propostas de formação continuada que promovem novos aprendizados, trocas de experiências e oportunidades de construção de conhecimento colaborativo e intercultural. Sob o prisma da relação rede social - formação continuada, fazer parte de uma comunidade de prática ubicada em uma plataforma digital como o Facebook, o Instagram e o Twitter permite ao professor entender e expressar opiniões sobre crenças e práticas culturais específicas.

As potencialidades e oportunidades das redes sociais como espaço de aprendizagem colaborativa, significativa, intercultural, personalizada e auto-determinada se materializaram no surgimento de trocas de postagens no Twitter em tempo real e organizada por uma hashtag, os bate-papos educacionais.

A partir dos pilares de colaboração e interculturalidade sob os quais as redes sociais se sustentam, a aprendizagem profissional encontrou no bate-papo

educacional[2] no Twitter um dos seus métodos mais populares. Aydin (2014) avaliou o microblogue como ambiente educacional e seu impacto positivo sobre a educação. Assim como outros estudos, os bate-papos educacionais proporcionam experiências proveitosas em termos de conteúdos, colaboração, protagonismo do professor, coaching entre pares e diálogos intermitentes.

Estes encontros em tempo real se destacam por oportunizar aprendizado contínuo sobre tópicos escolhidos pelos participantes em uma comunidade de educadores separados geograficamente. Acabam por disponibilizar ideias, sugestões e estratégias “prêt-à-porter”, prontas para serem implementadas em sala de aula, embora haja poucas oportunidades de desenvolvimento contínuo específico ou momentos limitados de diálogo com outros profissionais sobre problemas na prática.

Neste momento, cabem algumas considerações sobre a função dos bate papos educacionais no Twitter na aprendizagem colaborativa docente. A partir de 2009, usuários envolvidos direta ou indiretamente com a educação e o ensino resignificaram o microblogue Twitter, ao aproveitar as funcionalidades da rede social para estabelecer comunicação síncrona (e assíncrona), organizada em torno de hashtags para discutir assuntos relacionados à educação e ao ensino. Estes encontros, ou bate-papos educacionais transformaram em alternativas ou complementações a programas de formação continuada para aprimorar o aprendizado, construir redes profissionais; bem criar um meio para atingir metas de desenvolvimento profissional autodeterminadas.

A atualização de conhecimentos e a consolidação de aprendizados dos usuários participantes dos bate-papos educacionais pelo Twitter podem estar relacionados à participação consistente dos professores em vários chats educacionais do Twitter. Os chats demonstraram aprimorar as experiências de curso dos estudantes universitários e

Útil tanto para professores em exercício quanto para desenvolvimento profissional, bem como para professores em serviço em seus programas de

preparação, a sincronicidade dos chats do Twitter permite promover conexões sociais, diminuir o isolamento e construir comunidades de educadores (Carpenter & Krutka, 2014). Essas vias digitais de conexão podem tornar possível para os professores descobrirem que compartilham interesses comuns e podem oferecer oportunidades instantâneas e personalizadas para crescimento profissional.

Redes de Aprendizagem Profissional (em inglês, *Professional Learning Networks*), são entendidas aqui como vivências por meio de plataformas digitais. Ao pensar sobre quais fatores promovem o aprendizado profissional nesses contextos, podemos sugerir que as oportunidades online para desenvolvimento profissional devem incluir três componentes principais: conteúdo sobre assunto específico, a oportunidade de refletir sobre o aprendizado e uma oportunidade de colaborar com outros. Enquanto acreditamos que o uso da tecnologia não substitui a necessidade de uma pedagogia crítica, reconhecemos que os ganhos ao integrar recursos tecnológicos para fomentar um ensino baseado em aprendizagem ativa e cooperação.

Os resultados que serão apresentados e analisados mais adiante indicam que os chats do Twitter fornecem múltiplos componentes de aprendizagem profissional de alta qualidade, a saber um foco no conteúdo, colaboração e protagonismo do professor; em menor grau, eles podem fornecer treinamento de pares e permitir conversas por uma duração sustentada. No entanto, outros componentes de aprendizagem profissional significativa não são possíveis neste contexto, pois não é incorporada ao trabalho e não fornece aprendizagem ativa ou oportunidades apoiadas para a prática.

Como um modelo profissional alternativo, muitos educadores se voltaram para plataformas de mídia social como o Twitter como um espaço livre, informal e comunitário para aprendizagem profissional no qual os participantes podem criar redes personalizadas. Um método de aprendizagem profissional no Twitter é o uso de chats do Twitter, diálogos virtuais planejados que são organizados em torno de tópicos e incluem o uso de hashtags como seu método de organização.

Dado isso, os bate-papos educacionais realizados pelo Twitter apresentam pontos ao seu favor enquanto meio de aprendizagem informal e profissional: espaços de interação e troca de experiências que promovem a co-construção de

conhecimentos teóricos e práticos no que dizem respeito ao processo de ensino e aprendizagem. Em contrapartida, o efeito de universo bolha ou câmara de eco (em inglês, echo chamber) - a interação contínua com as mesmas pessoas em uma rede social tende a enviesar o olhar do usuário. Embora não haja estudos específicos sobre câmaras de eco em bate-papos educacionais no Twitter - a literatura disponível concentra-se principalmente em polarização política, desinformação e aspectos psicológicos das câmaras de eco - a polarização e o isolamento informacional podem afetar o desenvolvimento profissional e a troca de conhecimento entre educadores.

Com o advento das redes sociais profissionais online, plataformas como o Twitter oferecem possibilidades de aprendizagem profissional em tópicos selecionados pelos próprios usuários que participam dos bate-papos, formando uma comunidade de prática cujos membros, embora estejam separados geograficamente, criam uma relação dialógica que traz benefícios a todos: a sensação de pertencimento, de apoio mútuo e de uma fonte de ideias e trocas é inegável. As possibilidades se manisfestam de diversas formas, em especial:

- Coaching de pares: coaching de pares e feedback são características críticas de sistemas de aprendizagem profissional

- Aprendizagem ativa: essa reflexão sobre a ação pode então transformar uma experiência informativa, na qual os participantes ganham novos conhecimentos e habilidades, em uma experiência transformadora, na qual os professores mudam seus pontos de vista ou hábitos mentais.

No espaço coletivo de um bate-papo no Twitter, os professores podem compartilhar e obter o o que pode ser chamado de informações "just-in-time" que podem integrar à pedagogia. O bate-papo no Twitter serve como um terceiro espaço no qual os profissionais podem dialogar entre si e com organizações que atuam como corretoras de conhecimento em seus campos em relação a problemas de prática. Embora limitado, esse tipo de engajamento começa a atender à necessidade de aprendizagem ativa e colaboração do professor na aprendizagem profissional.

1.6. LINGUÍSTICA DE CORPUS

A análise dos dados obtidos depende diretamene do entrelaçamento do entendimento de discurso, conforme proposto por Berber Sardinha e Fitzsimmons Doolan (2025). Ao contrário das abordagens linguísticas tradicionais, que muitas vezes se concentram na análise de textos individuais ou em pequenos conjuntos de dados, a LC se ocupa de grandes volumes de texto. A Análise do Discurso assistida por corpus (Gillings et al., 2023) trabalha no desenvolvimento de métodos para identificar discursos subjacentes em um conjunto volumoso de textos. Para tanto, emprega um arcabouço interdisciplinar, integrando contribuições da Linguística de Corpus, Análise Multidimensional Lexical e Análise do Discurso Assistida por Corpus.

No que se segue, exploram-se as interfaces teóricas entre a LC, AMDL e a Análise do Discurso Assistida por Corpus, com ênfase em como suas perspectivas transdisciplinares podem ser mobilizadas para aplicações contemporâneas em análise de registros digitais. Tal abordagem evidencia o potencial da LC para ultrapassar fronteiras disciplinares, oferecendo subsídios metodológicos e analíticos que permitem uma investigação mais robusta dos fenômenos discursivos em redes sociais.

A Linguística de Corpus norteia os princípios de criação de corpora representativos da linguagem utilizada em determinado contexto comunicativo, bem como as bases de criação de sentido na língua em uso por meio da coocorrência sistemática de itens léxico gramaticais. A Análise Multidimensional Lexical, vertente discursiva da Análise Multidimensional, foi utilizada neste trabalho visando identificar os discursos que permeiam os bate-papos educacionais pelo Twitter. Por fim, foi usada a Análise do Discurso Assistida por Corpus, uma área multidisciplinar voltada à identificação de discursos por meio da análise de corpora (Friginal & Hardy, 2020).

Se analisados os bate-papos educacionais como sendo um ecossistema discursivo, eles devem ser espaços em que se consolidam visões de mundo, ideologias e representações que indexam valores, intenções, conceitualizações e modos de agir. O ecossistema discursivo pode ser definido como um ambiente aberto comunicativo e educacional, dinâmico e interdependente, no qual sujeitos, mídias, tecnologias, práticas pedagógicas e contextos culturais interagem de forma dialógica

e participativa. Ele emerge da articulação entre ecossistemas comunicativos e ecossistemas de aprendizagem, configurando-se como espaço de produção coletiva de sentidos, coautoria e construção democrática do conhecimento em rede, permitindo interatividade, conectividade e aprendizagem colaborativa.

A Linguística de Corpus é uma área dos estudos linguísticos que emprega corpora (plural de corpus), isto é, coleções de textos de diversos modos semióticos (falado, escrito, visual, sonoro, etc) armazenados em formato de computador, com a finalidade de descrever situações de uso das linguagens e entender como se produz a relação entre o uso sistemático desses recursos de expressão (Moreira, 2023, p. 3)

A Linguística de Corpus, portanto, se dedica ao estudo da linguagem natural por meio de corpora, isto é, conjuntos de textos escritos ou falados que são armazenados em bancos de dados eletrônicos. A Tecnologia Educacional é um campo que se dedica ao uso de tecnologias para melhorar o processo de ensino e aprendizagem. A interface entre essas duas áreas pode ser vista como uma oportunidade para o desenvolvimento de novas metodologias de ensino e aprendizagem que utilizem estes conjuntos de textos como ferramenta pedagógica.

Berber Sardinha (2000) explica que a Linguística de Corpus ocupa-se da coleta e exploração de corpora, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística e como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador.

A Linguística de Corpus vem mudando a maneira como se investiga a linguagem, nos seus mais diversos níveis, colocando à disposição do analista quantidades de dados antes inacessíveis. Um dos grandes agentes dessa mudança é a informática; sem a qual a Linguística de Corpus contemporânea não poderia existir. Assim, o linguista de corpus depende de programas de computador para lidar com corpora.

A LC se tornou a área dos estudos linguísticos e da Linguística Aplicada que se concentra na coleta e análise de grandes conjuntos de dados textuais, formatados para serem legíveis por computador, com o objetivo de pesquisar uma língua ou variedade linguística. Essa abordagem é fundamental para entender como a

linguagem é usada no dia a dia das pessoas, permitindo a formulação de teorias sobre o funcionamento da língua em uso. Analisar os impactos do uso dos bate-papos educacionais no Twitter na formação de professores de inglês por meio de metodologias da Linguística de Corpus como a coocorrência lexical revela os discursos e posicionamentos dos seus participantes nos exemplos encontrados no corpus.

A LC conta com recursos computacionais avançados para analisar grandes volumes de dados linguísticos, um processo que seria inviável com os equipamentos disponíveis nos anos 1950. Esta abordagem proporciona insights sobre o manejo e o uso da linguagem na sociedade. Ao analisar os tuítes de professores de inglês, por exemplo, pode-se obter uma compreensão mais profunda das tendências pedagógicas, desafios e soluções no campo do ensino de inglês como língua estrangeira.

Conforme explicitam Biber, Conrad e Reppen (1998), a abordagem baseada em corpus permite buscar respostas a respeito da linguagem utilizada em determinado contexto comunicativo com o uso do computador, o qual é naturalmente programado para detectar ocorrências e coocorrências. Ainda, esta abordagem centrada em corpus tem como arcabouço uma base empirista, ao utilizar técnicas quantitativas e qualitativas de interpretação dos dados. Contam ainda a própria definição de corpus: uma coletânea grande e criteriosa (`principled') de textos de linguagem natural (isto é, não artificiais como linguagem de programação de computador ou matemática). No contexto do Twitter, isso envolve analisar como os professores de inglês compartilham conhecimentos, discutem metodologias de ensino e interagem com outros profissionais da área.

1.7. OBJETIVOS DE PESQUISA

Esperamos explorar como o léxico contribui para a formação de discursos nas conversas entre os participantes dos bate-papos educacionais pelo Twitter. Assim sendo, o presente estudo tem como objetivo geral descrever os discursos subjacentes nos bate-papos educacionais online pelo Twitter, e como seu objetivo específico, identificar as dimensões subjacentes por meio da variação linguística dos bate-papos educacionais pelo Twitter.

1.8. PERGUNTAS DE PESQUISA

Feitas as considerações acima, convém explicitar as perguntas que motivaram o tema desta pesquisa:

1. Quais são as dimensões lexicais do Twitterchat corpus?

2. Quais discursos essas dimensões sinalizam?

Considerando que estas conversas em tempo real que ocorrem pelo Twitter se organizam em torno de uma hashtag para todos acompanharem e participarem da discussão, podemos identificar em termos gerais os assuntos abordados. Segundo Berber Sardinha (Berber Sardinha, 2022a, p. 657):

A pesquisa baseada em corpus sobre mídias sociais pode buscar respostas para perguntas importantes, como que tipo de linguagem é usada nessas comunidades de mídias sociais, como o uso da linguagem varia entre grupos e indivíduos, que variedades de texto existem e como elas se comparam às variedades não digitais e como as formas não padronizadas, como contrações, hashtags, emoticons e emojis, são difundidas[3].

Dizer que usuários participantes de encontros síncronos ou bate-papos educacionais pelo Twitter se reúnem virtualmente para discutir assuntos relacionados à educação, ao ensino, a teorias de ensino e aprendizagem nos priva de um panorama de conceitos, ideologias, valores e crenças que se encontram subentendidos nas postagens dos integrantes dessas comunidades de prática (estas são definidas como grupos cujos membros “se engajam frequentemente para o compartilhamento e a aprendizagem, baseados em seus interesses comuns” (Oliveira & Carvalho, 2023).

3 No original: Corpus-based research on social media can seek answers to key questions such as what kind of language is used in these social media communities, how language use varies across groups and individuals, what text varieties exist and how they compare with non-digital varieties and how widespread non-standard forms such as contractions, hashtags, emoticons and emojis are.

Com esta introdução, podemos posicionar a problemática dos bate-papos educacionais nas redes sociais e a perspectiva metodológica da Linguística de Corpus para a consecução da pesquisa e análise dos dados. O trabalho a seguir foi organizado da seguinte forma: no capítulo 2, expomos as principais ideias e teorias que apóiam a utilização do corpus, seu design de construção e a análise estatística quantitativa de dados linguísticos; o capítulo 3 destaca as principais técnicas empregadas na Análise Multidimensional Lexical, assim como detalhes da coleta do corpus e da definição das variáveis lexicais. Os capítulos 4 e 5 farão uma apresentação e subsequente discussão dos resultados gerados, a fim de responder às perguntas da pesquisa. No último capítulo faremos ponderações quanto à importância desses resultados para progamas e materiais de formação continuada de profissionais da educação.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta as áreas que forneceram embasamento teórico para o trabalho, orientado pela seguinte organização: primeiramente, são apresentados conceitos principais e trabalhos relevantes da LC. A seguir, são apresentados conceitos definidores da metodologia e estudos que também utilizaram a AMDL e que, portanto, têm relação direta com esta pesquisa.

Para que seja possível compreender a AMDL, deve-se, primeiramente, abordar a Linguística de Corpus. Conforme dito na Introdução, o trabalho aqui proposto tem como fundamentação teórica principal a Linguística de Corpus (doravante, LC), que pode ser definida como “uma área que trata do uso de corpora computadorizados (coletâneas de textos, escritos ou de transcrições de fala, mantidas em arquivo de computador)” (Berber Sardinha, 2004, p. xvii)

2.1. LINGUÍSTICA DE CORPUS

A pesquisa recorre ao arcabouço teórico da Linguística de Corpus (Berber Sardinha, 2004) por meio da abordagem metodológica da análise multidimensional (Berber Sardinha & Veirano Pinto, 2014, 2019). Mais especificamente, a análise multidimensional lexical (ou AMDL), que é uma metodologia da Linguística de Corpus aplicada ao estudo de variação lexical entre variedades textuais situacionalmente definidas, ou registros (Berber Sardinha & Fitzsimmons-Doolan, 2025).

As principais ocupações da LC são: (1) a compilação de corpora (construção de corpus especializado de uma língua ou variedade, tal como a linguagem jurídica, jornalística, médica etc.); (2) o desenvolvimento de ferramentas para análise de corpora; (3) a descrição de linguagem e (4) a exploração do uso de descrições baseadas em corpora para várias aplicações, tais como ensino-aprendizagem de línguas, processamento de linguagem natural por máquinas, reconhecimento de voz e tradução (Berber Sardinha, 2004)..

A Linguística de Corpus tem como preceito a linguagem enquanto sistema probabilístico. Segundo essa noção, os elementos gramaticais e lexicais não ocorrem de forma aleatória. Consequentemente, a Linguística de Corpus se encaixa no que pode ser chamado de Linguística Empírica. Na linguística, “empírico significa primazia

dos dados provenientes da observação da linguagem, em geral reunidos sob a forma de um corpus" (Berber Sardinha, 2004, p. 30).

A LC tem como principal objetivo, portanto, estudar a língua a partir da observação de grandes quantidades de textos autênticos, em língua natural, coletados criteriosamente e armazenados em formato eletrônico para fins de pesquisa linguística. Uma definição completa que envolve todas as características de um corpus é citada por Berber Sardinha (2004, p. 18):

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

A partir da definição acima, podemos salientar três atributos básicos da constituição de um corpus: autenticidade, legibilidade e representatividade dos dados. Estes devem ser legíveis por computador e representativos de uma linguagem ou variedade para estudo linguístico, de modo a embasar a pesquisa com sólidos fundamentos metodológicos.

2.2. BREVE HISTÓRICO DA LINGUÍSTICA DE CORPUS

Como visto anteriormente, a LC pode ser definida como a área da Linguística responsável pela coleta e análise de corpora. Essa definição, no entanto, revela pouco sobre a sua origem, que é mais antiga – proveniente, historicamente, do século XIII, período em que monges trabalhavam como copistas, extraindo concordâncias[4]. Antes da informatização, a coleta era, basicamente, feita em fichas de papel, que eram armazenadas em escaninhos. Há também documentos de

4 Definidas como "listagens das ocorrências de um item específico acompanhado do texto ao seu redor (o cotexto)." (BERBER SARDINHA, 2004, p.187) de textos bíblicos com o intuito de comprovar sua unidade harmônica por meio de padrões linguísticos (McCARTHY; O'KEEFFE, 2010).

lexicógrafos, datados do final do século XIX, que coletavam exemplos da língua em uso para ajudar na definição das palavras de forma mais precisa, assim como indícios da existência de coleções de textos sagrados que visavam a criação de índices por mais de cinco séculos, o que pode ser considerado como um trabalho precursor da LC.

Esse tipo de coleta de dados passou a fazer parte, mais adiante, também, de outras áreas do saber, à medida que teses como a de Halliday (1991) puderam contar com o avanço tecnológico de processamento de dados por computador. Foi este avanço, e não os estudos linguísticos dos anos de 1960, que impulsionaram o desenvolvimento das pesquisas baseadas em corpus. Assim, foi a partir do uso do computador que a coleta de dados evoluiu para o que chamamos, atualmente, de corpora.

Uma grande influência na pesquisa de Halliday é proveniente das ideias do linguista britânico John Rupert Firth, que atuou principalmente na década de 1950. Para Firth, a língua passa a ser vista enquanto parte integrante de uma situação comunicativa. Abandona-se a compreensão de língua como um sistema mental, não mais devendo essa ser estudada como entidade autônoma, mas, sim, percebida a partir de um conjunto de eventos linguísticos expressos pelo falante. A língua passa, portanto, a ser observada de forma significativa, a partir do seu contexto de produção. Para tanto, Firth relaciona o contexto de produção às categorias gramaticais, características dos participantes, objetos e o efeito da ação verbal, em um contexto que está diretamente atrelado ao âmbito da cultura. Segundo Firth (1957) , a significação parte tanto do sentido situacional como do sentido lexical. Ou seja, o sentido surge tanto do contexto em que são ditas, como das palavras que são ditas.

Na perspectiva firthiana, posto que a noção de significado surge da comunhão de vários níveis linguísticos, como o discursivo, o textual, o gramático, o lexical e o fonético, Firth (1957) expõe elementos que geram os seguintes princípios: (1) A Linguística é essencialmente uma ciência social e aplicada; (2) A língua deveria ser estudada por meio de exemplos de uso reais, atestados e autênticos, não por meio de sentenças intuitivas, inventadas, isoladas; (3) A unidade de estudo deve ser os textos inteiros; (4) Textos e tipos de textos devem ser estudados comparativamente entre os textos dos corpora; (5) A Linguística preocupa-se com o estudo do sentido:

forma e sentido são inseparáveis; (6) Não há limites entre léxico e sintaxe; léxico e sintaxe são interdependentes; (7) Muito do uso da língua é rotineiro; (8) A língua em uso transmite cultura; e (9) Dualismos saussurianos são concepções errôneas (Stubbs, Baker, & Tognini-Bonelli, 1993, p. 2)[5].

Esses princípios, que orientam a tradição neo firthiana da linguagem, lançam as bases para o que viria a ser a abordagem, hoje, conhecida, da Linguística de Corpus, assim como embasar pesquisas desenvolvidas por linguistas de diversas linhas, como M. A. K. Halliday e John Sinclair.

Halliday (1991) questionou, de modo direto, a posição de Chomsky, vigente na época, em relação à probabilidade relativa. Enquanto Chomsky, maior expoente do racionalismo na linguística, defendia a linguística gerativista, a qual enfatiza a determinação de quais agrupamentos sintáticos são possíveis dado o conhecimento que um falante nativo possui de sua língua, Halliday, seguindo a tradição empirista, descreve a probabilidade dos sistemas linguísticos, dados os contextos em que os falantes os empregam. Convicto da importância de um corpus para os estudos linguísticos, Halliday defende a característica probabilística da linguagem, classificando-a como algo inerente à língua, enquanto sistema, e afirmando que a frequência de um dado linguístico não pode ser interpretada meramente como acidente ou efeito circunstancial, tal como na teoria gerativa. Os estudos baseados em corpus, de fato, devem revelar esse caráter probabilístico da língua. Neles, a probabilidade relativa é tão somente uma questão de interpretação dos dados, e não

---

5 No original:

(1) Linguistics is essentially a social science and an applied science.

(2) Language should be studied in actual, attested, authentic instances of use, not as intuitive, invented, isolated sentences.

(3) The unit of study must be whole texts.

(4) Texts and text types must be studied comparatively across text corpora.

(5) Linguistics is concerned with the study of meaning: form and meaning are inseparable.

(6) There is no boundary between lexis and syntax; lexis and syntax are interdependent.

(7) Much language use is routine.

(8) Language in use transmits the culture.

(9) Saussurian dualisms are misconceived.

um construto sobre o funcionamento linguístico. De acordo com o raciocínio de Halliday, essa discordância no foco de discussão aparece de modo muito claro, quando o autor ressalta:

O fato de o positivo ser mais frequente do que o negativo seria uma propriedade essencial do sistema – tão essencial quanto os termos da oposição propriamente. Analiticamente, foi necessário separar as considerações sobre os termos do sistema das considerações sobre as suas probabilidades relativas; mas o que esteve envolvido foi um único fenômeno, não dois (Halliday, 1991, p. 31)[6].

O ponto fundamental na argumentação de Halliday não é o fato de determinada ocorrência ser ou não provável, mas o fato de sabermos com qual frequência ela acontece. Segundo Berber Sardinha (2004, p. 31), "a visão da linguagem como sistema probabilístico pressupõe que, embora muitos traços linguísticos sejam possíveis teoricamente, não ocorrem com a mesma frequência".

Entende-se, portanto, que a frequência observada dos dados não acontece de forma aleatória, ou seja, há repetição tanto da ocorrência de um dado linguístico quanto da situação em que é produzido. Essa noção se tornaria central à LC, sobretudo porque, a partir disso, tornou-se inevitável demonstrar que a língua é, em certa medida, "padronizada". Destacando-se, em LC, os estudos relacionados ao léxico, a questão da padronização aparece com força na detecção de elementos, que são de tal maneira regulares e sistemáticos, que se podem chamá-los de padrões "lexicogramaticais" (Halliday, 1991), em contraposição à divisão de gramática versus léxico da teoria gerativa.

A LC, desse modo, instaura-se com a clara convicção de que as línguas são sistemas probabilísticos. Desde então, os avanços das pesquisas envolvidas com corpus parecem indicar que, mais que um simples registro da performance do falante,

---

6 No original: It seemed to me self-evident that, given a system 'polarity' whose terms were 'positive/negative', the fact that positive was more frequent than negative was an essential property of the system – as essential as the terms of the opposition itself. Analytically, of course, it was necessary to separate the statement of the terms of the system from the statement of their relative probabilities; but what was involved was a single phenomenon, not two.

o corpus é um lugar de experimentação dos mais ricos e surpreendentes (Fillmore, 1992), tanto no que se refere aos resultados com ele obtidos quanto no que diz respeito ao posicionamento teórico que o investigador é levado a assumir ao construir ou utilizar um corpus.

Hoje cresce o interesse de grande número de linguistas em basear suas teses sobre a linguagem a partir de dados reais; ou seja, língua em uso. Estes podem vir a se manifestar nos textos escritos e/ou falados, de modo autêntico, em qualquer língua natural, o que recebe de corpus (McEnery & Wilson, 1996). O corpus apresenta-se na linguística da atualidade não só como meio de extração de informações sobre língua(s), mas também como objeto de pesquisa propriamente dito, devendo o linguista (de corpus) preocupar-se em conferir-lhe a confiabilidade científica até então questionada.

Em certo sentido, podemos dizer que um novo paradigma se coloca frente à visão chomskiana sobre a linguagem, indicando que mais do que descrever o que é possível na língua, seria investigar o que nela é provável (Halliday, 1991); que antes considerar a recorrência (padronização) de dados linguísticos do que apostar na ocasionalidade desses elementos no uso da língua (Biber et al., 1998; Sinclair & Jones, 1974/1996).

Faz-se necessário registrar que, ainda que esse referencial teórico incida de modo positivo sobre a consolidação da LC, ainda assim há fatores externos (ou contextuais) sem os quais a abordagem da LC não teria êxito. Esse componente contextual, que se mostra presente e constante na literatura, aparece como sendo de inteira relevância para o percurso da LC.

Sinclair, com base nas ideias de Firth, chega a sua concepção da língua e da linguagem a partir de dois princípios básicos: o princípio idiomático, no qual as palavras tendem a ocorrer juntas "e gerarem significados por meio dessa combinação" (Sinclair, 2004, p. 29)[7], o que faz com que elas sejam arquivadas mentalmente como unidades de sentido, equivalente a uma escolha única antes de serem proferidas; e o

7 Words tend to go together and make meanings by their combinations.

princípio da livre escolha, segundo o qual, nos casos raros em que as palavras têm um significado fixo em relação ao mundo, o falante as seleciona de modo individual e as combina em função de seu sistema de regras gramaticais mentais. Apesar de Sinclair pontuar, com frequência, em sua obra, que a maioria dos itens lexicais (vocabulário) segue o princípio idiomático, ele não se distancia por completo da distinção entre gramática e léxico, pois seu princípio da livre escolha admite a existência de uma gramática mental relacionada ao léxico. As noções de padrão de linguagem, lexicogramática e frequência de Sinclair, originadas da proposta de Firth, inspiram estudos na área de Lexicografia, no desenvolvimento de dicionários, na análise de discurso, na tradução e no ensino de línguas (Stubbs, 1995).

A Linguística de Corpus surge como resposta à necessidade de se trabalhar com uma linguagem empírica, a qual privilegia o estudo da língua em uso. A LC apoia-se na linguagem autêntica para, a partir dela, elaborar generalizações e delinear teorias a respeito do funcionamento linguístico como um todo. A LC tem suas bases na Linguística Aplicada e, enquanto área que se desdobra da LA, caracterizando-se, em linhas gerais, por se ocupar "da coleta e exploração de corpora, ou conjuntos de dados linguísticos textuais, em formato legível por computador, que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística" (Berber Sardinha, 2004, p. 3). Em outras palavras, não se trata de uma compilação qualquer de textos, mas da realização de uma coleta orientada por critérios previamente estabelecidos para que um corpus (no plural, corpora) seja considerado relevante e representativo da língua.

Assim, toda pesquisa em LC deve contemplar, necessariamente, um conjunto de textos coletados de maneira sistemática e criteriosa, ou simplesmente, um corpus. Hunston (Hunston, 2002, p. 2) aponta que, "um corpus é definido em termos de sua forma e seu propósito"[8]. Atualmente, essa forma está intrinsecamente ligada ao uso do computador, uma vez que é por meio da tecnologia que viabiliza-se o tratamento dos dados, sua coleta, armazenamento e processamento. O propósito de um corpus varia de acordo com a pesquisa que se pretende. Há relatos de corpora sendo usados

8 No original: A corpus is defined in terms of both its form and its purpose.

em diferentes áreas de atuação, como, por exemplo, no ensino, oferecendo oportunidades de observação da língua em uso; na tradução, com a possibilidade de se trabalhar com corpora paralelos; em lexicografia e na terminologia, na construção de glossários e dicionários; bem como na pesquisa, para buscar padrões linguísticos e recorrências que permitam classificar e descrever estilos e registros.

Para Hunston (2002, p. 23), as evidências de uso da língua devem, necessariamente, partir da análise de um corpus que permita descrever padrões recorrentes de uma determinada língua, e como ela é usada de forma contextualizada. O corpus permite identificar a frequência com que um determinado item lexical ou padrão lexicogramatical ocorre. Assim, um corpus pode ser definido como:

> […] um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de proporcionar resultados vários e úteis para a descrição e análise. (Sanchez, 1995, pp. 8-9).

Como podemos ver, o computador desempenha um papel primordial nas pesquisas da área. Ferramentas computacionais possibilitam a reorganização e extração de informações do corpus, a observação e interpretação de dados, além de viabilizar a identificação das regularidades e padrões[9].

Biber, Conrad e Reppen (1998), apontam o uso conjunto das técnicas de análise quantitativa e qualitativa como uma das principais características da abordagem baseada em corpus. As análises quantitativas (quando a contagem é parte fundamental da investigação) estão intimamente ligadas a estudos com base em corpus. Essa ligação acontece uma vez que, ao se trabalhar com uma grande quantidade de dados, torna-se possível obter frequências e aplicar testes estatísticos, bem como dispor os resultados em gráficos, além de realizar inúmeras comparações

9 Segundo Berber Sardinha (2005, p. 216), padrões são “associações entre itens lexicais, categorias gramaticais, semânticas ou pragmáticas, observadas num corpus”.

entre os números obtidos. Assim, no momento da análise qualitativa, interpretações funcionais (e lexicais, como é o caso desta pesquisa, especificamente) são aplicadas aos dados quantitativos identificados anteriormente. A aplicação da análise qualitativa com base na análise quantitativa torna-se imprescindível, na medida em que explica as razões de maior ou menor incidência de certos padrões de ocorrência de uma palavra – por exemplo, nos textos analisados.

Assim, o uso conjunto de análises tanto quantitativas quanto qualitativas na LC corrobora a necessidade de se realizar pesquisas que estejam voltadas para o que é provável, ao invés do que é teoricamente possível. Leech (1992) apresenta alguns aspectos que devem ser levados em consideração no trabalho com pesquisa empírica. Em linhas gerais, esses aspectos são: (1) por maior que seja um corpus, ele é apenas uma amostra da língua em uso; (2) os dados a analisar não devem ser escolhidos de acordo com a preferência do pesquisador, e sim aleatoriamente, e nenhum deles pode ser considerado irrelevante para a pesquisa; (3) teorias ou modelos podem ser criados para explicar os dados encontrados (a partir da intuição ou experiência do investigador, por exemplo), mas os valores quantitativos do modelo devem ser obtidos a partir dos dados do corpus; (4) a precisão do modelo pode ser testada em outro corpus; e (5) a princípio, a qualidade de um modelo pode ser medida e comparada com a de outros modelos (essa interação é importante para que os modelos de desempenho linguístico sejam progressivamente aperfeiçoados) e diferentes modelos podem ser testados com o mesmo corpus, para atestar a superioridade de um modelo em relação a outro.

2.3. CORPORA ELETRÔNICOS GERAIS E ESPECIALIZADOS

Sarmento (2009, pp. 264-269) elenca os tipos de corpora abaixo:

1. Corpus geral: contém diversos tipos de textos, sendo eles provenientes da linguagem falada, escrita ou ambas; produzidos em um determinado país ou em vários. Por ser um corpus de língua geral, é necessário ter a maior variedade de tipos de textos possível, com dimensões muito maiores do que um corpus específico. São muitas vezes chamados de Corpus de Referência por serem utilizados para contrastar estudos com corpora especializados. Um dos mais famosos é o BNC.

2. Corpus monitor: projetado para averiguar as mudanças recentes em uma língua. É alimentado com frequência (diária, mensal ou anualmente) e sempre com a mesma proporção de tipos de textos a fim de que se possa comparar cada período com o anterior. O COCA é um exemplo desse tipo, contendo, atualmente, 1 bilhão de palavras.

3. Corpus comparável: composto por dois ou mais corpora em línguas diferentes ou diferentes variedades de uma mesma língua, cuja compilação segue os mesmos critérios. Esses tipos de corpora são utilizados em sua maioria por tradutores e/ou aprendizes de uma língua para identificar suas diferenças e equivalências. Um exemplo é o CorTec, do Projeto COMET, da Universidade de São Paulo, que contém 19 corpora técnicos comparáveis.

4. Corpus paralelo: composto por dois ou mais corpora em línguas diferentes contendo textos originais, e suas respectivas traduções, ou textos produzidos ao mesmo tempo em duas ou mais línguas, como, por exemplo, as normas da União Europeia.

5. Corpus de aprendiz: composto por textos ou redações produzidas por aprendizes de uma determinada língua. Esse tipo de corpus tem o objetivo de identificar os mecanismos de aquisição de segunda língua ou língua estrangeira e/ou contrastar aspectos em que os aprendizes diferem entre si e em relação a falantes nativos. É provável que o maior corpus de aprendiz seja o International Corpus of Learner English (ICLE), contendo produções escritas de aprendizes de língua inglesa e falantes de 19 línguas nativas diferentes, incluindo o português, cujo subcorpus (BrICLE) está armazenado na PUC-SP.

6. Corpus pedagógico: composto pela linguagem que se expõe a um aprendiz. Pode conter livros didáticos e gravações.

7. Corpus histórico ou diacrônico: composto por textos provenientes de diferentes períodos de tempo. Seu propósito é identificar a evolução de determinados aspectos de uma língua através do tempo. Um exemplo é o corpus da revista TIME, que abrange o conteúdo publicado na revista de 1923, ano em que foi lançado, até 2006.

8. Corpus especializado: composto por um tipo específico de texto, registro ou gênero textual, definido de acordo com a tipologia de corpus, cujo objetivo é representar um determinado tipo de texto ou linguagem. Geralmente pode ser compilado pelo próprio pesquisador. Apesar de não haver limites para o grau de especialização envolvido, deve-se seguir os parâmetros da tipologia textual escolhida.

Com o advento da revolução tecnológica, em especial a partir da década de 1960, estudos linguísticos com corpora de maiores proporções tornaram-se possíveis. Um marco para a LC foi o lançamento do primeiro corpus eletrônico de linguagem escrita, o Brown University Standard Corpus of Present-Day American English, em 1964, contendo 1 milhão de palavras, distribuídas em 500 textos de diversos registros escritos em inglês americano e com data de publicação de 1961.

A criação do corpus Brown contribuiu para as futuras pesquisas na área, uma vez que foi estabelecido um padrão para corpora eletrônicos ao sistematizar e documentar a coleta de textos, bem como disponibilizado a outros pesquisadores. Sua criação incentivou a compilação do corpus Lancaster/Oslo/Bergen (LOB), um corpus espelho em inglês britânico. O corpus LOB foi coletado entre 1970 e 1978, em uma parceria entre pesquisadores da Universidade de Lancaster, Universidade de Oslo e do Centro de Computação para Humanidades Norueguês. Assim como o corpus Brown, O LOB também possui, aproximadamente, um milhão de palavras, distribuídas em 500 textos de linguagem escrita publicados em 1961, sendo que cada texto contém uma amostra de 2.000 palavras. Ambos os corpora estão agrupados em 15 categorias de texto dos mais diversos, contendo desde registros de imprensa, de cunho religioso, histórias populares, documentos governamentais, escritos científicos, até registros de ficção em geral, conforme descrito a seguir:

Quadro 1: *Timeline* dos principais corpora

Corpus	Data	Número de palavras	Variante do idioma
Brown Corpus	1964	1 milhão	Inglês americano e escrito
LOB (Lancaster-Oslo-Bergen) Corpus	1978	1 milhão	Inglês britânico escrito
LLC (London-Lund Corpus)	1980	500 mil	Inglês britânico falado

Birmingham Corpus	1987	20 milhões	Inglês britânico
TOSCA Corpus	1988	1,5 milhão	Inglês britânico escrito
SEU Corpus	1989	1 milhão	Inglês britânico escrito e falado
LCLE (Longman Corpus of Learner's English	1992	10 milhões	Inglês escrito por estrangeiros
SEC (Lancaster/IBM Spoken English Corpus)	1992	53 mil	Inglês britânico falado
English Corpus	-	-	-
Wellington Corpus (of Written New Zealand English)	1993	1 milhão	Inglês neozelandês, escrito
POW (Polytechnic of Wales Corpus)	1993	65 mil	Inglês infantil falado
BNC (British National Corpus)	1995	100 milhões	Inglês britânico escrito e falado
Wellington Corpus of Spoken New Zealand English	1995	1 milhão	Inglês neozelandês, falado
ICLE (International Corpus of Learner English	1997	2,5 milhões	Inglês escrito por estrangeiros
Bank of English	1997	450 mil	Inglês

Fonte: Adaptado pelo autor com dados extraídos de Berber Sardinha (2004)

Outro corpus de grande importância para a LC, e tido como referência para subsequentes corpora, é o Survey of English Usage (SEU), da University College London. Compilado manualmente a partir de 1953 por um time de pesquisadores liderados por Randolph Quirk, o corpus continha um milhão de palavras de linguagem escrita e oral quando finalizado. Seu desenho seguiu critérios específicos, sendo composto por um número fixo de textos (200), contendo 5.000 palavras cada. O SEU tornou-se um corpus eletrônico em 1989, contendo não somente os textos originais, mas também textos de linguagem falada de um projeto irmão, o Survey of Spoken English (SEE), da Lund University. A parceria entre os dois projetos, SEU e SEE, resultou na criação do London-Lund Corpus of Spoken English (LLC). O LLC continha 87 textos em sua versão original, porém 13 textos foram posteriormente adicionados, totalizando 100 textos de linguagem oral.

Nos dias de hoje, além das pesquisas baseadas nos corpora Brown, LOB e LLC, os corpora de inglês geral mais utilizados para pesquisa em LC são: o BNC (British National Corpus) de 1995, que contém 100 milhões de palavras e é composto por inglês britânico, falado e escrito; o Bank of English, lançado em 1987, que contém 450 milhões de palavras e é composto por textos do inglês britânico (e também contém um subcorpus com 56 milhões de palavras disponível para o ensino de língua inglesa); e o COCA (Corpus of Contemporary American English) lançado em 2008, que contém mais de 460 milhões de palavras e é composto por textos em inglês americano, falado e escrito. Há diversos tipos de corpora e a classificação deles varia conforme o tamanho, o propósito e a maneira como foram compilados.

Embora todos os corpora mencionados sejam igualmente importantes para a LC em geral, Biber (1988) baseou-se nos corpora LOB e LLC para determinar as dimensões de variação da língua inglesa falada e escrita. As pesquisas desenvolvidas pelo estudioso lançam bases teóricas e metodológicas para o desenvolvimento de diversas pesquisas, formando a base de estudos que norteia a presente pesquisa.

Berber Sardinha (2000) ressalta ainda que a Linguística de Corpus saiu dos centros universitários e está presente em diversas empresas de informática e editoras, resultado dos avanços da Linguística de Corpus.

2.4. CONCEITOS CHAVE DA LINGUÍSTICA DE CORPUS

Alguns dos conceitos-chave da LC tiveram sua origem nas ideias de John Rupert Firth, linguista inglês que influenciou uma geração inteira de linguistas britânicos por meio de suas aulas, por mais de 20 anos, na Universidade de Londres. Firth acreditava que a língua deveria ser estudada como um conjunto de eventos expressos pelo falante, que deveriam ser examinados em si. A significação se encontrava tanto nas palavras que haviam sido proferidas (sobre o que se falava, quem falava e como se falava) quanto no contexto situacional no qual haviam sido ditas. Sua noção de significado era tão abrangente que abarcava vários níveis linguísticos: o discursivo, o textual, o gramático, o lexical e o fonético.

O conceito de colocação postula que parte da significação de uma palavra se encontra nas palavras que coocorrem com ela de modo recursivo. Uma frase célebre do autor representa esse posicionamento, “uma palavra deve ser julgada por sua

companhia” (Firth, 1957, p. 11). Foi essa noção firthiana que fundamentou a visão que a LC possui sobre a linguagem: a de um sistema probabilístico (Berber Sardinha, 2004), o que prevê como a linguagem é utilizada em textos naturais, em oposição à qual a linguagem seria teoricamente possível.

Assim, a Linguística de Corpus, nos dias de hoje, fornece uma abordagem que possibilita a observação de redes semânticas e campos lexicais, o que facilita o trabalho do analista ao ter de manusear grande quantidade de dados. Dessa forma, a LC destaca-se pela capacidade de análise de padrões da linguagem em textos naturais em uma quantidade impossível de ser realizada manualmente.

De acordo com essa visão probabilística da linguagem, “embora muitos traços linguísticos sejam possíveis teoricamente, não ocorrem com a mesma frequência” (Berber Sardinha, 2004, p. 30), enquanto que, “o mais importante da diferença de frequências entre os traços é não serem aleatórias” (Berber Sardinha, 2004, p. 31). Isso significa que, embora muitas combinações e características da estrutura da língua sejam possíveis, não ocorrem com a mesma frequência. E, se essas frequências fossem aleatórias, não acrescentariam informações significativas em relação à lexicogramática (Berber Sardinha, 2012, 2020). Ou seja, existe uma variação sistemática de grupos de traços linguísticos, não-aleatória, em relação a textos provenientes de situações comunicativas específicas. Isso indica que há uma padronização da linguagem, que é evidenciada pela recorrência. Isto é, colocações, coligações ou prosódia semântica, que se repetem de modo significativo, parecem ser, na realidade, padrões lexicais ou lexicogramaticais (Berber Sardinha, 2004, p. 31).

Pesquisas baseadas em corpus requerem o uso de um grupo de ferramentas computacionais de análise quantitativa de dados, sendo que a análise qualitativa é baseada nos resultados quantitativos obtidos. De acordo com Berber Sardinha (2004), as ferramentas mais comuns à disposição da Linguística de Corpus são os programas para classificar palavras, que fazem a contagem das palavras em um corpus; os concordanciadores, que são programas que permitem que o usuário procure por palavras específicas em um corpus, e os etiquetadores, que fazem análises automáticas do corpus e inserem etiquetas (códigos) de ordem morfossintática, sintática, semântica ou discursiva. Essas ferramentas que a Linguística de Corpus

dispõe podem ser usadas na organização e extração de informações dos corpora, que viabilizam a observação e interpretação de dados, fornecendo novas perspectivas à análise linguística.

De acordo com Berber Sardinha (2004) há quatro tipos de anotação linguística: (1) morfossintática ou marcação de partes do discurso (part of speech – POS), (2) sintática (parsing), (3) semântica (semantic) e (4) discursiva (discourse). A etiquetagem ou marcação morfossintática, semântica e discursiva é realizada por programas denominados “etiquetadores” (taggers) e consiste em adicionar “um código a cada palavra do corpus, indicando a parte do discurso” (Hunston, 2002, p. 18). A marcação sintática é realizada por programas denominados “parsers”, os quais adicionam etiquetas contendo informações que identificam estruturas sintáticas, como por exemplo, sintagmas nominais e verbais.

Análises estatísticas também podem ser aplicadas às análises de dados provenientes de um corpus anotado. Para isso, utilizam-se pacotes estatísticos, tais como o Statistical Analysis System (SAS) e R. Nesse tipo de análise, não são os textos d o corpus que são inseridos na ferramenta, mas sim dados relativos às frequências das variáveis linguísticas que estão sendo analisadas nos textos.

Ferramentas computacionais denominadas counters são usadas para contabilizar as etiquetas presentes no corpus e os resultados são apresentados em documentos de texto sem formatação (.txt) que são transpostos para .xlxs (no Microsoft Excel, por exemplo). Essas planilhas, por sua vez, são inseridas no pacote estatístico escolhido para serem processadas. A partir dos dados importados para o pacote estatístico, torna-se possível verificar as semelhanças e diferenças entre os textos de acordo com padrões de coocorrência existentes entre as variáveis linguísticas do estudo. Para isso, estatísticas descritivas e análises fatoriais são empregadas.

2.5. COLOCAÇÃO

Talvez uma das contribuições mais duradouras da LC para a compreensão da língua em uso seja a colocação (Berber Sardinha, 2004). Ainda segundo o autor, a criação de concordanciadores - listas de palavras exibidas no ambiente linguístico em que elas são usadas - atesta para a utilidade da colocação como um conceito central

da LC. Os concordanciadores foram criados no início como uma espécie de índice à Bíblia e a outros textos tidos como importantes o suficiente para justificar a importância da colocação para a análise lexical. Os concordanciadores revelam o papel de colocação no desambiguamento do sentido e na interpretação do contexto em que as palavras se encontram.

Sinclair (1991) estabelece o fenômeno como sendo a ocorrência de dois vocábulos mais próximos um ou outro dentro de um texto[10]. O fato de a colocação extrapolar uma combinação randômica leva Berber Sardinha a mencionar as medidas estatísticas de associação resultantes da colocação.

O fenômeno da colocação manifesta-se de duas formas: ascendente e descendente. Sinclair detalha que quando duas palavras de diferentes frequências se combinam, a colocação tem um valor diferente na descrição das duas palavras. Se a palavra X é duas vezes mais frequente que a palavra Y, então cada vez que elas ocorrem juntas é duas vezes mais importante para Y do que para X. (Sinclair, 1991, p. 115): " Quando “X” é o nódulo e “Y” é o colocado, chamarei isso de colocação descendente – a colocação de X com a palavra menos frequente (Y). Quando “Y” é o nódulo e “X” o colocado, é classificada como uma colocação ascendente (Sinclair, 1991, p. 115).

Por fim, a colocação ilustra o princípio idiomático:[11]:

"[o] princípio idiomático consiste em o usuário da língua dispor de um vasto número de frases semi construídas que constituem escolhas únicas, ainda que possam ser analisadas em segmentos. De certa forma, isso pode refletir a recorrência de situações similares em assuntos humanos, ou pode ilustrar uma tendência natural para a economia de esforços, ou pode ainda ser motivada em parte pelas exigências da comunicação em tempo real.” (Sinclair, 1991, p. 110)

10 Collocation is the occurrence of two or more words within a short space of each other in a text.

11 Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text.

2.6. ANÁLISE DO DISCURSO NA LC

Nesta seção, tratamos do conceito de discurso no escopo da LC (Gillings et al., 2023).

O discurso, como fenômeno abstrato, reflete os valores e as ideologias historicamente atribuídas a grupos ou setores sociais. Apesar de sua natureza abstrata, os discursos se materializam na língua em uso. Berber Sardinha e Fitzsimmins-Doolan (2025) demonstram como o léxico serve de marcador da formação de discurso e alinhamento de ideologias. A detecção computacional dessa materialidade possibilita a quantificação dos índices de discurso. Contudo, a análise dos padrões levantados a partir dos dados empíricos provenientes da língua somente se valida com uma abordagem qualitativa, pois os discursos não emergem automaticamente de seus índices.

Os corpora se apresentam como recursos para a análise quantitativa de dados linguísticos, sendo assim possível identificar padrões, frequências e distribuições de características linguísticas dentro de um conjunto de dados. Com esses resultados em mãos se pode testar hipóteses de maneira sistemática e replicável e possuir uma base para atender às reivindicações teóricas da Análise do Discurso.

Entretanto, não existe uma abordagem definitiva para a Análise do Discurso, o que não significa que todos os métodos sejam viáveis. Além de aplicar conceitos, qualquer abordagem em AD requer uma reflexão mais abrangente de natureza histórica e filosófica, que vai além da simples aplicação de técnicas estanques e procedimentos seriados.

Nesse contexto, selecionamos algumas definições de discurso que mais atendem aos propósitos de nossa pesquisa. São elas:

Um conjunto de ideias, conceitos e categorizações que são produzidos, reproduzidos e transformados em um conjunto específico de práticas e por meio dos quais é atribuído significado às realidades físicas e sociais. (Hajer, 1993).[12]

Formas de olhar o mundo, de construir objetos e conceitos de certas maneiras, de representar a realidade (Baker & McEnery, 2015).[13]

É importante frisar que a Análise Multidimensional Lexical contribui metodologicamente aos estudos em Análise (Crítica) do Discurso assistida por Corpus em termos de detalhamento e amplitude estatística e reforça a natureza multidimensional do discurso. Entendemos então que a análise dos resultados aqui gerados só se concretiza através da aplicação da metodologia da AMD Lexical, com o suporte das teorias discursivas.

2.7. ANÁLISE MULTIDIMENSIONAL

A Análise Multidimensional é uma abordagem metodológica desenvolvida por Biber (1988). De acordo com Berber Sardinha (2004, p. 300), a Análise Multidimensional (AMD) é uma abordagem para a análise de corpus que usa procedimentos estatísticos, em especial a análise fatorial, com o intuito de mapear as associações entre um conjunto variado de características linguísticas dentro do corpus de estudo. Utiliza, também, procedimentos automáticos e semiautomáticos para análise do corpus, tais como etiquetagem morfossintática (part of speech tagging). Como explica Berber Sardinha (2004), a AMD formou a base para estudos de linguagem baseada em corpus eletrônico.

Segundo Biber (1988), a Análise Multidimensional é uma abordagem metodológica baseada em corpus voltada a: 1) identificar os padrões linguísticos

12 Original: “An ensemble of ideas, concepts, and categorizations that are produced, reproduced and transformed in a particular set of practices and through which meaning is given to physical and social realities”.

13 Original: “Ways of looking at the world, of constructing objects and concepts in certain ways, of representing reality”.

salientes coocorrentes em uma língua, em termos empíricos/quantitativos e 2) comparar registros no espaço linguístico definidos por esses padrões de coocorrência.

Considera-se registro um termo amplo para qualquer variedade da língua, definida por suas situações características, incluindo o objetivo do falante, a relação entre falante e ouvinte e as circunstâncias de produção. No caso desta tese, o trabalho possui o registro do Twitter.

A AMD se baseia no procedimento estatístico de análise fatorial, “usado para agrupamento de dados com base em sua coocorrência” (Berber Sardinha, 2004, p. 300). A análise fatorial revela os fatores, que representam conjuntos de características correlacionadas. Esses fatores, quando avaliados qualitativamente para determinar sua função comunicativa, tornam-se dimensões. De acordo com Berber Sardinha (2004, p. 304), as dimensões permitem “visualizar características partilhadas por uma porção significativa de dados”. A interpretação dos fatores faz surgir funções comunicativas na forma de dimensões de variação, que são padrões de coocorrência de elementos léxico-gramaticais subjacentes aos registros de uma língua.

Conforme destaca Kauffmann (2020), a análise fatorial é central na Análise Multidimensional, pois permite condensar um conjunto amplo e variado de variáveis em um número reduzido de fatores, os quais explicam grande parte da variação observada. Essa condensação seria impraticável manualmente, sendo viabilizada pela aplicação da análise fatorial multivariada, que identifica relações estatisticamente significativas entre as variáveis lexicais do corpus.

Na AMD, acredita-se que um grupo de características coocorre frequentemente nos textos de forma sistemática "porque atendem a alguma função comunicativa comum, e a interpretação da dimensão funcional subjacente às dimensões é uma hipótese e deve ser confirmada através da análise qualitativa dos textos que compõem um corpus” (Biber, 1988, p. 91).

Conforme coloca Berber Sardinha (2010), a AMD é uma metodologia que propicia um olhar em larga escala sobre questões importantes da constituição da língua e do discurso, na medida em que enfoca muitos textos de vários registros ao mesmo tempo. O objetivo da AMDL é interpretar uma seleção vasta e variada de corpora para identificar de forma precisa os discursos e/ou ideologias subjacentes

(Berber Sardinha & Fitzsimmons-Doolan, 2025). Embora a AMDL adote os mesmos procedimentos metodológicos e parta dos mesmos pressupostos teóricos sobre a variação como a AMD, a primeira lança mão apenas dos traços lexicais da linguagem, o que resulta em dimensões conceituados como construtos discursos, geralmente conhecidos como discursos, ideologias ou temas.

Alguns estudos realizados por pesquisadores do Grupo de Estudos da Linguística de Corpus (GELC) atestam para a eficácia da AMDL como metodologia de pesquisa para identificar discursos e temas semânticos recorrentes em um corpus.

Kauffmann (2020) analisou o estilo literário de Machado de Assis a partir de uma abordagem da Linguística de Corpus. Apesar de embasar seu estudo na análise multidimensional funcional, o autor recorrer à vertente lexical para examinar o uso de lemas frequentes e seus agrupamentos semânticos. Esta integração dessas abordagens permitiu uma análise mais profunda e estilisticamente sensível da obra de Machado de Assis.

Araújo (2025), por sua vez, investigou a variação lexical em reality TV shows norte-americanos para identificar os principais temas e discursos presentes na linguagem verbal nessa modalidade de entretenimento televisionado dos EUA por meio da Análise Multidimensional Lexical. Um corpus de mais de 2 milhões de palavras foi etiquetado com o TreeTagger, filtrado para palavras gramaticais com o uso de stoplists, utilizado para extrair listas de vocábulos mais frequentes e submetido a uma análise fatorial para determinar os padrões de coocorrência lexical. As dimensões lexicais resultantes foram interpretadas como campos semânticos representando discursos ou temas recorrentes. O estudo confirmou que a linguagem dos reality shows reflete variações temáticas distintas, conforme o tipo de programa, evidenciando que o léxico é altamente sensível ao tópico. Isso permitiu o mapeamento de discursos específicos e identificar padrões consistentes de variação lexical.

Outro exemplo do uso da AMDL é o estudo de Brogini (2022) sobre o termo sustentabilidade em postagens no Twitter entre 2018 e 2022. Usando as ferramentas Snscrape e TreeTagger para coletar e processar o corpus de aproximadamente 2,8 milhões de palavras respectivamente, a autora selecionou lemas, substantivos, verbos, adjetivos e hashtags, aplicou análise fatorial com rotação Promax para

detectar oito dimensões discursivas principais, cada uma representando sentidos distintos, e às vezes conflitantes, do termo "sustentabilidade". Com isso, concluiu-se que um único sentido para o termo “sustentabilidade” inexiste, e que na realidade há dezesseis sentidos distintos agrupados em pares antagônicos. O estudo evidencia um conflito discursivo entre abordagens ambientais e empresariais, destacando a complexidade e disputa de sentidos em torno do termo.

2.8. ANÁLISE MULTIDIMENSIONAL LEXICAL

Em 2014, Berber Sardinha propôs um novo modelo para a Análise Multidimensional apresentada por Biber (1988), chamada de Análise Multidimensional Lexical (AMDL), que foi desenvolvida com o intuito de observar a variação linguística identificando os campos semânticos subjacentes formados pelos itens que mais coocorrem em um corpus. O quadro a seguir traz um comparativo entre as duas vertentes a AMD Funcional e Lexical.

Quadro 2: Comparação entre AMD Funcional e Lexical

	Funcional	Lexical
Objetivo	Identificar parâmetros subjacentes de variação nos textos de um corpus	
Unidade de observação	Textos/segmentos	
Traços linguísticos	Lexico-gramaticais	Lexicais
Base da interpretação	Funcional, comunicativa	Campos semânticos, preferência semântica

Fonte: O autor

Na Análise Multidimensional Lexical, a análise fatorial começa com a definição das variáveis a serem analisadas, que são representadas pelos itens lexicais, definidos por seus lemas, oriundos de três classes gramaticais: substantivos comuns, adjetivos e verbos. A análise fatorial busca reduzir a dimensionalidade dos dados ao extrair fatores, facilitando a interpretação de complexas relações entre as variáveis. Após a identificação das variáveis, elas são agrupadas com base em características similares em fatores. Esses fatores são extraídos por meio de uma análise fatorial exploratória, técnica estatística que busca reduzir a dimensionalidade dos dados e identificar as relações subjacentes entre as variáveis.

Cada fator pode ter polo positivo e polo negativo, os quais não possuem um juízo de valor, ou seja, um não é superior ao outro, mas representam características complementares, sendo distribuídos de maneira oposta ao longo de um eixo (Delfino, Berber Sardinha, & Collentine, 2021). A partir dessa distribuição, é possível identificar a associação entre os itens lexicais e os fatores extraídos.

Na AMD funcional, as dimensões de variação são nomeadas, descritas e classificadas por meio das classes gramaticais, tipos de orações e categorias semânticas. Essas dimensões da AMD correspondem aos parâmetros funcionais de variação da linguagem presentes nos textos analisados. Na AMDL, as dimensões de variação são identificadas por meio das próprias palavras, lemas, n-gramas, ou colocações presentes nos textos.

Pensando justamente na possibilidade de identificar as dimensões de variação através do léxico, entendemos que se considera como sendo “texto”. Na Linguística de Corpus (LC), a expressão “porções de linguagem” (Berber Sardinha, 2004, p. 17) parece mais adequada para dar conta dos desafios em delimitar o conceito de texto, tais como um artigo científico, um resumo inicial, ou ainda um trecho de um diálogo. Deste modo, a AMDL possibilitou a identificação dos discursos (termo que engloba posicionamentos ideoógicos, valores e crenças que subjazem as conversas dos bate-papos educacionais do Twitter) ao detectar grupos de itens lexicais que coocorrem nos tuítes. Para a análise, 127.418 tuítes foram submetidos à etiquetagem e lematização, seguidas de análise fatorial para definir as sete dimensões lexicais.

Outra diferença importante entre a AMD funcional e a lexical é que durante a fase qualitativa da AMDL, o pesquisador interpreta as características lexicais a fim de detectar os temas ou discursos presentes no corpus, em vez de interpretar as características gramaticais a fim de detectar as funções comunicativas desempenhadas nos textos. Assim, na AMDL os fatores são interpretados em termos

de “preferências semânticas, conjuntos lexicais, campos lexicais, aboutness[14], tópicos e assunto” (Berber Sardinha, 2017c, p. 2).

Destacam-se três pesquisas feitas com a AMDL. A primeira identificou as dimensões de variação lexical subjacente e agrupou os textos em períodos históricos relevantes (Berber Sardinha, 2017a). Para isso, foi utilizado um corpus diacrônico composto por publicações da revista TESOL Quarterly de 1967 a 2014. A segunda pesquisa (Berber Sardinha, 2017b) identificou, a partir de um corpus diacrônico com mais de 11.000 textos dos principais periódicos de Linguística Aplicada dos últimos 70 anos, os temas mais recorrentes e as principais tendências e mudanças na área da Linguística Aplicada.

A terceira pesquisa explorou as representações culturais relacionadas aos adjetivos *American* e *Brazilian* em um corpus diacrônico formado por textos em inglês provenientes do corpus de n-gramas do Google Books, uma base de dados contendo documentos (na sua maioria livros) publicados de 1800 até 2008 (Berber Sardinha, 2019). Para tanto, Berber Sardinha investigou os colocados aos adjetivos American e Brazilian para identificar os parâmetros de representação de identidade nacional e cultural para entender o que significa ser americano ou brasileiro ao longo do tempo representado no corpus.

Conforme mencionado, a AMDL é uma abordagem que visa a detectar os parâmetros de variação lexical de um corpus, os quais podem sinalizar desde temas até discursos nos dados. Como a AMD Funcional, essa abordagem faz uso da análise estatística multivariada, particularmente a análise fatorial, para detectar as variáveis latentes, isto é, aquelas que operam abaixo da percepção imediata do falante. Essas variáveis subjacentes se concretizam em dimensões de variação, que são conjuntos de itens lexicais correlacionados, ou seja, que tendem a ocorrer simultaneamente nos textos, sejam eles escritos, falados, sonoros ou visuais (Moreira, 2023). Na AMD tradicional – também conhecida como Funcional – as principais categorias

14 O termo “aboutness” abrange, além do tópico dos textos, a representação construída através das características lexicais. Nesse sentido, pode-se determinar o posicionamento do(s) autor(es) do(s).

gramaticais, estruturas de sentença e classes semânticas são empregadas para rotular, descrever e categorizar as dimensões de variação que correspondem aos parâmetros funcionais da linguagem presentes nos textos analisados. Já a AMD Lexical considera como variáveis, ou unidades de análise, palavras de conteúdo (substantivos, verbos, adjetivos e advérbios), grupos de palavras (colocações e n-gramas) e elementos extralinguísticos (hashtags e emojis) para identificar as dimensões de variação. Os emojis são automaticamente convertidos em rótulos descritivos, permitindo que sejam tratados como itens lexicais.

Associações com valores, crenças e ideologias específicas podem não se manifestar com clareza em uma análise puramente estrutural da gramática. O emprego de determinadas palavras e expressões muitas vezes é moldado por eventos históricos ou tendências sociais, elementos que nem sempre se refletem nas estruturas gramaticais. Conceitualizações abstratas como representações culturais ou ideologias são, admitidamente, mais difíceis de serem identificadas em textos devido à sua natureza oculta. Embora os princípios sejam semelhantes aos da AMD Funcional, a abordagem lexical adota um critério específico que seleciona palavras lematizadas, ou seja, palavras agrupadas pelo mesmo núcleo significativo, incluindo todas as suas formas flexionadas, tanto verbais quanto nominais. Enquanto a AMD Funcional possui uma lista de variáveis limitada pelo número de categorias gramaticais levantadas pelos programas etiquetadores, na AMDL a lista de variáveis é aberta, variando de acordo com os textos componentes do corpus estudado. Essa abordagem orientada por dados, permite que as dimensões lexicais surjam dos textos, capturando os principais parâmetros lexicais subjacentes à variação intertextual. As dimensões lexicais resultantes desse enfoque lançam luz sobre uma série de fenômenos linguísticos realizados pelo léxico. O passo-apasso para a realização da AMDL nesta pesquisa envolve:

- Identificação e contagem das palavras;
- Normalização das frequências das variáveis lexicais;
- Extração fatorial inicial não rotacionada baseada nas frequências normalizadas para identificar os fatores a serem utilizados;
- Scree plot: definição do número de fatores para análise por meio de um gráfico de análise de sedimentação;

- Eliminação das variáveis lexicais com comunalidades menores que 0,15 (Cf. BIBER, 2006, p. 183);
- Extração fatorial final rotacionada contendo o número de fatores estabelecidos para análise;
- Cálculo da quantidade de variação compartilhada pelos fatores extraídos;
- Checagem da variância dos fatores;
- Cálculo dos escores de fator de cada texto;
- Interpretação dos fatores em termos de seus discursos subjacentes através da observação dos textos, registros e variáveis.

Uma gama de estudos de LC realizou suas análises utilizando a AMD Lexical, abordando uma variedade de registros linguísticos. Muitos são os exemplos dessa diversidade, como o trabalho de Mayer (2018), que investigou a variação lexical em conteúdo gerado por usuários na web, provenientes de diferentes sites e redes sociais em inglês. Para tal, compilou um corpus diversificado, que abrangia 15 registros diferentes, com textos escritos por usuários com propósitos variados. A AMD Lexical revelou quatro dimensões temáticas: avaliação e conjecturas sobre pessoas; interlocução crítica; interlocução e conjecturas com foco informacional; e interlocução descritiva.

Romeiro (2020) examinou a obra da fotógrafa Sally Mann através de sua produção textual e das críticas sobre seu trabalho. O corpus reuniu 12 registros diferentes, incluindo livros de fotografia escritos pela artista, textos de parede de suas exposições, artigos da imprensa geral e especializada, entre outros. Esses registros foram extraídos da biblioteca oficial de Mann e de arquivos públicos, abrangendo um período superior a 30 anos. Foram encontradas sete dimensões lexicais relacionadas aos temas subjacentes à obra fotográfica de Mann, centrados no Sul norte-americano, no fascínio pela mortalidade e na questão familiar.

Kauffmann (2020), analisou a variação linguística na prosa ficcional de Machado de Assis, com foco no estilo, visando destacar as principais dimensões estéticas do autor, tanto em termos de comunicação quanto de temática, manifestadas através da língua. Para conduzir a pesquisa, foram coletados dois corpora: o Corpus Literário de Machado de Assis, abrangendo 9 romances e 76 contos do escritor, e o

Corpus Literário Congênere, uma coleção de referência composta por 92 obras de 23 escritores do período de 1850 a 1910. O estudo identificou três dimensões estéticas distintas, nomeadas como: romantismo introspectivo formal; narrativa oralizada sentimental; e representação dramática.

Veiga (2021) realizou um estudo abrangente dos livros sagrados das principais religiões, traduzidos para o inglês, para identificar seus temas mais representativos. O corpus consistiu nos textos dogmáticos de sete religiões: budismo, espiritismo kardecista, hinduísmo, islamismo, judaísmo, mormonismo e protestantismo. Os principais livros de cada uma delas foram coletados e armazenados em formato eletrônico. O estudo identificou seis dimensões lexicais: o mundo dos espíritos e a evolução moral; fluidez, adoração e celebração à força divina; a retidão para esclarecimento espiritual versus a dádiva da terra e o poder do Senhor; crer ou sofrer as consequências versus a casa do Senhor e dos povos; devoção e respeito temente a Deus versus sacrifícios para proteção e espaços celestiais; e ritos sacrificiais de adoração.

Brogini (2022) investigou as práticas discursivas associadas à sustentabilidade, a fim de compreender como o termo é utilizado atualmente no português brasileiro. Para tal, foi compilado um corpus de 127.418 tuítes postados por 42.503 usuários diferentes, no período de 2009 a 2022. Como resultado, foram identificadas sete dimensões lexicais, cada uma compreendendo discursos distintos porém complementares: cultura corporativa versus recurso escasso/insustentável; esfera de poder político versus modelo de negócio; critério de metas corporativas versus tema da educação; matriz energética limpa versus instrumento de marketing; tópico de conhecimento versus crédito tangível; desenvolvimento local versus desenvolvimento global; proteção ambiental versus oportunidade empresarial inovadora; e agronegócio versus filosofia de vida.

Whiteman (2024) investigou o universo discursivo do movimento antivacina brasileiro na plataforma Twitter. Um corpus de cerca de 8 mil tuítes foi coletado, contemplando o período de 2020 a 2022. As sete dimensões encontradas foram rotuladas como defesa das redes sociais; alertas sobre efeitos colaterais das vacinas de Covid-19; crítica ao governo, veículos de mídia oficiais e medidas de saúde pública;

relatos sobre efeitos colaterais das vacinas de Covid-19; resistência à vacinação obrigatória; defesa da autonomia parental; e defesa da liberdade de escolha.

Em suma, a AMDL desempenha um papel central na identificação das dimensões discursivas presentes no corpus compilado para este estudo, viabilizando nossa análise dos discursos que subjazem as postagens nos bate-papos educacionais no Twitter. Para isso, faz-se importante compreender o tuíte como registro.

Observou-se que os rótulos atribuídos às dimensões refletem discursos diretamente relacionados ao processo educacional e à formação continuada docente. Esses encontros síncronos no Twitter têm como objetivo central a construção de um espaço de aprendizagem profissional colaborativa.

A Análise Multidimensional, por definição, emprega um conjunto de métodos estatísticos que permite a análise simultânea de diversas medidas associadas a cada fenômeno linguístico observado. Seu principal objetivo é identificar correlações e padrões entre as variáveis, possibilitando a obtenção de resultados mais abrangentes e generalizáveis.

Nesse contexto, a LC se destaca como a área dos estudos linguísticos que mais se beneficia de ferramentas de processamento de dados e procedimentos estatísticos, especialmente quando se trata da análise de grandes volumes de textos. Essa abordagem probabilística da linguagem tem como marca central a investigação sistemática de padrões de variação linguística, os quais são identificados, descritos e interpretados a partir dos dados empíricos.

O foco da Análise Multidimensional é, portanto, descrever a variação entre diferentes variedades textuais situacionalmente definidas, chamadas de registros (Biber, 1988; Biber et al., 1998). Esses registros podem incluir, por exemplo, a conversação (Biber, Egbert, Keller, & Wizner, 2021), artigos acadêmicos (Gray, 2013; Hardy, 2015), e mensagens em redes sociais (Berber Sardinha, 2021, 2022a, 2022b). À medida que essa variação é descrita com base em dimensões identificadas estatisticamente, pode-se compreender a dimensão como o parâmetro explicativo da variação entre os textos.

Na presente pesquisa, no entanto, conforme colocado, empregamos uma vertente diferente da Análise Multidimensional, que se destina à identificação de grupos correlacionados de itens lexicais, compartilhados entre os textos. Essa vertente, conhecida como Análise Multidimensional Lexical, permite identificar textos que compartilham conjuntos lexicais correlacionados. Tais conjuntos lexicais, os quais são identificados por meio de análise fatorial, assim como na Análise Multidimensional de base gramatical, são interpretados como índices que ajudam o analista a chegar aos temas ou discursos subjacentes.

No âmbito da AMDL, entendemos discurso como uma atividade social que produz sentido, situada historicamente e que pode ser traçada a partir de escolhas lexicais. Tal entendimento se apoia em concepções de discurso como a de Baker e McEnery (2015, p. 5), segundo a qual os discursos são maneiras de olhar o mundo, construir objetos e conceitos de determinadas maneiras, representando a realidade[15] , na de Burr (1995, p. 48), segundo a qual discursos são o conjunto de significados […], representações […], afirmações que ao estarem juntos produzem uma visão específica de eventos[16], na de (Hajer, 1993, p. 44), que entende discursos como um conjunto de ideias, conceitos e categorizações que são produzidas, reproduzidas e transformadas em um determinado grupo de práticas pelas quais o significado é dado para realidades físicas e sociais[17] , bem como na de HALL (1992).

A análise fatorial ao qual o Twitterchat Corpus foi submetido é “um procedimento estatístico usado para revelar estruturas salientes em um conjunto de variáveis, isto é, descobrir padrões simples no relacionamento entre as variáveis”[18]

---

15 No original: Ways of looking at the world, of constructing objects and concepts in certain ways, of representing reality.

16 No original: the set of meanings, […] representations, […] statements and so on that in some way together produce a particular version of events.

17 No original: an ensemble of ideas, concepts, and categorizations that are produced, reproduced and transformed in a particular set of practices and through which meaning is given to physical and social realities.

18 No original: is a statistical procedure used to reveal latent structure of a set of variables, that is, to discover simple patterns in the relationships among variables.

(Cantos Gómez, 2013, p. 113). Em nosso caso, a análise fatorial analisa e observa os conjuntos correlacionados de características léxico-discursivas do corpus, a coocorrência dessas variáveis agrupando-as em fatores.

Com o arquivo de resultado da contagem das frequências normalizadas das variáveis de cada um dos textos do Twitterchat Corpus, realizou-se a análise fatorial do corpus utilizando o pacote estatístico SAS on Demand for Academics.

2.9. TUÍTE COMO REGISTRO

Na LC, em termos gerais, registro se refere a qualquer variedade textual associada a contextos situacionais ou propósitos comunicacionais específicos. Embora as distinções de registro sejam definidas em termos não-linguísticos, geralmente existem diferenças linguísticas importantes entre os múltiplos registros, que se manifestam através de escolhas específicas no léxico, na gramática e na estrutura do texto. Portanto, as relações funcionais entre o contexto situacional e as características linguísticas emergem como o principal componente na descrição de um registro. Um dos principais argumentos sustentados nessa definição é que as características linguísticas são sempre funcionais. Em outras palavras, certas características linguísticas são recorrentes em determinado registro porque se mostram especialmente adequadas aos seus propósitos comunicativos ou ao contexto situacional. Logo, essa análise funcional é um componente essencial em qualquer descrição de registro. Em muitos casos, os registros são identificados como variedades dentro de uma cultura – romances, cartas, editoriais, sermões e debates. Podem ser definidos em diferentes níveis de generalidade: por exemplo, a prosa acadêmica representa um registro bastante amplo, enquanto os resumos de artigos acadêmicos constituem um registro mais específico.

No que concerne ao Twitter, textos gerados em contextos semelhantes têm a tendência de apresentar padrões linguísticos correlatos. As postagens (tuítes) produzidas pelos usuários do Twitter exibem um conjunto de características que definem a plataforma como um registro singular, demarcando-a de maneira única em relação a outras formas de expressão. Ao examinarmos as dinâmicas do Twitter, torna-se evidente que diversos recursos-chave exercem uma influência significativa sobre a linguagem. Cada um desses elementos impõe coerções distintas à expressão

do usuário, moldando, assim, o estilo comunicativo geral da plataforma. Primordialmente, as características situacionais do Twitter desempenham um papel essencial no delineamento do uso da linguagem na plataforma. A brevidade das postagens, uma característica distintiva da plataforma, teve sua origem no limite inicial de 140 caracteres, que refletia a adaptação da rede para mensagens de SMS (Zappavigna, 2017). Essa limitação compele os usuários a destilar suas ideias de forma sucinta, o que resulta no uso de abreviações, siglas, frases condensadas, gramática e pontuação simplificadas e emojis, priorizando a brevidade sobre a complexidade intrincada. O aumento do limite de caracteres do Twitter para 280 em 2017 também foi um marco significativo que impactou diretamente a dinâmica da plataforma. Essa alteração não apenas dobrou o espaço disponível para cada tuíte, mas também desencadeou uma série de transformações na maneira como os usuários se engajam e compartilham informações. Com o novo limite, a natureza das postagens tornou-se mais diversificada, abrindo a possibilidade para maior contextualização. A presença marcante da hashtag exerce grande influência no cenário comunicativo do Twitter. Adicionar hashtags implica incluir uma palavra ou frase precedida pelo símbolo cerquilla (#) em um tuíte, viabilizando a categorização e facilitando a busca por outros usuários. Esse uso estratégico de hashtags não só expande consideravelmente o alcance de um tuíte, abarcando novas audiências, mas também desempenha um papel vital no estabelecimento e na manutenção de conexões sociais (Zappavigna, 2017).

O símbolo @ é usado para a menção explícita de um usuário por meio de um tuíte. Trata-se de uma referência amplificada que pode ser uma ferramenta valiosa de autopromoção, uma vez que outros usuários que seguem o perfil mencionado podem visualizar a interação. Além disso, as menções podem ser consolidadas e pesquisadas por outros usuários. É possível recuperar todas as instâncias de menções para um usuário específico em um intervalo de tempo determinado usando a interface de busca do Twitter, metadados e sua Application Programming Interface (API – Interface de Programação de Aplicação).

O retuite “permite que os membros retransmitam ou encaminhem um tuíte por meio de sua rede”, marcando o texto citado como notável e recomendando-o efetivamente aos seus seguidores. Isso pode ampliar significativamente o alcance de

um tuíte, principalmente quando um usuário com um grande número de seguidores, como uma celebridade, opta por retuitar algo. A convenção marca um tuíte como como digno de atenção dentro desse contexto de conversação e permite que o usuário exiba uma postura em relação ao texto retuitado e o projete como valioso para a comunidade, seja em termos de mérito ou notoriedade.

Além da retransmissão, o retuite contribui para o ecossistema comunicativo da rede, na qual as conversas são compostas por uma interação pública de vozes que dão origem a uma sensação emocional de contexto conversacional. Os retuites são marcados pela sigla RT na maioria dos casos, seguida pelo caractere @ para atribuir o texto ao seu autor original. Retuites também podem ser usados de maneira semelhante à função de resposta, bem como trazer vozes externas para um tuíte. Além disso, o retuite é frequentemente utilizado para sinalizar concordância ou endosso, já que os usuários têm maior probabilidade de retuitar conteúdos que consideram especialmente relevantes e interessantes. Para além das características situacionais, os usuários do Twitter perseguem propósitos comunicativos específicos A qualidade efêmera dos tuítes exige o compartilhamento em tempo real, privilegiando a expressão imediata de pensamentos, sentimentos e notícias em detrimento de comentários reflexivos. Embora propósitos como o compartilhamento de informações sejam comuns em diversas plataformas de comunicação, a forma como os usuários os realiza linguisticamente no Twitter segue determinadas convenções.

A ausência de recursos extensivos de edição, aliada à brevidade dos tuítes, fomenta um registro linguístico informal. Essa configuração tonalidade alinha-se mais estreitamente com a fala do que com a escrita formal, tornando gírias, palavrões e emoticons, partes inerentes da linguagem dessa rede social. Ultrapassando as fronteiras da linguística e dos propósitos comunicativos, as funções interativos da plataforma exercem um papel fundamental na configuração dos discursos na plataforma.

3. METODOLOGIA

Neste capítulo é feita a descrição dos processos que conduziram a pesquisa, desde a coleta dos dados linguísticos que compuseram o Twitterchat corpus até a extração da análise fatorial final a partir de lemas e palavras de conteúdo, que abarcam as palavras mais frequentes do corpus associadas à educação.

Nesta etapa da pesquisa, foram testadas diferentes soluções fatoriais para a AMDL, com base nos valores de eigenvalues. A solução com sete fatores foi considerada a mais interpretável e, portanto, adotada como definitiva. A extração foi rotacionada pelo método promax, assumindo correlação entre os fatores. A partir dos resultados quantitativos, iniciou-se uma análise qualitativa focada nos padrões de uso das variáveis lexicais em textos representativos do corpus. Com base nessa análise, identificaram-se dimensões predominantes, às quais foram atribuídos rótulos interpretativos, refletindo os discursos evidenciados pelas coocorrências lexicais. As próximas seções apresentam em detalhe esse processo de rotulação e interpretação.

Após a identificação das dimensões lexicais por meio da análise fatorial, procede-se à atribuição de rótulos interpretativos. Esses rótulos têm como objetivo refletir os discursos indicados pelas variáveis lexicalmente coocorrentes. Considerando que as dimensões discursivas abrangem uma multiplicidade de sentidos interligados, optou-se por dois tipos de rótulo: um rótulo longo, mais elaborado e abrangente, que expressa de maneira densa o discurso interpretado; e um rótulo curto, utilizado para facilitar a referência no corpo do texto. Ambos os rótulos se referem à mesma dimensão e servem para dar conta da complexidade discursiva evidenciada na análise.

Esta base permitiu a posterior análise qualitativa dos resultados e sua interpretação fundamentada, de acordo com a AMD Lexical (BERBER SARDINHA, 2021).

3.1. COLETA E PRÉ-PROCESSAMENTO DO CORPUS

O presente capítulo mostra todos os passos seguidos para a confecção do corpus aqui apresentada e, para tal, divide-se em três seções principais. A primeira apresenta o corpus de estudo (Twiterchat Corpus), incluindo considerações acerca

dos critérios para escolha do referido corpus, assim como a sua coleta. Já a segunda seção trata da análise dos padrões encontrados no corpus de estudo, relevantes para a análise do discurso assistida por corpus. A terceira seção apresenta os passos metodológicos da AMDL utilizada nesta pesquisa, capaz de permitir um estudo tanto quantitativo como qualitativo dos tuítes que compõem o Twitterchat Corpus. Para uma melhor compreensão, a seção também trará, quando necessário, os parâmetros utilizados na elaboração dos cálculos estatísticos, bem como na interpretação dos resultados obtidos.

Assim como na Análise Multidimensional Funcional, a Análise Multidimensional Lexical utiliza-se de conceitos-chave, os quais são apresentados e descritos aqui a fim de subsidiar o entendimento da exposição:

Registro: usado para definir uma variedade linguística definida por aspectos situacionais, que inclui o propósito do falante, sua relação com o ouvinte, e o contexto de produção. A perspectiva de registo combina uma análise das características linguísticas, que são comuns em uma variedade de textos, com os aspectos situacionais, ou seja, a análise da situação da utilização da variedade (Biber & Conrad, 2009).

Gênero: "a perspectiva de gênero é semelhante à perspectiva de registro, na medida em que inclui a descrição dos propósitos e contexto situacional de uma variedade de texto, mas a sua análise linguística contrasta com a perspectiva de registro, uma vez que incide sobre as estruturas convencionais utilizadas para construir um texto completo dentro de sua variedade" (Biber & Conrad, 2009, p. 2)[19]

Fator: um conjunto de traços que coocorrem significativamente em termos estatísticos. Eles são extraídos por meio da análise fatorial, procedimento estatístico

19 No original: The genre perspective is similar to the register perspective in that it includes description of the purposes and situational context of a text variety, but its linguistic analysis contrasts with the register perspective by focusing on the conventional structures used to construct a complete text within the variety.

em que um grande número de variáveis, os traços linguísticos, são reduzidas a um pequeno conjunto de variáveis subjacentes derivadas.

Dimensões: definidas como conjuntos de características linguísticas coocorrentes, que exercem funções comunicativas. “As dimensões emergem da coocorrência das características linguísticas observadas nos textos e são produto da análise; não são, portanto, postas a priori” (Berber Sardinha, Kauffmann, & Acunzo, 2014, p. 2).

Análise fatorial: técnica estatística, que identifica padrões de coocorrências. Na análise fatorial, uma grande quantidade de variáveis originais é reduzida a um conjunto de variáveis subjacentes, chamado de fator. Por meio da análise fatorial é possível identificar grupos de traços linguísticos que coocorrem com bastante frequência nos textos distribuídos nos fatores. Esses grupos são interpretados funcionalmente, como dimensões textuais. De acordo com o autor, o uso da análise fatorial para textos requer dois pressupostos, a saber: são relativamente poucos os parâmetros funcionais subjacentes de variação linguística em inglês; a coocorrência frequente de traços linguísticos em textos indica a existência de uma função comunicativa subjacente que aqueles traços compartilham.

Padrão fatorial: as variáveis que carregam em cada fator. Cada fator foi interpretado funcionalmente, tendo como base os atributos discursivos, sociais e comunicativos comuns entre as suas características. A interpretação dos fatores leva à definição das dimensões e é portanto, um passo crucial na análise multidimensional.

O pesquisador deve se preocupar em escolher os tipos de textos que vão compor o corpus, o número de textos, a seleção de determinados textos, a seleção de amostras retiradas de textos e o tamanho das amostras de texto. Estes aspectos devem ser levados em consideração na elaboração de um desenho de corpus eficaz.

O desenho de corpus é um dos primeiros passos para o planejamento, a compilação e organização de um corpus de estudo, artefato central e norteador em estudos da LC, que devem atender a procedimentos teórico-metodológicos claros e definidos a priori da coleta, uma vez que de acordo com Sinclair (1991, p. 13) todas as decisões que são feitas sobre o que estará em um corpus, e como esta seleção será organizada, controlam tudo que acontecerá subsequentemente no estudos

baseados em corpora e afirma que “os resultados são somente tão bons quanto o corpus.”

Segundo Berber Sardinha (2004, p. 29), “além de representativo, o corpus deve ser adequado aos interesses do pesquisador, que deve ter uma questão a investigar para a qual necessite de um corpus específico”. Desta forma, o desenho de corpus torna-se crucial para a construção dos corpora, como coluna vertebral, dando-lhe estrutura e forma, garantindo sua representatividade (uma boa amostragem de determinada variedade linguística) e sua relevância às questões de pesquisa e interesses do pesquisador.

O desenho apropriado para um corpus dependerá do que ele pretende representar. Nesta pesquisa, buscou-se obter um gama de bate-papos educacionais identificados pelas suas respectivas hashtags ao longo de quinze anos. Segundo Berber Sardinha (2004), um corpus deve ser representativo de uma língua ou variedade linguística, pois é uma amostragem de certa variedade textual que contém características linguísticas articuladas ao contexto de uso e atendendo a funções comunicativas específicas.

O corpus gerado para fins deste estudo, o Twitterchat Corpus, contém 127.418 tuítes marcados por um ou mais de 982 hashtags educacionais no Twitter entre 2009 e 2022. As hashtags foram criadas pelos próprios usuários e tiveram como objetivo reunir profissionais da educação para a partilha de experiências e troca de informações, formando o que a literatura define como comunidades de prática e/ou aprendizagem. As conversas geradas durante estes bate-papos foram arquivados e disponibilizadas em um repositório digital, sendo este hospedado em uma plataforma digital de desenvolvimento profissional contínuo (Participate Learning). Dessa forma, o Twitterchat Corpus se caracteriza por ser especializado, escrito, sincrônico, e estático.

Para coletar o corpus, foram compilados tuítes contendo as hashtags educacionais no Twitter usando a ferramenta snscrape, por meio do comando:

```
snscrape --jsonl --progress --max-results 100000 twitter-hashtag  "$hashtag"
```

A saída foi gravada em um arquivo do tipo JSON, estruturado em forma de campos delimitados. Em seguida, esse arquivo foi processado por um script em Unix Shell desenvolvido pelo professor orientador, que realizou as operações descritas no Quadro 3:

Quadro 3: Funções do script de processamento do corpus para a pesquisa

Função	Descrição	Função	Descriçao
Cleantweets	Limpeza dos arquivos JSON, reduzindo-os aos campos essenciais para a análise, que são o texto, o usuário, a data e o ID do tuíte	SAS	Criação de arquivos de entrada para o pacotes estatístico SAS OnDemand
removedupes	Remoção de de tuítes duplicados do corpus, inserção de espaço entre ao redor de cada palavra para isolá-la de outras e dos demais elementos ortográficos, como emojis, hashtags e pontuação.	datamatrix	Cálculo de correlação entre as contagens dos lemas por meio de rotinas em Python
tokenizing	Tokenização dos tuítes; isto é, inserção de espaço ao redor de cada palavra para isolá-la de outras e dos demais elementos, como emojis, hashtags e pontuação.	correlationmatrix	Geração de uma matrix de correlação para o SAS OnDemand
emoji	Conversão de emoji em uma etiqueta textual descritiva com a bilioteca em Python demoji	dates	Listagem das datas de cada tuíte em número de palavras (tokens) e criação de arquivo de metadados formatado para SAS OnDemand

treetagging	Etiquetagem morfossintática e lematização de cada tuíte com o etiquetador TreeTagger, de tal forma a atribuir a cada palavra uma classe gramatical e uma forma vocabular canônica (lema)	wcount	Listagem de extensão de cada tuíte em número de palavras (tokens) e criação de arquivo de metadados formatado para o SAS OnDemand
tokenstypes	Listagem de itens (tokens) e vocábulos (types) associados a classes gramaticais de conteúdo; isto é, substantivos, verbos e adjetivos, além de emojis e hashtags	formats	Criação de arquivos de formato para o SAS OnDemand
removedupes	Remoção de tuítes parcialmente repetidos com base nos itens (tokens)	jqlisting	Geração de arquivo de referência dos dados em JSON para facilitar extração de dados de rotina seguinte
toplemmas	Contagem dos lemas e listagem dos mil lemas mais recorrentes, com base na contagem do número de tuítes em que ocorrem	examples	Listagem de exemplos de cada dimensão

Fonte: Berber Sardinha e Moreira (2023)

Após a coleta das postagens de forma automatizada, o corpus foi processado no etiquetador TreeTagger para língua inglesa. Além de checagens manuais efetuadas no corpus, foram mantidos os substantivos, verbos e adjetivos, assim como as hashtags contidas nas postagens, através do uso de um script desenvolvido pelo professor orientador para realizar este trabalho. Os lemas referentes às categorias gramaticais – substantivos (incluindo os nomes próprios encontrados), verbos e

adjetivos – e hashtags presentes no corpus foram contados e organizados em planilhas em formato csv (comma separated values).

A seleção final das variáveis utilizadas na análise fatorial utilizou o critério de frequência. Para a análise fatorial, foi utilizado um conjunto inicial de variáveis lexicais, entre lemas (N = 127.418) e hashtags (N = 928). A lista das variáveis lexicais utilizadas e suas respectivos valores de carregamento na análise fatorial que contribuíram para a criação do Twitterchat Corpus encontra-se no Apêndice.

3.2. COMPOSIÇÃO DO CORPUS

O Quadro 4 traz a composição final do corpus.

Quadro 4: Composição do Twitterchat Corpus

Composição	Total
Postagens	127418
Palavras (tokens)	4.190.493
Palavras únicas (types)*	3.142.870

Fonte: O autor

No que concerne ao Twitter, textos gerados em contextos semelhantes têm a tendência de apresentar padrões linguísticos correlatos. As postagens produzidas pelos usuários do Twitter exibem um conjunto de características que definem a plataforma como um registro singular, demarcando-a de maneira única em relação a outras formas de expressão. Ao examinarmos as dinâmicas do Twitter, torna-se evidente que diversos recursos-chave exercem uma influência significativa sobre a linguagem. Cada um desses elementos impõe coerções distintas à expressão do usuário, moldando, assim, o estilo comunicativo geral da plataforma. Primordialmente, as características situacionais do Twitter desempenham um papel essencial no delineamento do uso da linguagem na plataforma. A natureza curta das postagens, uma característica distintiva da plataforma, teve sua origem no limite inicial de 140 caracteres, que refletia a adaptação da rede para mensagens de SMS.

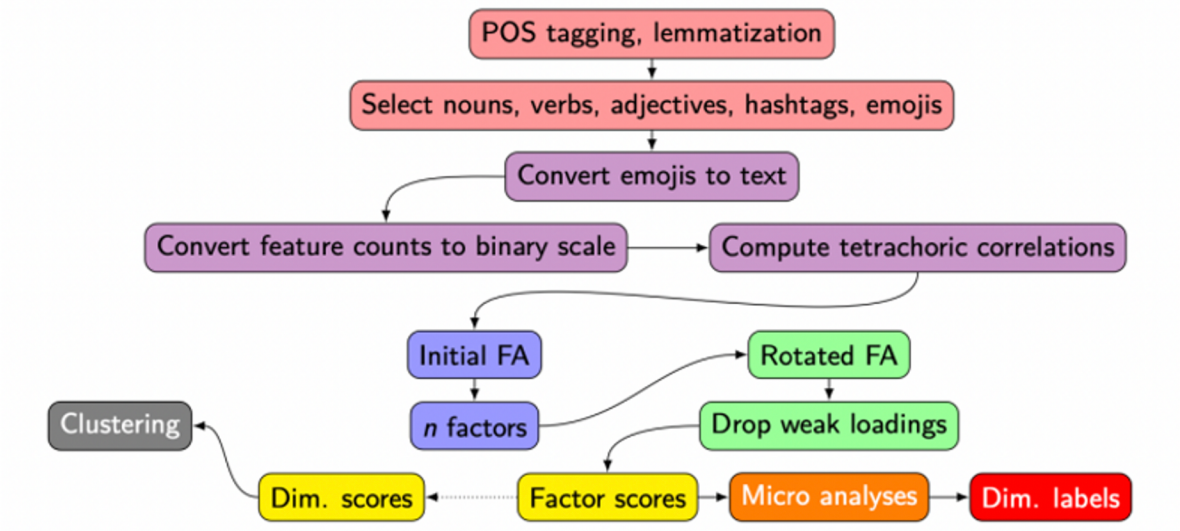
A caracterização do tuíte como registro é fundamental para a compreensão dos bate-papos educacionais no Twitter, marcados por brevidade, imediatismo, informalidade e performatividade, mas também atravessados por intertextualidade e

estratégias de autopromoção (Clarke, Brookes, & McEnery, 2022). Essas características configuram o tuíte como um registro linguístico próprio, cujos traços discursivos não apenas distinguem o Twitter de outros ambientes digitais, mas também justificam sua inclusão em análises de variação lexical-discursiva em larga escala. Nesse sentido, a presente pesquisa insere o tuíte educacional em um quadro multidimensional de coocorrências lexicais, demonstrando como tais singularidades estilísticas se articulam a propósitos comunicativos específicos no contexto da formação docente online.

3.3. ANÁLISE FATORIAL

A AMDL, conforme descrita na Figura 1, compreende três etapas fundamentais: o pré-processamento do corpus, abarcando etiquetagem, normalização e seleção das variáveis (traços linguísticos) a serem investigadas; uma análise fatorial inicial (não rotacionada), na qual o analista determina o número de fatores a serem extraídos por meio da análise do gráfico de escarpa, e por fim, uma segunda análise fatorial (rotacionada), que implica extração do número definitivo de fatores, eliminação de variáveis com baixa carga nos fatores, cálculo de escores e, por fim, nomeação das dimensões por meio de microanálises.

Figura 1: Procedimentos envolvidos em AMDL



Fonte: Berber Sardinha, Romeiro, Marcondes, Ribeiro, Oliveira et al. (2023)

Antes de computar os escores das dimensões, os valores têm de ser padronizados. A padronização é necessária para que as características de alta e baixa

frequência tenham um status similar na computação dos escores. Sem a padronização, a presença de características de alta frequência (por exemplo, substantivos e verbos) definiriam os escores da dimensão, tornando características menos frequentes, porém importantes. A extração fatorial não rotacionada gerou o scree plot, ou gráfico de sedimentação, apresentado na Figura 2.

Figura 2: Scree plot da extração factorial

Fonte: O autor

Com base na interpretação dos valores dos eigenvalues, que indicam a contribuição de cada fator para a variação total observada — conforme ilustrado no gráfico apresentado —, foram consideradas três soluções possíveis para a extração final: com quatro, sete e nove fatores. A solução com sete fatores foi a que apresentou maior interpretabilidade, conforme indicado no gráfico acima, sendo, portanto, adotada como a extração definitiva neste estudo.

A matriz fatorial resultante foi então submetida à rotação oblíqua do tipo promax, adequada à suposição de que os fatores podem estar correlacionados entre si, como é comum em dados linguísticos.

Até este ponto, os procedimentos adotados foram de natureza predominantemente quantitativa. A seguir, a pesquisa adota uma abordagem qualitativa. Nessa etapa, o pesquisador examina os contextos de uso das variáveis (isto é, os itens lexicais com altos pesos fatoriais), a partir da análise das linhas de concordância. Essa investigação qualitativa visa compreender como os padrões lexicais contribuem para a construção de sentidos discursivos em contextos específicos de uso.

A partir dessa análise, foram identificadas as dimensões predominantes no corpus e, posteriormente, atribuíram-se rótulos interpretativos a essas dimensões. Esses rótulos buscam refletir os discursos característicos emergentes, com base nas coocorrências lexicais observadas. As seções seguintes descrevem em detalhes o processo de rotulação das dimensões.

4. RESULTADOS

Neste capítulo, apresentamos e discutimos os resultados obtidos com o método explicitado acima.

4.1. APRESENTAÇÃO DOS RESULTADOS

O Quadro 5 apresenta as sete dimensões definidas e suas respectivas rotulações. Para tornar a análise as mais abrangente possível, atribuímos rótulos mais extensos de cada polo das sete dimensões.

Quadro 5: Rótulos das dimensões lexicais

Dimensão	Polo positivo	Polo negativo
Dimensão 1	Promovendo a Competência Comunicativa, o Pensamento Crítico e a Aprendizagem Prática para Aplicação em Contextos Reais e Profissionais	Engajamento Afetivo em Redes Online de Educadores: Expressão Emocional e Interação entre Pares no Discurso Profissional Colaborativo
Dimensão 2	Aprendizagem Centrada no Estudante na Educação Infantil: Integrando o Jogo, a Prática e o Desenvolvimento de Habilidades por Meio de Atividades Interativas e Acessíveis	Ampliando Oportunidades para a Aprendizagem, o Desenvolvimento Profissional e a Prática Colaborativa de Educadores
Dimensão 3	Avançando na Aprendizagem Digital, Inovação Pedagógica e Desenvolvimento Profissional para Educadores Globalmente Conectados	Pedagogias Metacognitivas e Reflexivas para a Aprendizagem Orientada à Resolução de Problemas: Estratégias Instrucionais Fundamentadas em Práticas Educacionais Focadas no Processo
Dimensão 4	Empoderamento Educacional por meio da Chamada para Ação e Inovação em Comunidades Profissionais de Aprendizagem	Curadoria Colaborativa e Disseminação de Recursos Educacionais em Comunidades de Aprendizagem Digitais em Rede
Dimensão 5	Promoção de Reconhecimento, Engajamento e Oportunidades para Motivação e Participação Ativa nas Comunidades Educacionais	Promoção do Diálogo Colaborativo e das Práticas Reflexivas na Educação: Potencializando o

		Engajamento e a Aprendizagem Interativa
Dimensão 6	Comunicação Profissional de Apoio em Redes Educacionais: (Expressões de) Incentivo, Reconhecimento e Propósito Compartilhado	A Co-construção do Conhecimento por Meio do Diálogo Colaborativo e do Engajamento Criativo
Dimensão 7	Promovendo a Excelência Educacional por meio das Interações de Coaching, Apoio Instrucional e da Integração de Práticas de Feedback e Reflexão	Cultivando a Liderança Jovem, o Desenvolvimento Profissional e a Diversidade por meio de Práticas Educacionais Criativas e Inovadoras

Fonte: Elaborado pelo autor

DIMENSÃO 1

Polo positivo: Promovendo a Competência Comunicativa, o Pensamento Crítico e a Aprendizagem Prática para Aplicação em Contextos Reais e Profissionais

Exemplo 1
A1 Gives Ss key info while also engaging them in **critical thinking** & **concept application**, placing **learning on Ss through interaction** #dlhisd

Exemplo 2
This model assumes that four factors prevailed: cognitive skills, **critical thinking**, **problem solving**, they increasingly interact and develop a high-level synthesis of single-subject designs. #dojochatEU

Polo negativo: Engajamento Afetivo em Redes Online de Educadores: Expressão Emocional e Interação entre Pares na Discurso Profissional Colaborativo

Exemplo 3

@jenorr This is **amazingly beautiful**. Sometimes we forget the strength in owning weakness. Thank you for **sharing** so **openly**. #PPL1chat

Exemplo 4

Thank you @iamsturchie for hosting these chats! Very **powerful**! I **appreciate** all of the @CESTinyTigers who participated too! Ready for 2018-2019! #cisdtigerchat

DIMENSÃO 2

Polo positivo: Aprendizagem Interativa Centrada no Estudante para Crianças: Enfatizando o Jogo, a Prática e o Desenvolvimento de Habilidades por Meio de Atividades Simples e Envolventes

Exemplo 5

You can learn **content without learning skills**. How are we **preparing our kids for college**. @Edunautics #edcampsalinas

Exemplo 6

5th graders participated in a writing **carousel activity** to **rock their writing**. Students got up and moved around in **collaborate groups** to build a piece of original writing. #RockOutMCPS

Polo negativo: Ampliando Oportunidades para a Aprendizagem, o Desenvolvimento Profissional e a Prática Colaborativa de Educadores

Exemplo 7

Join NCASCD today! District members are **eligible** to **win** #NCASCD Awards, to **present** at conferences and online events, to **receive** reduced registration rates and much more! #Networking #PeerSupport

Exemplo 8

Apple opens registration for **professional learning** virtual conferences with iPad - 9to5Mac @AppleEDU #AppleEDUchat

DIMENSÃO 3

Polo positivo: Avançando na Aprendizagem Digital, Inovação Pedagógica e Desenvolvimento Profissional para Educadores Globalmente Conectados

Exemplo 9

@iconsproject @P21centskills featured in today's #web20wednesday **highlighting global awareness**

Exemplo 10

Digital learning can help us close the global education gap World Economic Forum

Polo negativo: Pedagogias Metacognitivas e Reflexivas para a Aprendizagem Orientada à Resolução de Problemas: Estratégias Instrucionais Fundamentadas em Práticas Educacionais Focadas no Processo

Exemplo 11

Was definitely **reminded** that I **know** stuff that is valuable to the **training creation process.** Sometimes that imposter syndrome hits hard #lrnchat

Exemplo 12

@evernote allows me to **reflect** via voice memos, collecting pictures, etc. it's incredibly **useful** and **empowering** and can happen on the run with the app. #lcpsedchat

DIMENSÃO 4

Polo positivo: Empoderamento Educacional por meio da Chamada para Ação e Inovação em Comunidades Profissionais de Aprendizagem

Exemplo 13

Proud to join hands with @DiscoveryEd Cha-Ching #MoneySmartKids to celebrate National #SummerLearningWeek **empowering** youth through high-quality financial literacy education in elementary school. Discover this and more at https://t.co/JYqOxh5ONu #DiscoverSummer

Exemplo 14

A5: Talk with school administrators & express need and desire to change pattern to **engage** and **empower** Ss #techcoachBC

Polo negativo: Curadoria Colaborativa e Disseminação de Recursos Educacionais em Comunidades de Aprendizagem Digitais em Rede

Exemplo 15

We've some **excellent, free online PD options** for school teachers and librarians this term: *Growing and shaping your school library **collection** (online course starting 23 May) *Inclusive school libraries for LGBTIQA+ students.

Exemplo 16

Hey #DeafEdAcademics #DeafTwitter, we are **updating** our #DeafEd lending library **collection**. What current texts are you using and would **recommend**? TIA

DIMENSÃO 5

Polo positivo: Promoção de Reconhecimento, Engajamento e Oportunidades para Motivação e Participação Ativa nas Comunidades Educacionais

Exemplo 17

Introducing the 2020 Golden Archer **Award winner**! WI students **love** @DrewDaywalt's The Legend of Rock Paper Scissors! #wemta #goldenarcheraward

Exemplo 18

We ❤️ #educators! **Congratulations** to our 2022 @SpaceStnExplore **award winners** - **Exceptional** Educators Becky Busby & Mary Vaughn & our 1st ever Tony So **Excellence** in Education award winner, Lauren Parker! #thankateacher #STEM Learn about them here:

Polo negativo: Promoção do Diálogo Colaborativo e das Práticas Reflexivas na Educação: Potencializando o Engajamento e a Aprendizagem Interativa

Exemplo 19

Loving this conversation and the **sharing** of ideas! #EdcampEVA

Exemplo 20

Let our hands-on, **interactive** workshops help you build **meaningful**, **positive** student **relationships**. #profdev #onlinelearning #CollaborativePD

DIMENSÃO 6

Polo positivo: Comunicação Profissional de Apoio em Redes Educacionais: Expressões de Incentivo, Reconhecimento e Propósito Compartilhado

Exemplo 21

Attention all #**EduHeroes**! We are gearing up for #EdCampJoCo on September 15th. We love to have Ed Gifts on hand for our attendees as door prizes and gift bags. We would love to **promote your PD books, children's lit, gadgets**, etc. Send me a direct message to reach out!

Exemplo 22

States *can* **improve #edschools**--it's "a matter of **discipline**," @ArthurELevine tells @TeacherBeat

Polo negativo: A Coconstrução do Conhecimento por Meio do Diálogo Colaborativo e do Engajamento Criativo

Exemplo 23

Resilience, risk, **collaboration** and creativity - how do mobile devices promote **independent thinking** & learning. Find out tomorrow #mite2017

Exemplo 24

"@spinedu: @LeneJensbyLange thx f **awesome insight** f this piece of **creative** learning spaces http://t.co/NR4EO39uJi" #eddesignchat #skolchatt

DIMENSÃO 7

Polo positivo: Promovendo a Excelência Educacional por meio das Interações de Coaching, Apoio Instrucional e da Integração de Práticas de Feedback e Reflexão

Exemplo 25

Your **Coaching** Toolbox—**resources**, **tips**, and **reflections** for instructional coaches, by instructional coaches. #educoach #educoachOC

Exemplo 26

Instructional **coaching** doesn't solely need to mean **coach** / teacher. Learn why opening up the definition to **include peer-to-peer feedback** is also important.

@DrMichaelMoody @danielson_group #PrinLeaderChat #EdLeaders

Polo negativo: Cultivando a Liderança Jovem, o Desenvolvimento Profissional e a Diversidade por meio de Práticas Educacionais Criativas e Inovadoras

Exemplo 27

On #YouthDay we celebrate #SDGYOUTHACTION & call on leaders to **realize young women's** roles in **achieving progress**! #whatwomenwant #IYD2017

Exemplo 28 Introducing @BenDavisHS class of 2016 Grad, @IamAllyJ Ally Johnson, BFA, MFA!! We are so **proud** of this **young woman**. #Classof2022 #wearewayne #straightoutofthewestside #blackandhooded

4.2. DISCUSSÃO DOS RESULTADOS

As dimensões acima representam posicionamentos distintos, porém não necessariamente contrastivos, sobre a educação:

A docência é vista como prática humanizada ao passo que o professor deve atuar como formador de habilidades para o mundo além dos muros da escola (Dimensão 1);

O educador é um aprendiz em constante evolução ao passo que o professor deve ser agente inovador e constantemente atualizado (Dimensão 2);

O conhecimento se forma através da reflexão e da resolucção de problemas ao passo que o conhecimento se constrói por meio da troca de experiências e informações (Dimensão 3);

Ensinar se apoia na escuta ativa do docente ao passo o ato de aprender depende quase que exclusivamente do protagonismo do aluno (Dimensão 4);

Os professores como comunidades de prática exercem papel transformador ao passo que o pertencimento a essas comunidades mantem o docente motivado (Dimensão 5);

O conhecimento é co-construído e colaborativo ao passo que o professor precisa ser incentivado e seu trabalho reconhecido (Dimensão 6);

As comunidades online de professores são espaços de acolhimento e apoio mútuo ao passo que a escola deve ser espaço de excelência, em que se forma líderes (Dimensão 7).

Esses discursos indicam que os bate-papos educacionas possuem um repertório definido, que reflete diferentes posicionamentos e práticas no campo educacional. O Quadro 6 apresenta as representações acerca de docência refletidas nas dimensões identificadas.

Quadro 6 – Síntese dos discursos e das representações

Dimensão	Discursos	Representações de docência
Dimensão 1 – Polo positivo	Promovendo a Competência Comunicativa, o Pensamento Crítico e a Aprendizagem Prática para Aplicação em Contextos Reais e Profissionais	O professor como formador de habilidades para o mundo real.
Dimensão 1 – Polo negativo	Engajamento Afetivo em Redes Online de Educadores: Expressão Emocional e Interação entre Pares no Discurso Profissional Colaborativo	A docência como prática sensível e comunitária.
Dimensão 2 – Polo positivo	Aprendizagem Centrada no Estudante na Educação Infantil: Integrando o Jogo, a Prática e o Desenvolvimento de Habilidades por Meio de Atividades Interativas e Acessíveis	A criança como protagonista da própria aprendizagem.
Dimensão 2 – Polo negativo	Ampliando Oportunidades para a Aprendizagem, o Desenvolvimento Profissional e a Prática Colaborativa de Educadores	O educador como aprendiz em constante evolução.
Dimensão 3 – Polo positivo	Avançando na Aprendizagem Digital, Inovação Pedagógica e Desenvolvimento Profissional para	O professor como agente inovador e conectado.

	Educadores Globalmente Conectados	
Dimensão 3 – Polo negativo	Pedagogias Metacognitivas e Reflexivas para a Aprendizagem Orientada à Resolução de Problemas: Estratégias Instrucionais Fundamentadas em Práticas Educacionais Focadas no Processo	Aprender é refletir e resolver problemas com autonomia.
Dimensão 4 – Polo positivo	Empoderamento Educacional por meio da Chamada para Ação e Inovação em Comunidades Profissionais de Aprendizagem	Comunidades de docentes são promovem mudança.
Dimensão 4 – Polo negativo	Curadoria Colaborativa e Disseminação de Recursos Educacionais em Comunidades de Aprendizagem Digitais em Rede	O conhecimento se constrói por meio da troca livre entre educadores.
Dimensão 5 – Polo positivo	Promoção de Reconhecimento, Engajamento e Oportunidades para Motivação e Participação Ativa nas Comunidades Educacionais	A sensação de pertencimento impulsiona a participação.
Dimensão 5 – Polo negativo	Promoção do Diálogo Colaborativo e das Práticas Reflexivas na Educação: Potencializando o Engajamento e a Aprendizagem Interativa	Ensinar é escutar, dialogar e construir juntos.
Dimensão 6 – Polo positivo	Comunicação Profissional de Apoio em Redes Educacionais: (Expressões de) Incentivo, Reconhecimento e Propósito Compartilhado	O espaço educacional é lugar de incentivo e cuidado mútuo.
Dimensão 6 – Polo negativo	A Coconstrução do Conhecimento por Meio do Diálogo Colaborativo e do Engajamento Criativo	Aprender é criar com os outros.
Dimensão 7 – Polo positivo	Promovendo a Excelência Educacional por meio das Interações de Coaching, Apoio Instrucional e da	A qualidade educacional vem do acompanhamento reflexivo

	Integração de Práticas de Feedback e Reflexão	
Dimensão 7 – Polo negativo	Cultivando a Liderança Jovem, o Desenvolvimento Profissional e a Diversidade por meio de Práticas Educacionais Criativas e Inovadoras	A escola forma líderes diversos e atuantes.

Fonte: Elaborada pelo autor

As representações sociais que emergem das dimensões lexicais extraídas de bate-papos educacionais no Twitter constituem um ecossistema discursivo complexo, que tanto revela pontos de convergência quanto marca áreas de tensão e contraste. Em vez de configurar categorias estanques, tais representações interagem entre si em um contínuo dinâmico de significações sociais, discursivas e ideológicas. A identificação dessas aproximações e distanciamentos não apenas fundamenta uma análise qualitativa e quantitativa robusta de um grande volume de dados, como também permite compreender as formas pelas quais sentidos compartilhados sobre educação, docência e aprendizagem são produzidos e compartilhados em ambientes digitais.

Em suma, as representações sociais extraídas dos dados revelam não apenas diferentes perspectivas sobre o papel do professor, do aluno, da escola e da rede, mas também um campo discursivo em disputa, onde valores como colaboração, autonomia, inovação e cuidado são reconfigurados conforme os contextos e os interlocutores. A ADML mostra-se, assim, uma ferramenta fundamental para captar tais representações, ao evidenciar os posicionamentos ideológicos que sustentam a linguagem dos bate-papos educacionais no Twitter.

5. CONCLUSÃO

Como pesquisadores de Linguística de Corpus, encontramo-nos inseridos na visão que entende os sentidos da língua não por meio de regras, mas por meio de padrões de uso. Entendemos que o uso da língua é mediado pelo registro, pelos discursos e por outras variáveis que resultam nos diferentes usos, mais do que na sua homogeneidade. Por isso, o estudo de língua em uso classifica como a condição primeira da linguagem a de ser uma incompletude, em que tanto os sujeitos, tanto os discursos como os sentidos estão em constante mudança.

Esta pesquisa explora um fenômeno contemporâneo, ao abordar um objeto de estudo original, investigando a linguagem dos bate-papos educacionais no Twitter, um fenômeno relativamente recente e pouco explorado pela Linguística de Corpus.

A identificação dos discursos e posicionamentos ideológicos que permearam os bate-papos educacionais no Twitter contribui para a compreensão do impacto das novas tecnologias na linguagem e na comunicação. Estes bate-papos educacionais, que já foram um centro vibrante de aprendizagem profissional e networking entre educadores, sofreram um declínio significativo no Twitter (agora X) devido às mudanças na plataforma e à dinâmica mutável da comunidade. Entretanto, o conceito de bate-papos com educadores em tempo real e orientados por hashtag migrou para outras plataformas sociais, entre elas o BlueSky, o Mastodon e o Thread.

Esta análise revelou sete dimensões principais de variação discursiva, que, por meio de seus polos positivos e negativos, permitiram observar a coexistência de perspectivas complementares, por vezes tensionadas, que configuram a prática discursiva dos professores em rede. De um lado, evidenciou-se a valorização da afetividade, do pertencimento e da prática docente colaborativa; de outro, destacou-se a ênfase em competências técnicas, inovação pedagógica, curadoria de recursos e excelência instrucional. Esse leque de discursos mostra que os participantes dos bate-papos educacionais não se limitam a reproduzir temas curriculares ou conteúdos programáticos, mas constroem coletivamente representações sobre o ser professor, o ensinar e o aprender em contextos digitais conectados.

As dimensões identificadas sintetizam aspectos como: (i) a construção de comunidades afetivas e de apoio profissional; (ii) a valorização da aprendizagem ativa e reflexiva; (iii) o compartilhamento de recursos e experiências; (iv) a promoção da liderança docente e da diversidade; (v) o reconhecimento da formação como processo contínuo, situado e coconstruído. Assim, o Twitterchat Corpus revelou-se um território discursivo multifacetado, onde valores, crenças, práticas e ideologias educacionais são constantemente (re)negociados.

Em termos interpretativos, os dados apontam para a emergência de um novo modelo de desenvolvimento profissional docente — horizontal, em rede, baseado na colaboração e na reciprocidade. Os bate-papos educacionais, embora informais e autogeridos, configuram-se como ecossistemas discursivos relevantes para a constituição de comunidades de prática que transcendem as fronteiras institucionais e geográficas da escola. Ao engajarem-se nessas interações, os professores compartilham suas vozes, exercem sua agência e refletem criticamente sobre sua prática.

Diante disso, algumas recomendações podem ser formuladas. Primeiramente, destaca-se a pertinência de integrar os achados desta pesquisa a propostas de formação continuada de professores, especialmente aquelas que promovam o uso reflexivo e intencional das redes sociais como espaços de desenvolvimento profissional. Sugere-se, também, a criação de materiais didáticos e formativos que utilizem os discursos mapeados como ponto de partida para a discussão de crenças docentes, metodologias pedagógicas e práticas colaborativas. Por fim, recomenda-se a ampliação das investigações para outros contextos linguísticos e educacionais, com o intuito de compreender como diferentes comunidades de educadores negociam seus saberes e identidades discursivas em rede.

Em síntese, este trabalho demonstrou que os bate-papos educacionais no Twitter/X constituem práticas discursivas ricas em significados e potentes em termos formativos. Por meio da análise multidimensional lexical de um corpus representativo desses encontros, foi possível delinear sete grandes dimensões discursivas que refletem a complexidade, a diversidade e a intencionalidade dos usos da linguagem entre professores em rede. Mais do que um simples espaço de socialização, os bate-papos educacionais configuram-se como espaços de construção de conhecimento,

de identidade e de pertencimento profissional. A Linguística de Corpus, ao fornecer ferramentas para a descrição empírica e rigorosa desses fenômenos, afirma-se como um caminho promissor para novas investigações sobre a linguagem em uso nos ambientes educacionais digitais.

Os resultados possibilitaram a emergência de temas caros à uma leitura crítica sobre o conceito de educação, já que alguns polos nas dimensões se voltam para a face mercadológica da educação formal enquanto outros tratam da capacidade de solucionar problemas e resolver conflitos que surgem naturalmente em torno das diferentes teorias de ensino e aprendizagem. As próprias hashtags, além de exercer o papel de localizar e organizar as postagens sobre assuntos de educação, denotam os propósitos dos grupos organizadores dos bate-papos e pontuam a inserção de discursos diferentes, às vezes opostos, em comunidades discursivas diversas, sejam elas ligadas ao ensino, à formação e ao desenvolvimento do professor, ao empoderamento do aluno, à escola como espaço de aprendizagem por excelência. Vale notar que adicionar uma hashtag a um tuíte faz com que aquele discurso seja ligado, por esse símbolo, a outros usuários e outras comunidades que fazem uso do mesmo sinal e assunto, o que pode gerar sentidos bem diferentes.

A pesquisa encontrou 14 sentidos ou entendimentos diferentes para a educação, os quais podem ser vistos como metafóricos – educação como processo de desenvolvimento e rede de conexões; educação como habilidades e interatividade; educação como bem social e prática; educação como transformação e estrutura; educação como desafio e transformação; educação como impacto e celeiro de criatividade; educação como ensino direcionado e avaliação.

Não cabe, assim, a afirmação de que exista um “verdadeiro” sentido para a palavra educação se a análise fatorial apontou para esses 14 efeitos de sentido associados. Não se pode afirmar que um sentindo é mais real do que outro sem aplicarmos juízos de valor, o que envolve as conclusões e os próprios preceitos da pesquisa. Não há nenhum mérito ou “superioridade” de sentido se o discurso faz parte do polo positivo ou negativo. Os polos são indicativos de que certas palavras tendem a não circular nos mesmos discursos que outras. Tratam-se de dados concretos, obtidos pela análise fatorial, de que certas palavras se atraem na formação de certos discursos, enquanto outras tendem a se “repelir”.

Se fizermos uma síntese das dimensões, notamos uma coexistência de um discurso tradicional da educação formal, de uma capacidade empoderadora e solucionadora da educação, da característica colaborativa, comunitária e acolhedora da educação, da fala neoliberal da educação como privilégio (e consequentemente da falta de equidade) entre outros. Isso não deslegitima um discurso face a outro, mas é possível observar agrupamentos de efeitos de sentido, o que se apresenta como oportunidade para um avanço da pesquisa posteriormente. Essa coexistência representa as contradições inerentes à questão da educação, visto que interesses políticos se contrapõem à visões idealistas do papel da educação na sociedade moderna.

É possível observar pelos dados, por exemplo, que se o “discurso” educacional estivesse de fato abordando as teorias e metodologias de aprendizagem, os dados mostrariam só um discurso, reduzido ao efeito de sentido da educação como resultados. As evidências desveladas nas dimensões, todavia, apontam que há certos efeitos de sentido dos discursos dos usuários incorporando a “educação” a questões além do contexto da sala de aula.

Em suma, a pesquisa joga luz sobre os usos educacionais de redes sociais por um lado e os fundamentos epistemológicos da LC por outro. Apresenta os bate-papos educacionais como ecossistemas discursivos com potencial formativo ao se propor a dialogar com os desafios da formação continuada docente. Assim, a tese pretende ter contribuido para o avanço do conhecimento na Linguística ao tratar um "novo registro linguístico": os bate-papos educacionais.

REFERÊNCIAS BIBLIOGRÁFICAS

Aydin, S. (2014). The effect of educational microblogging tool twitter on the students' academic success and satisfaction. *Procedia - Social and Behavioral Sciences*, *143*, 470-474.

Baker, P., & McEnery, T. (2015). *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Basingstoke: Palgrave Macmillan.

Berber Sardinha, T. (2000). Linguística de corpus: Histórico e problemática. *DELTA*, *16*(2), 323-367.

Berber Sardinha, T. (2004). *Linguística de Corpus*. São Paulo: Manole.

Berber Sardinha, T. (2010). Abordagem metodológica da Análise Multidimensional [Method and procedures in Multidimensional analysis]. *Gragoatá*, *29*, 107-125.

Berber Sardinha, T. (2012). Lexicogrammar. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 3365–3370). Hoboken, NJ: Wiley.

Berber Sardinha, T. (2017a). *Applied Linguistics history in TESOL Quarterly.* presented at 18th World Congress of Applied Linguistics (AILA), Rio de Janeiro, RJ, Brazil.

Berber Sardinha, T. (2017b). *A corpus-based history of Applied Linguistics.* presented at 18th World Congress of Applied Linguistics (AILA), Rio de Janeiro, RJ, Brazil.

Berber Sardinha, T. (2017c). Lexical priming and register variation. In M. Pace-Sigge & K. Patterson (Eds.), *Lexical Priming: Applications and Advances* (pp. 190-230). Amsterdam: John Benjamins.

Berber Sardinha, T. (2019). Using multi-dimensional analysis to detect representations of national culture. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 231-258). London: Bloomsbury.

Berber Sardinha, T. (2020). Lexicogrammar. In C. Chapelle (Ed.), *The Concise Encyclopedia of Applied Linguistics* (pp. 701-705). Hoboken, NJ: Wiley.

Berber Sardinha, T. (2021). *Going multimodal in corpus linguistics: The case of social media.* Talk presented at International perspectives on corpus technology for language learning, University of Queensland, Australia.

Berber Sardinha, T. (2022a). Corpus linguistics and the study of social media: a case study using multi-dimensional analysis. In A. O'Keeffe & M. McCarthy (Eds.),

The Routledge Handbook of Corpus Linguistics (2nd ed., pp. 656-674). New York: Routledge.

Berber Sardinha, T. (2022b). A text typology of social media. *Register Studies*, *4*(2), 138-170.

Berber Sardinha, T., & Fitzsimmons-Doolan, S. (2025). *Lexical Multidimensional Analysis: Identifying Discourses and Ideologies*. Cambridge: Cambridge University Press.

Berber Sardinha, T., Kauffmann, C., & Acunzo, C. M. (2014). Dimensions of register variation in Brazilian Portuguese. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber* (pp. 35-80). Amsterdam/Philadelphia, PA: John Benjamins.

Berber Sardinha, T., Romeiro, Y., Marcondes, L. N. L., Ribeiro, N. L., Oliveira, M., Tavares Pinto, P., Araujo, R. F. d., Oliva, K., Schmitz de Almeida Lopes, T., Dutra, D., Whiteman, M., Brogini, A., Hughes, S., Soares da Silva, C., Chiarelo Boldarine, A., Zamboni Milanez, A., Silva, E., Nunes Delfino, M. C., Ferraz Escarbelin, L., . . . Ferreira Lopes, M. (2023). *The coronavirus infodemic: A multidimensional, discourse-based perspective.* Panel presentation presented at The Twelfth International Corpus Linguistics Conference (CL 2023), Lancaster University.

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2014). *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber*. Amsterdam/Philadelphia, PA: John Benjamins.

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2019). *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics - Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., Egbert, J., Keller, D., & Wizner, S. (2021). Towards a taxonomy of conversational discourse types: An empirical corpus-based analysis. *Journal of Pragmatics*, *171*, 20–35.

Britt, V. G., & Paulus, T. (2016). Beyond the four walls of my building: A case study of #Edchat as a community of practice. *American Journal of Distance Education*, *30*(1), 48-59.

Brogini, A. A. (2022). *Representações contemporâneas da sustentabilidade: Uma análise multidimensional lexical discursiva como contribuição para o portal multimodal/multilíngue para avanço da ciência aberta nas humanidades [Contemporary representations of sustainability: A multidimensional lexical discursive analysis as a contribution to the multimodal/multilingual portal for the advancement of open science in the humanities]*. MA Thesis. Sao Paulo, Brazil: Graduate Program in Applied Linguistics and Language Studies (LAEL), Pontifical Catholic University of Sao Paulo (PUCSP).

Burr, V. (1995). *An Introduction to Social Constructionism*. London: Routledge.

Cantos Gómez, P. (2013). *Statistical methods in language and linguistic research*. Sheffield: Equinox.

Carpenter, J. P., & Krutka, D. G. (2014). How and why educators use Twitter: A survey of the field. *Journal of Research on Technology in Education*, *46*(4), 414-434.

Clarke, I., Brookes, G., & McEnery, T. (2022). Keywords through time: Tracking changes in press discourse of Islam. *Internation Journal of Corpus Linguistics*, *27*(4), 399-427.

Delfino, M. C. N., Berber Sardinha, T., & Collentine, J. G. (2021). *Tipologia multidimensional multimodal big data da música pop em inglês*. LAEL, PUCSP.

Fillmore, C. (1992). 'Corpus linguistics' or 'computer corpus linguistics'. In J. Svartvik (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 35-60). Berlin, New York: De Gruyter.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-55. In F. R. Palmer (Ed.), *Selected Papers of J. R. Firth 1952–59* (pp. 168-205). London: Longmans.

Fonseca de Araujo, R. (2025). *Linguistic layers of Reality TV: A Multidimensional approach to American Reality TV Shows.* Paper presented at Corpus Linguistics 2025 International Conference (CL2025), Aston University, Birmingham, UK.

Friginal, E., & Hardy, J. (Eds.). (2020). *Routledge Handbook of Corpus Approaches to Discourse Analysis*. London: Routledge.

Gillings, M., Mautner, G., & Baker, P. (2023). *Corpus-assisted discourse studies*. Cambridge: Cambridge University Press.

Gray, B. (2013). More than discipline: Uncovering multi-dimensional patterns of variation in academic research articles. *Corpora*, *8*, 153-181.

Greenhalgh, S. P., & Koehler, M. J. (2017). Tweet, tweet, teach: Educators' perspectives on using Twitter for professional development. *Journal of Digital Learning in Teacher Education*, *33*(1), 20-28.

Hajer, M. (1993). Discourse coalitions and the institutionalization of practice. In *The argumentative turn in policy analysis and planning* (pp. 43-76). Durham, NC: Duke University Press.

Hall, S. (1992). *Formations of Modernity*. Cambridge: Polity Press.

Halliday, M. A. K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik* (pp. 30-43). London: Longman.

Hardy, J. A. (2015). Multi-Dimensional analysis of academic discourse. In P. Baker & T. McEnery (Eds.), *Corpora and Discourse Studies: Integrating Discourse and Corpora* (pp. 155-174). Basingstoke: Palgrave Macmillan.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kauffmann, C. (2020). *Linguística de corpus e estilo: análises multidimensional e canônica na ficção de Machado de Assis [Corpus Linguistics and style: Multi-dimensional and canonical analyses of Machado de Assis's fiction]*. Sao Paulo: LAEL, Graduate Program in Applied Linguistics, Catholic University of Sao Paulo.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 105-127). Berlin, New York: De Gruyter.

Mayer, C. (2018). *O que e como escrevemos na Web: Um estudo multidimensional de variação de registro em língua inglesa [What and how we write on the Web: A multi-dimensional study of register variation in English].* São Paulo: Graduate Program in Applied Linguistics, Sao Paulo Catholic University.

McEnery, T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Moreira, M. M. F. P. (2023). *Deus, Pátria e família: Os discursos bolsonaristas na rede social Twitter*. Relatório de Iniciação Científica. São Paulo: Pontifícia Universidade Católica de São Paulo.

Oliveira, J. M., & Carvalho, A. B. G. (2023). Comunidade de prática de redes sociais como processo de formação para professores e licenciandos. *Em Teia: Revista de Educação Matemática e Tecnológica Iberoamericana*, *14*(1), 11.

Romeiro, Y. (2020). *A linguagem verbal das artes visuais: uma análise multidimensional do discurso sobre a fotografia de Sally Mann*. São Paulo: LAEL, PUCSP.

Sanchez, A. (1995). Definicion e historia de los corpus. In A. Sanchez, R. Sarmiento, P. Cantos, & J. Simon (Eds.), *CUMBRE - Corpus Linguistico de Espanol Contemporaneo* (pp. 7-24). Madrid: SGEL.

Sarmento, S. (2009). Corpora e ensino de línguas. In *Enciclopédia do português do Brasil*. São Paulo: Universidade de São Paulo.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford, New York: Oxford University Press.

Sinclair, J. M. (2004). Trust the Text: Language, Corpus and Discourse. vi, 212.

Sinclair, J. M., & Jones, S. (1974/1996). English lexical collocations: A study in computational linguistics. In J. A. Foley (Ed.), *J M Sinclair on Lexis and Lexicography* (pp. 22-68). Singapore: UniPress.

Stubbs, M. (1995). Collocations and Cultural Connotations of Common Words. *Linguistics and Education*, *7*, 379-390.

Stubbs, M., Baker, M., & Tognini-Bonelli, G. F. a. E. (1993). British traditions in text analysis - From Firth to Sinclair. In *Text and technology: In honour of John Sinclair* (pp. 1-35). Philadelphia/Amsterdam: John Benjamins.

Veiga, A. T. (2021). *As dimensões da fé: Sete religiões mundiais em análise multidimensional lexical*. Tese de doutorado. São Paulo: LAEL, PUCSP.

Whiteman, M. (2024). *Os discursos em torno do movimento antivacina no Brasil durante a pandemia de COVID-19*. MA Thesis. Sao Paulo: LAEL, Graduate Program in Applied Linguistics and Language Studies, Pontifícia Universidade Católica de São Paulo. https://repositorio.pucsp.br/jspui/handle/handle/42478

Williams, T. (2025). Void left by decline of academic Twitter ‘will be hard to fill’. Retrieved 21 Jun 2025, from https://www.timeshighereducation.com/news/void-left-decline-academic-twitter-will-be-hard-fill

Xing, W., & Gao, F. (2018). Exploring the relationship between online discourse and commitment in Twitter professional learning communities. *Computers & Education*, *126*, 388-398.

Zappavigna, M. (2017). Twitter. *Pragmatics of Social Media*, *11*, 201.