

CALIENT: CORPUS DE APRENDIZES DA LÍNGUA INGLESA DO ENSINO MÉDIO TÉCNICO

CALIENT: Learner Corpus of English Language from Technical High School Education

DOI: 10.14393/LL63-v41S-2026-06

Lucas Renato dos Santos Murta^{1*}

Shirlene Bemfica de Oliveira^{2**}

RESUMO: Este artigo apresenta o corpus CALIENT, um conjunto de dados composto por transcrições de discurso oral e textos escritos em inglês, produzidos por alunos do ensino médio técnico de um Instituto Federal. A coleta de dados é integrada a um projeto pedagógico, de pesquisa e extensão que, fundamentado em teorias socioconstrutivistas e de letramento crítico (Vygotsky, 1978; Storch, 2005; Street, 2014), incentiva o uso da língua inglesa através de tarefas de resolução de problemas, discussões sobre temas transversais e escrita colaborativa com recursos tecnológicos. O projeto visa desenvolver o pensamento crítico, a interlíngua e a autonomia dos alunos, encorajando-os a se engajar em atividades sociais. O CALIENT se destaca por fornecer registros linguísticos de aprendizes de inglês em diferentes níveis. Este trabalho detalha a organização do corpus e a seleção de metadados e apresenta uma análise baseada na Análise Multidimensional (Biber, 1988; Biber; Conrad, 2009; Sardinha, 2000; Aluísio; Almeida, 2021; González, 2007).

PALAVRAS-CHAVE: Corpus de aprendizes. Ensino médio técnico. Língua inglesa. Ensino de línguas. Análise multidimensional.

ABSTRACT: This article introduces the CALIENT Corpus, a dataset comprising transcriptions of oral discourse and written texts in English, produced by technical high school students from a Federal Institute. Data collection is integrated into a pedagogical, research, and extension project that, grounded in socio-constructivist and critical literacy theories (Vygotsky, 1978; Storch, 2005; Street, 2014), encourages the English language use through problem-solving tasks, discussions on transversal themes, and collaborative writing with technological resources. The project aims to develop students' critical thinking, interlanguage, and autonomy, encouraging them to engage in social activities. CALIENT stands out by providing linguistic records of English learners at different proficiency levels. This paper details the corpus organization, metadata selection, and presents an analysis based on Multidimensional Analysis (Biber, 1988; Biber; Conrad, 2009; Sardinha, 2000; Aluísio; Almeida, 2021; González, 2007).

KEYWORDS: Learner corpus. Technical high school education. English language. Language teaching. Multidimensional analysis.

^{1*} Graduação em Letras – Licenciatura em língua inglesa. Universidade Federal de Minas Gerais. ORCID: 0009-0005-3640-7616. E-mail para contato: lucasrenatomurta(AT)gmail.com.

^{2**} Doutorado em Estudos Linguísticos. Instituto Federal de Minas Gerais / Campus Ouro Preto. ORCID: 0000-0003-1435-1728. E-mail para contato: shirlene.o(AT)ifmg.edu.br.

1 Introdução

Neste artigo, apresentamos o Corpus de Aprendizizes da Língua Inglesa do Ensino Médio Técnico (CALIEMT) que vem sendo compilado e atualizado anualmente com o apoio de professores e estudantes brasileiros. Esse corpus inovador contém textos escritos de diversos registros e transcrições de aulas e sessões colaborativas para produções escritas. Ele foi projetado para ser disponibilizado para pesquisadores em Educação e Linguística, especialmente aqueles focados no ensino de inglês como língua adicional e na criação de corpora eletrônicos (Aluísio; Almeida, 2021; Dutra et al., 2022; Kaatari et al., 2024; Wang et al., 2025). Adicionalmente, o CALIEMT fornece metadados detalhados sobre os estudantes, permitindo o estudo empírico das relações entre propriedades linguísticas, variáveis extralinguísticas e de aprendizagem.

Estudos com corpora de aprendizes do Ensino Médio Técnico ainda são incipientes no contexto brasileiro, e os corpora existentes internacionalmente geralmente focam estudantes universitários ou de alta proficiência. No contexto nacional, embora o Corpus do Inglês para Fins Acadêmicos - CorIFA (Dutra et al., 2022) e o *Brazilian Sub-Corpus of the International Corpus of Learner English - Br-ICLE* (Sardinha, 2001) contribuam para a área, ainda há uma escassez de corpora abrangentes de inglês escrito por jovens aprendizes do ensino médio técnico profissional no contexto de sala de aula brasileira (Aluísio; Almeida, 2021; Dutra et al., 2022). O CALIEMT ajuda a preencher a lacuna de dados empíricos sobre a linguagem produzida por adolescentes brasileiros desse nível educacional. Ao disponibilizar textos orais e escritos desse público, o estudo permite mapear a interlíngua de base escolar pública e fundamentar materiais didáticos específicos para essa faixa etária.

Alguns exemplos de corpora internacionais são: *The International Corpus of Learner English – ICLE* (Granger et al., 2020), *Louvain International Database of Spoken English Interlanguage – LINDSEI* (Gilquin et al., 2010), *Michigan Corpus of Upper-level students papers - MICUSP* (University of Michigan, 2009), *The British Academic Written English - BAWE* (Nesi et al., 2008-2009) e *Locness Louvain Corpus of Native English Essays - LOCNESS* (Granger, 1998). Embora os corpora nacionais e internacionais apresentados anteriormente mapeiem o nível universitário e acadêmico, eles não cobrem produções de adolescentes de ensino médio integrados à educação profissional técnica, criando uma lacuna metodológica sobre como

esse segmento específico desenvolve sua interlíngua. Os dados disponibilizados pelo CALIEMT podem ajudar os pesquisadores a investigar a variação linguística presente nas produções escritas de alunos de diversos níveis de proficiência em textos de registros variados “a fim de entender as escolhas dos usuários sob perspectivas de dados transversais ou longitudinais” (Dutra et al., 2022, p. 5).

O CALIEMT fornece informações detalhadas sobre os aprendizes e suas produções, e esses metadados são importantes para possibilitar o estudo empírico das relações entre as propriedades linguísticas dos textos e outras variáveis extralinguísticas que podem interferir nas produções orais e escritas na língua-alvo, um campo de pesquisa que ainda carece de maior investigação, conforme Kaatari et al. (2024) e Wang et al. (2025). O uso de corpora de aprendizes, como o CALIEMT, pode informar currículos e possibilitar a criação de materiais pedagógicos relevantes, oferecendo exemplos autênticos da interlíngua³ dos alunos (Vinokurova et al., 2023). Neste estudo, foi utilizada a primeira versão do corpus escrito, que inclui 1091 textos (181437 palavras) e o corpus oral é composto por 50 transcrições de gravações (54175 palavras). Esse material possibilita aos pesquisadores entender as regularidades e variações da linguagem em uso, por meio da detecção e análise de padrões linguísticos (vocabulário, sintaxe, gramática e estilo) na escrita dos alunos adolescentes (Biber et al., 1998). Este artigo descreve o processo de compilação das versões oral e escrita do CALIEMT e apresenta uma descrição baseada na Análise Multidimensional⁴ de Biber (1988) e Biber e Conrad (2009).

Nosso objetivo é demonstrar como o CALIEMT pode ser utilizado para analisar o inglês de aprendizes em contexto de ensino médio técnico e aprofundar a compreensão do impacto do engajamento oral e escrito⁵, considerando diferentes contextos culturais e educacionais

³ A interlíngua pode ser entendida como a competência do aluno na língua adicional. Selinker (1972) a define como uma língua intermediária ao observar que os enunciados produzidos por um aprendiz na língua-alvo não são idênticos aos produzidos pelo falante nativo para expressar o mesmo significado. O autor estabelece a hipótese da existência de um sistema linguístico em separado, o qual se apreende quando o aprendiz tenta, em seu output, produzir uma norma na língua-alvo. Dessa forma, a interlíngua é a produção oral ou escrita que o aprendiz efetivamente alcança (Selinker, 1972, p. 214).

⁴ A análise Multidimensional será definida na seção de metodologia.

⁵ O conceito de *engagement* (engajamento) é um construto dinâmico e multidimensional que reflete a manifestação ativa da motivação do aprendiz. Esse fenômeno transcende a mera atenção passiva, demandando um investimento holístico de esforço mental, emocional e social direcionado à apropriação do novo idioma. De acordo com Sulis (2022), esse conceito é fundamentado no modelo canônico da psicologia educacional e é

(Wang et al., 2025). Dessa forma, o valor do corpus estaria mais vinculado ao processo de construção textual dos alunos, pois o CALIEMT oferece detalhes sobre como a aprendizagem de inglês realmente acontece, em vez de apenas medir o resultado da proficiência. Ele serve como uma ferramenta de análise para entender como e por que os aprendizes constroem a linguagem da maneira que fazem em diferentes situações.

2 Pressupostos teóricos

Este projeto propõe uma interface teórico-metodológica entre a Linguística de Corpus (LC) e a Linguística Aplicada (LA). Para isso, são explorados os estudos baseados em corpora, os pressupostos de ensino-aprendizagem de inglês como língua adicional, e as investigações acerca da compilação e análise de corpora e suas funcionalidades.

2.1 Linguística de corpus

A Linguística de Corpus (LC) é uma área que estuda empiricamente padrões probabilísticos da linguagem em uso, utilizando grandes corpora de dados reais analisados por computadores (Sardinha, 2004; González, 2007). Caracteriza-se por metodologias quantitativas para investigar textos em larga escala, garantindo confiabilidade e validade empírica. O foco principal da LC reside na identificação de regularidades e padrões linguísticos, incluindo a carga avaliativa de combinações de palavras e a constituição de discursos através do uso repetido e sistemático de certas palavras e construções em diferentes registros (Stubbs, 2001, p. 215; Baker, 2006, p. 13). Além de identificar padrões, a LC possui potencial heurístico, contribuindo para pesquisas em áreas como análise do discurso e de sentimentos (Baker, 2006), semântica, estilística, variação e mudança linguística (Sardinha, 2004), linguística forense, aquisição de línguas e sociolinguística (González, 2007).

estruturado em três dimensões interdependentes que tocam em questões comportamentais, como a prontidão dos alunos em participar das tarefas. O engajamento também envolve questões de cunho emocional, como, por exemplo, as reações afetivas que englobam o interesse, o entusiasmo e a segurança psicológica necessária para assumir riscos linguísticos (o erro) sem ansiedade. E, por fim, o engajamento prevê o investimento cognitivo ou intelectual e a autorregulação, evidenciados pelo esforço deliberado em processar regras complexas e praticar a autoavaliação (Sulis, 2002).

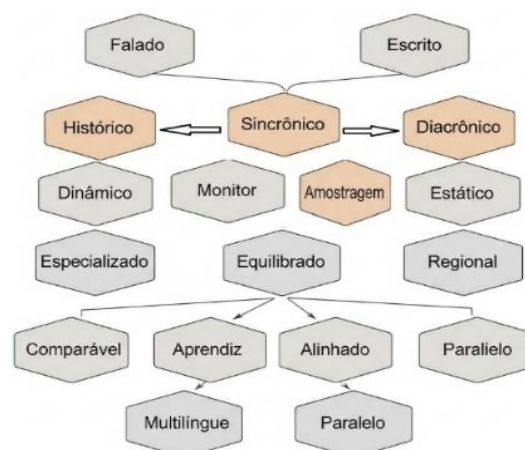
Embora muitas vezes vista como puramente quantitativa, a LC combina componentes qualitativos (estudo da linguagem em uso, interpretação humana e contextual, análise de variação e registro, análise das intenções comunicativas, relações de poder), permitindo a visualização de grandes corpora por lentes quantitativas e qualitativas. Isso envolve a coleta, organização e análise de grandes volumes de dados linguísticos autênticos para descobrir padrões gramaticais e semânticos, com análise aprofundada em seus contextos de ocorrência (Gillings et al., 2023). E essa capacidade de lidar com dados autênticos em escala é o que torna a LC uma área importante, fornecendo evidências concretas para teorias e análises.

A principal característica da LC é a observação de dados empíricos em corpora com ferramentas eletrônicas para verificar fenômenos da língua em uso (González, 2007, p. 8). Para esse tipo de análise, os pesquisadores recorrem a um corpus que é entendido como um conjunto de dados linguísticos (escritos ou orais) sistematizado, extenso, representativo de um domínio do uso da língua e processável por computador para descrição e análise (Sinclair, 1991; Sanchez, 1995, p. 8-9). A distinção crucial entre um corpus compilado com base nas premissas da LC e um simples banco de textos reside na ênfase na autenticidade⁶, representatividade⁷, balanceamento e diversidade dos textos (Sinclair, 1991; Gillings et al., 2023), guiada por critérios bem definidos para refletir as características da variedade de linguagem em foco. Resende e Maverick (2016, p. 8) oferecem uma caracterização multifacetada das tipologias de corpora, essencial para compreender sua diversidade e aplicações. Eles não se limitam a uma única forma de classificar, mas abordam diferentes critérios que ajudam a definir a natureza e o propósito de um corpus, conforme apresenta a Figura 1, abaixo:

⁶ Por textos naturais entende-se 'autênticos', isto é, aqueles que existem na linguagem e que não foram criados com o propósito de figurarem no corpus. Além disso, amplia-se a ideia de 'natural' para incluir somente aqueles textos produzidos por humanos. Desta forma está excluída a produção provinda de programas de geração de textos (Sardinha, 2000, p. 336).

⁷ O corpus é uma amostra de uma língua ou variedade linguística, contendo as principais propriedades, características e variações do todo que se pretende estudar, permitindo que as conclusões tiradas sejam válidas para a língua real e não apenas para os textos selecionados (Sardinha, 2000).

Figura 1: Tipologia dos Corpora



Fonte: Resende; Maverick (2016, p. 8)

Sardinha (2000, p. 339) afirma que essa classificação define o conteúdo e o propósito dos corpora. Ela é extensa e engloba os seguintes critérios: modo, tempo, seleção, conteúdo, autoria, disposição interna e finalidade. Em relação ao modo, a maioria dos corpora é formada por textos escritos (*ICLE*, *BAWE*, *CorIFA*, *ICLE*), embora existam os corpora orais, como o *LINDSEI*, constituídos por falas transcritas. Há também os corpora multimodais, como o *IFA Dialog Video Corpus* (Son, 2007), composto por transcrições e descrições da interação face a face, e o *Natural Media Motion-Capture Corpus* (Tillmann et al., 1995), que integra diferentes modos de comunicação, como texto, imagem e áudio/vídeo, acompanhados de transcrições, descrições e anotações gestuais. Em relação à temporalidade, os corpora podem ser sincrônicos, quando incluem textos de um período; diacrônicos, quando se referem a vários períodos; contemporâneos, em se tratando de um tempo atual; e históricos, que refletem um tempo passado.

No que diz respeito ao processo de seleção, os corpora podem ser de amostragem, monitores, dinâmicos ou orgânicos, estáticos ou equilibrados (Sardinha, 2000, p. 340). Os corpora de amostragem (*sample corpus*) são elaborados com porções de textos ou de variedades textuais, planejados para serem uma amostra finita da linguagem como um todo. Um exemplo é o *The Helsinki Corpus of English Texts* (Rissanen, 1991), um corpus diacrônico multi genérico estruturado, que inclui amostras de textos organizadas periodicamente em inglês antigo, médio e contemporâneo.

Os corpora monitores são geralmente projetados para verificar mudanças atuais em uma língua, são alimentados de forma recorrente e aumentam de tamanho gradualmente. Entretanto, segundo Sarmiento (2010, p. 91), a proporção de tipos de texto mantém-se constante, de forma que cada período possa ser comparado com o anterior. Um exemplo, segundo a autora, é o *The Bank of English - BoE* (Sinclair; Karlsson, 1993), que atualmente conta com cerca de 200 milhões de palavras.

Sardinha (2000) afirma que os corpora dinâmicos ou orgânicos qualificam os corpora monitores, pois o crescimento e a diminuição são permitidos e os corpora estáticos caracterizam os de amostragem. Nos corpora equilibrados (*balanced*), os componentes (gêneros, textos etc.) são distribuídos em quantidades semelhantes (por exemplo, mesmo número de textos por gênero). O *Brown corpus* é um exemplo, pois é focado no inglês americano contemporâneo, dividido por gêneros textuais (como notícias, ficção, religião etc.) e com anotações de classes gramaticais (PoS tagging) e lematização, servindo como um marco para estudos de frequência e padrões linguísticos (Sardinha, 2000, p. 340).

Os corpora paralelos são constituídos por textos em dois idiomas, geralmente traduções. Quando alinhados, apresentam correspondência direta entre trechos do texto original e seu equivalente traduzido, permitindo comparações entre línguas. Um exemplo é o Corpus Português-Italiano Literário - COR P IL, que é “um conjunto aberto de textos em português do Brasil alinhado com as suas traduções para o italiano, e de textos em italiano alinhados com as suas traduções para português (Zavaglia; Ferraresi, 2006).

Em relação ao conteúdo, os corpora podem ser especializados e seus textos são de tipos específicos (em geral, gêneros ou registros definidos). O *Corpus of Agrarian Sciences Scientific Articles - CorAgrarian* (Marques, 2024) é um corpus especializado de artigos científicos da área de Ciências Agrárias publicados em inglês em revistas de alto impacto (Qualis A). O corpus geral ou de referência, por outro lado, contém muitos tipos de texto, como o *British National Corpus* (BNC).

Além da classificação proposta, os corpora também podem ser classificados quanto ao tamanho (pequenos, médios, grandes), ao estágio de desenvolvimento (fechados/estáticos ou abertos/dinâmicos), ao tipo de anotação (*Part of speech tagging - POS tagging*,

lematização, sintaxe, semântica, entidades nomeadas, metadados), e à autoria - de aprendizes ou de falantes nativos da língua (Biber et al., 1998; Stubbs, 2001; Resende; Maverick, 2016).

2.2 Linguística Aplicada

Este projeto integra-se à Linguística Aplicada com foco no ensino de línguas, pois promove encontros que ativam conhecimentos prévios dos alunos que debatem e escrevem sobre temas transversais e os Objetivos de Desenvolvimento Sustentável (ODSs). O foco é dado na interação e colaboração entre os discentes que usam a língua inglesa para a construção/reconstrução de saberes em diversos registros (Storch, 2005; Vygotsky, 1987; Oliveira et al., 2018). Pesquisas sobre a produção escrita e oral de alunos são valiosas para a Educação, Linguística Aplicada e Linguística de Corpus, pois permitem o entendimento dos padrões linguísticos no uso da linguagem (Sardinha, 2004). Além disso, a compreensão dos ODSs e dos temas transversais por alunos do ensino médio técnico auxiliam na formação cidadã e na promoção de ações de justiça social. Tais estudos também podem fomentar novas metodologias de investigação e ensino de línguas em Institutos Federais e na Educação Profissional.

A abordagem heurística proposta para o ensino de inglês na escola técnica integral busca a discussão e escrita de textos em inglês em uma atmosfera de descoberta acadêmica, seguindo a máxima latina *“Docendo discimus”* – “Ensinando aos outros aprendemos” (atribuída a Sêneca em suas Epístolas Morais a Lucílio.)⁸.

A pesquisa adota uma abordagem de multiletramentos, desenvolvendo habilidades integradas (leitura, escrita, audição, fala) e a produção/análise de diversos registros textuais. O objetivo é fomentar o posicionamento crítico dos alunos sobre temas, tornando o ensino de idiomas orientado por dados e a aprendizagem centrada no aluno, por meio da descoberta de padrões linguísticos, do uso de ferramentas CALL (*Computer-Assisted Language Learning*), e do desenvolvimento do pensamento crítico e da interlíngua (Street, 2014; Mattos, 2014; Oliveira, 2017). Mais detalhes sobre a especificação do CALIEMT serão descritos na metodologia.

⁸ Do Latim *“Homines dum docente discunt”*, que significa “Os homens aprendem enquanto ensinam” (Sêneca – Epístolas Moais a Lucilio – Carta VII, seção 8).

3 Metodologia

O CALIEMT (Corpus de Aprendizes da Língua Inglesa do Ensino Médio Técnico) está sendo compilado em um Instituto Federal em Minas Gerais, com a colaboração anual de dois bolsistas do ensino médio técnico e uma pesquisadora. Os dados são provenientes de aproximadamente 850 alunos de cursos técnicos integrados (Administração, Automação Industrial, Edificações, Mineração e Metalurgia), coletados em ambientes de sala de aula presenciais e virtuais. O projeto foi aprovado pelo Comitê de Ética (CAAE: 60114922.4.0000.8507), e a inclusão dos dados no corpus foi precedida da assinatura de Termos de Consentimento Livre e Esclarecido (TCLEs) e Termos de Assentimento Livre e Esclarecido (TALEs).

A compilação e organização do CALIEMT seguem as diretrizes da Linguística de Corpus, com foco na autenticidade, representatividade, balanceamento e diversidade dos textos (Sinclair, 1991, Biber et al., 1998; Sardinha, 2004; González, 2007). O processo adota os quatro passos propostos por Resende e Maverick (2016, p. 11): seleção de textos e obtenção de permissões; desidentificação (Anexo); manipulação e limpeza dos dados; etiquetagem estrutural e nomeação de arquivos.

O CALIEMT é um corpus sincrônico composto por textos orais e escritos em inglês, produzidos por alunos do ensino médio técnico. Ele se concentra em temas transversais contemporâneos e de cunho social, refletindo a interlíngua dos aprendizes. Os dados escritos são coletados individualmente ou em pares, digitalizados e armazenados em um banco de dados próprio. Já os textos orais são gravados em encontros presenciais e virtuais em um Instituto Federal em Minas Gerais, utilizando filmadoras, e transcritos via plataforma *TurboScribe*. O corpus é atualizado anualmente e, apesar de dinâmico e representativo da interlíngua, não apresenta equilíbrio, pois a distribuição dos registros não é uniforme e alguns registros foram mais fáceis de coletar do que outros. O projeto CALIEMT articula estratégias pedagógicas voltadas à discussão de temas transversais e aos ODSs, culminando na produção de textos em inglês que contempla diversos registros, como cartas de apresentação, ensaios, texto expositivo de definição, argumentação, dissertação, poesia, narrativa e depoimento.

3.1 Metodologia de coleta e análise de dados

A pesquisa adota uma abordagem mista de coleta de dados (presencial e virtual), que consiste em gravações de aulas e textos de alunos, resultantes de discussões temáticas transversais. Em encontros semanais de 1h40, são desenvolvidas habilidades linguísticas integradas e aspectos como pronúncia e vocabulário, com uma abordagem global da língua (Cristal, 2003) que valoriza a multiculturalidade e experiências prévias dos alunos. O professor pesquisador foca o empoderamento dos jovens, promovendo a formação cidadã e justiça social via ODSs (Rajagopalan, 2003). Adicionalmente, o laboratório é usado para trabalhos em grupo, com gravações de discussões que apoiam as produções escritas. Essas atividades, focadas em Letramentos Sociais (Street, 2014), incentivam a expressão, criatividade, colaboração e resolução de problemas.

Os metadados advêm de um questionário respondido por 853 alunos participantes que também realizaram um exame de proficiência aplicado pela pesquisadora. A caracterização da proficiência desses estudantes, conforme o Quadro Comum Europeu de Referência para Línguas (QCER) (Conselho da Europa, 2001)⁹, revelou uma distribuição diversificada: 482 alunos no nível A1 (44,6%), 297 no A2 (27,2%), 192 no B1 (17,6%) e 117 no B2 (10,7%), com três alunos alcançando o nível C1 (0,3%) e nenhum C2. Demograficamente, a amostra incluiu 527 meninas (61,8%), 320 meninos (37,4%) e 6 participantes não binários (0,7%). Em relação à idade, 2 alunos tinham entre 10 e 13 anos, 270 entre 14 e 16 anos, 567 entre 17 e 20 anos, e 10 entre 21 e 24 anos. Quanto aos cursos, 160 alunos eram de Administração, 7 de Automação Industrial, 248 de Metalurgia e 434 de Mineração.

Os arquivos coletados (transcrições das gravações e textos), armazenados em um servidor local, são convertidos manualmente para o formato .txt. Em seguida, são refinados com o *Corpus Script* (Alves, 2024) e anotados com o *MAT Tagger* (Nini, 2015) para auxiliar no processo analítico exploratório. Essas ferramentas, segundo seus criadores, automatizam a identificação de traços linguísticos específicos associados aos diferentes tipos textuais. Essas

⁹ O Quadro Europeu Comum de Referência para Línguas (CEFR) é um marco internacional de descrição de competências linguísticas, classificadas em uma escala de seis níveis, do A1 (para principiantes) ao C2 (para quem já domina uma língua). Isso permite que todos aqueles que estão envolvidos no ensino de línguas e avaliação de competências linguísticas, como professores e alunos, vejam facilmente a que nível corresponde cada certificação (Conselho da Europa, 2001, p. 19). A1- Iniciante, A2 - Básico, B1 - Intermediário, B2 - Usuário independente, C1 - Proficiência operativa eficaz e C2 - Domínio pleno.

informações são vitais para o CALIEMT e para a reutilização do corpus em outros projetos, facilitando a recuperação de dados para análise. Os dados orais e escritos são analisados quanto à produção textual e construção colaborativa, empregando concordanciadores eletrônicos para identificar padrões lexicais e colocações. Os concordanciadores são programas de computador usados para listar as ocorrências dos itens estudados, o qual fica centralizado, com uma quantidade definida de contextos, tanto à esquerda como à direita, fornecendo uma visualização privilegiada do item. Eles permitem que o pesquisador empreenda uma observação mais qualitativa dos dados do corpus analisando o contexto (Viana, 2010, p. 71)

Para o recorte deste artigo, as análises foram realizadas com o apoio do *AntConc* e baseadas na Análise Multidimensional de Biber (1988) e Biber e Conrad (2009). A Análise Multidimensional (doravante AMD) é uma metodologia que possibilita a investigação e codificação textual de traços linguísticos em grandes corpora sob múltiplas perspectivas (dimensões). Para Sardinha (2004, p. 300), a AMD é definida como:

Uma abordagem para análise de corpus que usa procedimentos estatísticos, visando ao mapeamento das associações entre um conjunto variado de características dentro do corpus de estudo. Também usa procedimentos automáticos e semiautomáticos para análise do corpus, tais como etiquetagem morfossintática.

O conceito de dimensão de variação é associado ao conjunto de traços que subjazem a um corpus e fornecem as ferramentas para a identificação de padrões de coocorrências de todos os tipos de características, para a descrição mais precisa e ampliada da língua em corpora, de modo abrangente e com mais capacidade de comparação, conforme mostra o Quadro 1 abaixo.

Quadro 1 - Análise Multidimensional (AMD) de textos orais e escritos

Dimensão	Descrição	Indicadores Linguísticos
1. Informatividade vs. Envolvimento	Grau de densidade informativa vs. traços conversacionais A primeira dimensão da AMD de Biber (1988) refere-se à distinção entre linguagem espontânea, geralmente oral e	Substantivos, preposições, pronomes, verbos modais, conjunções

	voltada à interação, e linguagem mais planejada, frequentemente escrita e revisada.	
2. Narratividade	Presença de estrutura narrativa A dimensão da narratividade proposta por Biber (1988, p. 150) avalia a presença de estrutura narrativa nos textos, com foco em verbos no passado, advérbios temporais e conectores sequenciais que indicam eventos, ações e progressão cronológica.	Verbos no passado, advérbios de tempo
3. Elaboração	Estrutura sintática elaborada A terceira dimensão da AMD avalia o grau de complexidade estrutural e desenvolvimento sintático presente nos textos (Biber, 1988, p. 165). Essa dimensão se manifesta por meio do uso de conectores subordinativos, modificadores complexos, cláusulas dependentes e expressões que indicam esforço argumentativo ou explicativo.	Cláusulas dependentes, modificadores complexos
4. Persuasão/Expressão subjetiva	Opiniões, avaliação pessoal Essa dimensão avalia o grau de envolvimento subjetivo e expressividade pessoal presentes no texto (Biber, 1988, p. 170). Ela está relacionada ao uso de recursos linguísticos que revelam opinião, julgamento ou envolvimento emocional, como pronomes pessoais da 1ª pessoa, verbos modais epistêmicos e adjetivos avaliativos que expressam valor ou intensidade. Esses elementos indicam a presença de uma voz discursiva pessoal, com intenção persuasiva ou avaliativa.	Adjetivos avaliativos, pronomes da 1ª pessoa, modais epistêmicos
5. Referência abstrata vs. situacional	Grau de abstração A quinta dimensão distingue textos que fazem referência direta a tempo e espaço específicos daqueles cuja linguagem se orienta para abstrações e conceitos generalizados.	Palavras ligadas ao tempo/lugar concreto vs. termos abstratos

6. Coesão lexical	Uso de vocabulário semanticamente conectado para reforçar a progressão temática dos textos. Coesão Lexical: refere-se aos recursos lexicais que estabelecem relações semânticas entre elementos do texto, funcionando como complemento à coesão gramatical tradicional (como pronomes e conjunções). Entre os mecanismos característicos dessa dimensão, destacam-se: reiterações lexicais (repetição ou uso de sinônimos, hiperônimos e hipônimos, colocações semânticas, nominalizações e recursos estilísticos como metonímia e antonomásia).	Repetição de termos, colocações, sinonímia, nominalizações.
7. Formulaicidade ¹⁰	Presença de expressões fixas ou padrões recorrentes que facilitam a fluidez do discurso Essa dimensão avalia o uso de expressões fixas, fórmulas e padrões linguísticos recorrentes, que facilitam a fluidez e a naturalidade do discurso (Biber; Conrad, 2009).	Marcadores conversacionais, fórmulas acadêmicas.

Fonte: Adaptado de Biber (1988) e Biber e Conrad (2009)

Cada dimensão identificada na AMD corresponde a um agrupamento independente de traços linguísticos que ocorrem em conjunto e que podem ser interpretados em termos funcionais. Essa abordagem permite uma avaliação empírica do número de dimensões independentes, da relação entre funções associadas ou distintas, e da relevância relativa de cada dimensão na estrutura linguística analisada (Biber, 1988, p. 14). Para mapear as dimensões estatisticamente de acordo com a AMD de Biber (1988), os dados foram processados no MAT Tagger. Os Scores Z serão apresentados por dimensão na seção de resultados, evidenciando o contraste funcional entre as vertentes oral e escrita do CALIEMT. Os Scores Z das dimensões textuais foram calculados individualmente por arquivo através do Multidimensional Analysis Tagger (Nini, 2015), o qual utiliza como vetores de padronização as

¹⁰ Wray (2002, p. 9) conceitua a sequência formulaica como “(...) uma sequência de palavras ou de outros elementos, contínua ou descontínua, que é ou parece ser pré-fabricada, isto é, armazenada e recuperada como um todo na memória no momento do uso, em vez de ser gerada ou analisada a partir do processamento de regras gramaticais da língua”.

médias e desvios-padrão do macrocorpus de referência de Biber (1988). Posteriormente, para comparar os dados neste artigo, os coeficientes individuais foram agrupados por modalidade, calculando-se a média aritmética dos Scores z dos arquivos componentes do subcorpus escrito (N=1.091) e do subcorpus oral (N=50) de forma isolada.

4 Resultados

Na estruturação do CALIEMT, os dados e os metadados são separados. Os metadados dividem-se em duas partes distintas: a primeira parte está relacionada aos dados pessoais dos participantes, conforme apresentado na metodologia. A segunda parte abrange informações sobre os textos, conforme mostra o Quadro 2 abaixo:

Quadro 2: Composição do CALIEMT

Crítérios	corpus de língua inglesa como língua adicional
Abrangência	regional: região sudeste do Brasil
Meio	escrito e oral
Tempo	contemporâneo
Renovação	dinâmico /anual
Línguas	inglês
Produtores	professores e aprendizes de língua inglesa do Ensino Médio Técnico
Emprego	estudo
Estratégias de produção dos textos	- Produzidos individualmente resultantes de processo de reescrita pedagógica (versão preliminar e versão final) - Elaborados a partir de mapas conceituais e produção escrita colaborativa - Escritos em pares ou trios com gravação das interações - Transcrições dos encontros de conversação e de colaboração para a escrita
Gêneros / Registros	- Gravações dos encontros - Textos expositivos de definição, cartas de apresentação pessoal, mapas conceituais, ensaios, redações argumentativas, poemas, narrativas, depoimentos, transcrições de encontros de conversação e interações de discussão em pares e grupos.
Temas / ODSs	- Distúrbios alimentares - <i>Bullying</i> - Violência na sociedade moderna - Arte e sociedade - Racismo - Temas dos ODS (Objetivos de Desenvolvimento Sustentáveis) - Relações de poder - <i>Black History Month</i>

	- Notícias falsas - Positividade do corpo - Temas propostos pelo Br-ICLE (Sardinha, 2001)
--	---

Fonte: Dados do projeto

Os dados para este artigo foram coletados antes da popularização das ferramentas de inteligência artificial generativa e os alunos mencionaram os materiais de consulta e o apoio de colegas e professores para a elaboração dos textos. Os textos são produzidos a partir de estratégias pedagógicas, incluindo a reescrita, o uso de mapas conceituais e a escrita colaborativa em coautoria. Os gêneros textuais e registros englobam os Temas Contemporâneos Transversais, temas propostos pelo Br-ICLE (Sardinha, 2001)¹¹ e os temas relacionados aos ODSs, refletindo a diversidade de contextos e interesses dos participantes. A seleção dos temas para as redações do corpus é majoritariamente influenciada pelo Br-ICLE, conforme lista abaixo:

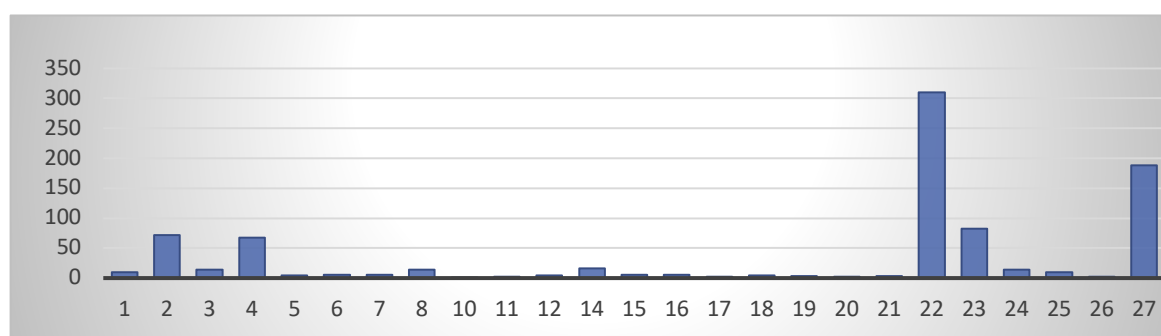
- (1) Crime does not pay
- (2) The prison system is outdated. No civilised society should punish its criminals: it should rehabilitate them
- (3) Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value
- (4) A man/woman's financial reward should be commensurate with their contribution to the society they live in
- (5) The role of censorship in Western society
- (6) Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television
- (7) All armies should consist entirely of professional soldiers: there is no value in a system of military service
- (8) The Gulf War has shown us that it is still a great thing to fight for one's country
- (9) Feminists have done more harm to the cause of women than good.
- (10) In his novel *Animal Farm*, George Orwell wrote "All men are equal: but some are more equal than others." How true is this today?
- (11) In the words of the old song "Money is the root of all evil"
- (12) In the 19th century, Victor Hugo said: "How sad it is to think that nature is calling out but humanity refuses to pay heed. "Do you think it is still true nowadays?"
- (13) Some people say that there is no longer a place for dreaming and imagination in our modern world, dominated by science, technology, and industrialisation. What is your opinion?
- (14) Eating disorder: whose fault?
- (15) Environmental development and preservation: how to reconcile these interests in conflict:
- (16) Childhood work in Brazilian society
- (17) The challenge of living with differences
- (18) Children and adolescent rights: how to face this national challenge?
- (19) Violence in Brazilian society: how to change the rules of this game?
- (20) Abortion Controversy: Pro-Life or Pro-Choice?
- (21) Bullying, crime at schools: children and teenagers have been isolating themselves, insulting each other, attacking colleagues, and exposing an alarming reality; parents and schools do not know how to deal with assaults that start earlier and earlier.

¹¹ Temas propostos pelo ICLE.

- (22) Art in modern society
- (23) The power of art and culture
- (24) Art can benefit health
- (25) Racism: does it really exist in Brazil?
- (26) Intolerance
- (27) Introducing yourself

Além disso, três temas específicos foram propostos diretamente em sala de aula (Art in modern Society – tema 22, Introducing yourself – tema 27 e The power of art and culture – tema 23). O gráfico 1 apresenta os temas mais recorrentes no corpus CALIENT.

Gráfico 1 - Temas propostos escolhidos



Fonte: dados do estudo.

A distribuição quantitativa dos textos revela a participação dos alunos nas atividades propostas em sala com os temas 22, 23 e 27. Complementarmente, os alunos tiveram a oportunidade de escolher temas para a produção de textos argumentativos e, entre as escolhas, destacam-se o tema 2, sobre o sistema prisional, e o tema 4, que trata da necessidade de igualdade salarial entre homens e mulheres. A maioria dos textos tem menos de 500 palavras¹² e se associa ao nível de proficiência dos alunos que estão no nível A1 ou A2, pois aprendizes iniciantes tendem a produzir períodos mais simples e de menor fôlego estrutural devido às limitações em sua competência lexical e sintática corrente. Em contraste, a quantidade de textos nas categorias de 500 a 1200 palavras é significativamente menor, e provavelmente está relacionada aos níveis B1 e B2. Os textos com mais de 1500 palavras

¹² Não havia número mínimo de palavras para a elaboração dos textos escritos. Os alunos foram orientados acerca das características do registro a ser produzido e os textos não foram corrigidos para serem adicionados ao corpus.

apresentam uma ocorrência mínima, decorrente do fato de apenas 0,3% da amostra situar-se no nível C1, patamar onde a fluência e a autonomia discursiva viabilizam produções textuais de maior extensão em ambiente de sala de aula.

Para fins de análise quantitativa do CALIEMT, utilizou-se o software AnConc para identificar características dos subcorpora escrito e oral, como a extensão dos subcorpora, o inventário de formas únicas (*types*) e a razão *type-token* (TTR). O corpus escrito apresenta maior extensão, com 1.091 textos, 10.847 formas e 181.437 palavras, enquanto o corpus oral conta com 50 textos, 5.104 formas e 54.175 palavras.

4.1 Análise das dimensões no CALIEMT

Com o auxílio do *AntConc* foram geradas as listas das palavras (*WordLists*) mais frequentes dos dois subcorpora, e, em seguida, foi feita uma breve análise das dimensões da AMD. Para validar estatisticamente essas frequências e mapear o perfil macrogramatical do corpus de acordo com a metodologia de Biber (1988), os dados foram processados no MAT Tagger. Os Scores Z, consolidados por dimensão, serão apresentados na Tabela 3 (seção 4.1.1), evidenciando o contraste funcional entre as vertentes oral e escrita do CALIEMT.

4.1.1 Subcorpus escrito

A análise do subcorpus escrito inicia-se pela investigação *top-down* da lista de palavras, cujo perfil de frequências absolutas e normalizadas sinaliza o cumprimento das restrições de registro da modalidade escrita e planejada (BIBER, 1988). A Tabela 1 mostra as 20 palavras mais frequentes na modalidade escrita.

Tabela 1: *WordList* do subcorpus escrito

Type	Rank	Freq	Range	NormFreq	NormRange ¹³
the	1	11194	1015	61696.346	0.930
and	2	6869	984	37858.871	0.902
of	3	5198	921	28649.063	0.844
to	4	5023	963	27684.541	0.883
is	5	4351	954	23980.776	0.874
in	6	3953	928	21787.177	0.851
a	7	3682	888	20293.545	0.814
art	8	3103	404	17102.355	0.370
that	9	2366	707	13040.339	0.648
it	10	1913	667	10543.605	0.611
for	11	1668	705	9.193.274	0.646
with	12	1492	624	8.223.240	0.572
are	13	1475	624	8.129.544	0.572
can	14	1352	565	7.451.622	0.518
this	15	1276	559	7.032.744	0.512
people	16	1210	565	6.668.982	0.518
be	17	1152	548	6.349.311	0.502
I	18	1141	234	6.288.684	0.214
have	19	1122	553	6.183.965	0.507
violence	20	1081	343	5.957.991	0.314

Fonte: Dados do estudo

Observa-se a proeminência de palavras gramaticais ou funcionais altamente recorrentes no topo do ranking (*the*, *and*, *of*, *to*, *in*, *a*), traço que atesta o encadeamento sintático complexo e a alta densidade informacional típica de textos dissertativos. A evidência de palavras gramaticais no topo do ranking é comum à maioria dos corpora. No entanto, na AMD, o peso positivo combinado de substantivos, preposições e conjunções atesta uma alta densidade informacional e o uso frequente de frases preposicionais modificadoras, o que é típico do encadeamento sintático complexo de textos formais e planejados. No domínio lexical, o surgimento do item *art* como a oitava palavra mais frequente do subcorpus (17.102,35) e de *violence* na vigésima posição (5.957,99) reflete diretamente o impacto dos

¹³ Legenda: Type (Item lexical único); Rank (Posição no ordenamento); Freq (Frequência absoluta); Range (Dispersão de arquivos); NormFreq (Frequência normalizada por 1 milhão de palavras); NormRange (Dispersão normalizada).

comandos de produção textual e dos Temas Contemporâneos Transversais mobilizados nas oficinas pedagógicas.

A saturação desses substantivos concretos e abstratos aponta para uma orientação temática voltada à discussão de conceitos sociológicos complexos. Ademais, a presença de verbos de estado estáticos no topo da lista (*is, are, be, have*) sinaliza um estilo eminentemente descritivo e expositivo, focado na definição de problemas em detrimento da ação dinâmica. Por outro lado, a emergência do pronome de primeira pessoa do singular (*I*) na posição 18 indicia traços pontuais de subjetividade e envolvimento, mapeados em gêneros específicos como as cartas de apresentação pessoal, revelando a coexistência de marcas de interlíngua e hibridismo de registros na escrita dos aprendizes.

Contudo, embora a inspeção linear da lista de palavras (Tabela 1) forneça indícios sobre a orientação lexical do subcorpus escrito, o mapeamento definitivo de sua natureza funcional exige uma modelagem multidimensional. Para detalhar como esses padrões se materializam, investigam-se a seguir os traços linguísticos das dimensões propostas por Biber (1988) identificados nas produções escritas do CALIEMT.

A Dimensão 1, associada à informatividade, se evidencia no subcorpus escrito por meio de recursos próprios da comunicação objetiva. Como grande parte dos textos são expositivos de definição e de argumentação, as palavras funcionais (*the, of, and, to, is, in, with, for, a*) e nominalizações (*education, information, solution, discussion, development, discrimination, expression*) aparecem com alta frequência, condensam ideias complexas e aumentam a abstração lexical, favorecendo descrições conceituais em vez de ações diretas. A presença dos verbos (*be* e *have*) tem foco na definição, caracterização ou exposição de fatos, enquanto substantivos concretos e abstratos (*art, people, violence, society, life, world*) revelam um vocabulário associado às propostas pedagógicas. Embora sejam verbos de alta frequência geral na língua, a presença marcante de *be* e *have* atuando como verbos principais estáticos - associada ao Score Z marcadamente negativo da Dimensão 2 (Narratividade, -1,84) - confirma o perfil predominantemente descritivo e expositivo das produções escritas analisadas. Com a ocorrência do pronome pessoal (*I*) e contrações (*can't*), que apontam para traços de oralidade, registros interativos ou textos narrativos mais informais, há indícios de envolvimento interpessoal no corpus escrito. Esses usos revelam que marcas subjetivas da

modalidade falada coexistem com tentativas de formalização escrita, indicando que a separação plena entre os registros ainda está em processo de consolidação pelos aprendizes (Botelho, 2012).

A Dimensão 2, que trata da narratividade, no subcorpus escrito se revela pelo uso de verbos no passado (*was, had, went*) com baixa frequência e marcadores temporais (*today, when, years* e *nowadays*) e sugere uma estrutura expositiva, voltada à descrição e argumentação, com pouca construção de enredos ou sequência de eventos. Nos excertos abaixo o autor do curso de Administração expõe sua argumentação acerca do impacto da arte nas comunidades mais vulneráveis. A dimensão 2 é explícita no exemplo 1, abaixo:

Exemplo 1:

Additionally, it's possible to say that across History, art has been used as a response to conflict and violence, mainly as a form of denunciation and protest, for example in the piece Guernica by Picasso, where the artist aims to shed light and judgment upon the violence of war. This way, art can become the driving force in a community that experiences violence every day, as it gives voice to those suffering and help rebuild the social cohesion that is perhaps necessary to leave the conflict or at very least, it becomes a way of condemning the social injustices present and with that attracting attention and assistance to that community. [...] Furthermore, the expression of art can be a key element in the healing process of those who have suffered traumatic experiences or still live with social, cultural or crime related violence, as it can be a way of reintegrating these people into a community by creating a nonviolent alternative for social interaction and inciting peace based values that re-create a sense of communal trust and help reduce interpersonal violence, as well as social and criminal conflicts.
(CALIENT_TA_2023_0028_2_ENG4_ADM_B2)

Além disso, a predominância de verbos funcionais (*is, are, be, have*) no topo da *WordList* reforça esse estilo mais descritivo. Os dados empíricos apresentados na Tabela 3 destacam que a baixa frequência de verbos dinâmicos e os scores padronizados cancelam o estilo descritivo-expositivo. Os dados mostram que os autores recorrem à narratividade como forma de compartilhar experiências, especialmente no registro escrito mais informal e autobiográfico, como a carta de apresentação. O exemplo 1 demonstra como o aprendiz constrói sua reflexão mobilizando conceitos históricos e sociais em detrimento de uma ordem cronológica de eventos, utilizando construções como "art has been used" e "art can become" para universalizar seu argumento, o que justifica os baixos índices de narratividade típicos desta dimensão. Desse modo, embora os dados mostrem que os autores recorrem à narratividade como forma de compartilhar experiências, especialmente no registro escrito

mais informal e autobiográfico, como a carta de apresentação, textos acadêmicos ou de opinião como o do Exemplo 1 priorizam a exposição conceitual.

Na Dimensão 3 de referência explícita vs. dependência situacional, foram encontrados poucos nomes próprios (*Pedro, Sofia, Mariana*), alta frequência de substantivos concretos (*school, problem, student*), pronomes de terceira pessoa com referência clara (*they, it, them, he*), orações relativas explicativas (*who, which, that*), adjetivos atributivos que qualificam substantivos (*important, human, social*) e uso de *to* em orações complementares. Eles sinalizam a natureza descritiva, explícita e referencial dos textos, uma tentativa de desenvolver ideias com abstração e precisão.

A presença de traços linguísticos relacionados à Dimensão 4 (Persuasão/Subjetividade) revela uma tendência à avaliação subjetiva, posicionamento argumentativo e envolvimento dos autores. Há alta frequência de adjetivos avaliativos (*important, good, bad, necessary*), verbos opinativos (*think, believe, feel, want*), modais epistêmicos (*can, should, must, may, might*), pronomes da 1ª pessoa (*I, we, my, our*). Esses elementos, em conjunto, reforçam as características dos registros do corpus, por exemplo, cartas, editoriais ou textos opinativos, nos quais o autor se posiciona de forma explícita diante dos temas abordados.

A Dimensão 5 avalia o grau de abstração conceitual dos textos e, no subcorpus escrito, há alta frequência de conceitos e nominalizações, que favorecem a generalização, reflexão, análise e argumentação, com menor presença de elementos sensoriais ou descritivos (Biber, 1988). A frequência de substantivos (*violence, society, problem, world, social, culture, expression*) evidenciam o foco em temas amplos e sociológicos.

A Dimensão 6 avalia a coesão lexical dos textos e, neste subcorpus, a densidade de coesão lexical é evidenciada pela alta frequência de termos abstratos relacionados aos Temas Transversais e ODSs (Biber; Conrad, 2009). O corpus escrito apresenta alta coesão lexical, também com uso frequente de nominalizações e colocações temáticas, como mostra o excerto abaixo, que traz uma discussão acerca do aborto.

Exemplo 2:

Currently the question of abortion has been much discussed and has become quite controversial. Some of the reasons that make a woman practice abortion are: financial condition, rejection of her partner or even disapproval of the family. The point is that people should be more concerned with the woman who is generating the baby, even when she wants to abort and when she chooses to have the child. Before imposing any opinion on the subject, people must understand that there is a difference between being in favor of abortion and being in favor of legalizing

abortion. A person can be against abortion and at the same time accept that he cannot impose his/her opinion on the life of another person, deciding, then, to let the woman decide what she wants to do. Even so, abortion brings much discussion between who is for and who is against. The religious question is always used as an argument against abortion. The Bible says that human life is sacred and shouldn't be taken away from anyone except through God's will. Some people who are also against, argue in defense of life saying that if the woman was irresponsible for not being careful, now she must accept the consequences and take care of your baby. In a society where casual sex has become increasingly common and well promoted by the media, many people are liable to have an unwanted child and are forced to create their child because the same hypocritical society desires. We can't ignore the fact that abortion poses a major risk to women's health, especially women with low financial conditions. This is because, with the criminalization of abortion, disadvantaged women end up going to illegal clinics that don't provide good conditions for their patients and this can pose a big risk to their lives. (CALIEMT_2020_TA_1_00022_ENG1_MIN_B1)

Nesse contexto, as colocações temáticas e as nominalizações evidenciam os grupos de palavras que se combinam recorrentemente para discutir temas como aborto, legislação, moralidade e saúde pública. Há forte presença de campos semânticos recorrentes (*question of abortion, practice abortion, religious issues, illegal clinics etc*), o que reforça a densidade conceitual dos textos. Além disso, a repetição de termos-chave e o uso de sinônimos contribuem para a progressão textual e a manutenção de tópicos.

Na Dimensão 7, que trata da formulaicidade, o subcorpus escrito também utiliza fórmulas, mas de natureza mais acadêmica e estruturada. Utiliza fórmulas recorrentes que se dividem em pacotes lexicais multifuncionais gerais (*In conclusion, According to, It is important to note that, On the other hand*) e *construções de linguagem formulaica tipicamente acadêmica destinadas a estruturar objetivos metadiscursivos* (*This paper aims to, The results indicate that*).

No que tange ao léxico específico empregado pelos aprendizes para a discussão de temáticas sociais, a formulaicidade manifesta-se de forma expressiva em torno de palavras-chave como *violence, society* e *problem*. Longe de emergirem como escolhas lexicais isoladas, esses itens funcionam como núcleos de pacotes lexicais e padrões combinatórios recorrentes no subcorpus escrito. No caso de *violence*, observa-se uma forte tendência à coocorrência em colocações do tipo adjetivo + substantivo, tais como (*physical violence, psychological violence* e *verbal violence*), além de coligações gramaticais rígidas que se repetem de forma idêntica entre diferentes textos, a exemplo das construções estruturadas (*living with violence* e *resolve the situation with violence*). Da mesma forma, o item *society* aciona pacotes lexicais

mobilizadas em introduções textuais para situar o problema abordado, destacando-se padrões formulaicos como (*In the present society...*) e a interlíngua recorrente (*in an actually society*) (desvio induzido pelo falso amigo latente em que o aprendiz transfere erroneamente o advérbio *actually* para exercer a função do adjetivo *current* ou *modern*), além da colocação consolidada (*social problem*). Por fim, a palavra *problem* atua como um dos principais núcleos de sintagmas geradores de sequências formulaicas e blocos de sentido nos dados analisados. O termo opera tanto em colocações adjetivais que qualificam a natureza do fenômeno (*global problem, widespread problem, serious problem*), quanto em binômios coordenados de apoio (*verbal or physical, teachers and students*) e sintagmas verbais fixos de resolução e identificação (*solve the problem, identify the problem* e *detect the problem*). Essa padronização combinatória evidencia que a construção textual do aprendiz, nessa dimensão, ancora-se em blocos de linguagem internalizados e replicados para assegurar a coesão e a legitimidade do registro dissertativo-argumentativo.

Com base nos traços recorrentes identificados nas sete dimensões funcionais, o subcorpus escrito revela predomínio de uma estrutura expositiva, com alta informatividade, coesão lexical densa, linguagem abstrata e uso de recursos persuasivos marcados por envolvimento pessoal. Esses padrões indicam que os textos foram produzidos em contextos formais e planejados, com foco na construção argumentativa, reflexão e aprofundamento temático.

4.1.1 Subcorpus oral

Após detalhar o comportamento das dimensões no texto escrito, passa-se à análise do subcorpus oral. Para iniciar essa investigação, a Tabela 2, a seguir, mostra as 20 palavras mais frequentes na modalidade oral.

Tabela 2: *WordList* do subcorpus oral

Type	Rank	Freq	Range	NormFreq	NormRange
like	1	1216	43	22.445.778	0.843
think	2	938	43	17.314.259	0.843
s	3	920	43	16.982.003	0.843
know	4	879	42	16.225.196	0.824
yeah	5	818	41	15.099.216	0.804
people	6	760	41	14.028.611	0.804
that	7	733	40	13.530.226	0.784
I'm	8	535	42	9.875.404	0.824
so	9	505	39	9.321.643	0.765
and	10	492	42	9.081.680	0.824
good	11	465	40	8.583.295	0.784
one	12	416	43	7.678.819	0.843
okay	13	393	38	7.254.269	0.745
right	14	374	42	6.903.553	0.824
would	15	343	35	6.331.334	0.686
it	16	334	39	6.165.205	0.765
going	17	327	38	6.035.994	0.745
say	18	320	37	5.906.784	0.725
yes	19	319	43	5.888.325	0.843
really	20	307	34	5.666.820	0.667

Fonte: Dados do estudo

A análise da lista de palavras do subcorpus oral (Tabela 2) revela um perfil de frequências que reflete as condições de produção da linguagem falada. No topo da lista, observa-se a saturação de pronomes pessoais e verbos de ação ou de processo mental associados a esses pronomes (*think, know, say, going*) na primeira e segunda pessoa (*I, you, we*), o que sinaliza o alto grau de envolvimento interpessoal e subjetividade característicos da fala espontânea (Biber, 1988). Outro traço marcante é a alta recorrência de marcadores conversacionais e conectivos de preenchimento, como *yeah, like, okay, know* e *think*. A presença expressiva desses itens no topo do ranking evidencia o processo de planejamento em tempo real que ocorre durante a interação oral, servindo tanto para manter o turno da fala quanto para monitorar a atenção do ouvinte. No domínio lexical, ao contrário do subcorpus escrito, os substantivos temáticos abstratos aparecem em posições muito mais periféricas. A linguagem oral dos aprendizes organiza-se de maneira menos densa e mais focada na dinâmica da interação imediata.

Da mesma forma como fora feito com o subcorpus escrito, para detalhar como esses padrões se materializam, investigam-se a seguir os traços linguísticos das dimensões propostas por Biber (1988) identificados nas produções orais do CALIEMT.

Na Dimensão 1, o subcorpus oral se revela interativo, subjetivo e emocionalmente marcado, com foco na construção de vínculo e troca interpessoal (Biber, 1988). O envolvimento interpessoal se constitui por meio de pronomes pessoais (*I, we, you*) marcando a presença e subjetividade no subcorpus. Há uma estrutura sintática simples e espontânea, com conectores (*so, and, but*) que organizam o discurso em tempo real, muitas contrações (*can't, don't, I'm*), marcadores conversacionais e expressões formulaicas¹⁴ (*yeah, you know, I mean*) que criam proximidade e fluidez. Predominância de verbos modais (*can, should, would*) e marcadores de opinião e de emoção (*I think, you know, really, feel, want*), que revelam posicionamento pessoal. Esses traços revelam um estilo conversacional, espontâneo e voltado para a interação direta com o interlocutor (típico da fala natural).

O comportamento na Dimensão 2 sugere que, embora o subcorpus oral exiba registros próximos aos relatos pessoais e às reconstruções de experiências pontuais, as baixas frequências globais absolutas de verbos dinâmicos no passado indicam uma estrutura narrativa ainda episódica e circunscrita, sem um encadeamento temporal complexo sistematizado. Há verbos de ação e marcadores temporais (*said, went, today, now, last*) que sugerem sequências de ações e construção de enredos, porém com frequência limitada. Expressões como (*I think, I said* ou *I saw*) podem indicar sequências narrativas curtas, geralmente inseridas em interações espontâneas ou em referências a experiências pessoais (Biber, 1988). No entanto, não se verifica uso consistente de verbos narrativos ou estruturação temporal elaborada.

No subcorpus oral, a Dimensão 3 da elaboração sintática é mais limitada. A estrutura do discurso tende à espontaneidade, com pausas, *fillers*, presença reduzida de conectores subordinativos (*because, if, when, although, so that*) e modificadores elaborados (Biber, 1988). A oralidade privilegia construções simples, coordenações básicas e formulações

¹⁴Referem-se a sequências de palavras que são armazenadas e recuperadas como unidades únicas e pré-fabricadas da memória de longo prazo, em vez de serem geradas palavra por palavra usando regras gramaticais (Wray, 2012, p. 9).

diretas, o que é natural na fala interacional e adequado ao nível de proficiência linguística dos alunos. Ainda que ocorram estruturas subordinadas introdutórias ou *chunks* de enquadramento opinativo, como *I think that*, elas não chegam a formar períodos sintáticos longos ou profundamente ramificados. No entanto, o uso recorrente dessas expressões indica um esforço argumentativo e explicativo por parte dos alunos, característico de interações sob pressão do planejamento em tempo real. Assim, embora o subcorpus oral exiba certas fórmulas organizadoras de opinião, as restrições de tempo real e o nível de proficiência em desenvolvimento limitam o encadeamento gramatical complexo, resultando em uma estrutura predominantemente justaposta e fragmentada

Na Dimensão 4, no subcorpus oral, observa-se uma presença mais visível de *I, we* e *my*, especialmente em sequências interacionais e relatos. Os verbos modais epistêmicos - *would, can, could* - ocorrem com destaque, frequentemente associados à expressão de opinião ou possibilidade. Além disso, aparecem adjetivos subjetivos (*good, bad, nice, important*) e expressões avaliativas (*I think, really, so cool*). Isso aponta para uma linguagem mais expressiva, emocional e posicionada, compatível com a natureza espontânea do discurso oral.

A Dimensão 5 avalia se os textos se ancoram em tempo e espaço específicos (situação concreta) ou se se orientam para conceitos abstratos e generalizações (Biber, 1988). O subcorpus oral apresenta alta densidade de referências situacionais, com uso frequente de marcadores temporais e espaciais (*now, today, here, there, years, nowadays*). Os falantes se referem a experiências vividas, momentos específicos e lugares concretos, o que reforça o caráter contextual e experiencial da fala. A linguagem abstrata é menos presente, indicando que os textos orais se ancoram mais em eventos concretos e imediatos.

A Dimensão 6 é elaborada pelos mecanismos lexicais que conectam semanticamente os elementos do texto, como repetições, sinônimos, colocações e nominalizações (Biber e Conrad, 2009). A coesão lexical no subcorpus oral é construída por meio de repetições temáticas, colocações informais e campos semânticos recorrentes (emoção, opinião, desejo), que, no subcorpus oral, se manifestam pelo uso de *like, think, thought, want, need, feel, good* e *bad*. Há pouco uso de nominalizações e sinônimos técnicos, o que reforça o caráter espontâneo e não planejado da fala. A coesão é mantida por encadeamentos temáticos simples, com forte presença de verbos opinativos e avaliativos (*know, mean, like*). Além de

grande frequência como verbo, *like* auxilia na coesão e encadeamento temático como substantivo, conjunção e como advérbio marcador do discurso, como mostram as linhas de concordância abaixo.

And people, *like*, started to understand more what she was doing and started to, *like*, to understand

Because racism is not, *like*, just prejudice

>Show me what skin color you would *like* to have as your heart. (verbo de desejo)

It looks *like* that England is a very rich country (conjunção comparativa/conformativa informal)

I think someone would *like* to say something.

Is it *like* a national stereotype or a cultural stereotype

I *like* to do my best to help my mother and my sisters

Por fim, a Dimensão 7, da formulaicidade, do uso de expressões fixas, fórmulas linguísticas e padrões recorrentes, que tornam o discurso mais fluido e natural (Biber e Conrad, 2009). No subcorpus oral, a formulaicidade é alta, com uso intenso de expressões fixas e padrões conversacionais. Fórmulas (*I mean, you know, I think, kind of*) funcionam como estratégias de fluidez, envolvimento e reformulação. Os marcadores conversacionais (*yeah, okay, you see, so, right*) reforçam a naturalidade e espontaneidade do discurso oral.

A fim de consolidar e validar estatisticamente o contraste funcional observado ao longo da descrição isolada de ambos os subcorpora, os dados macrogramaticais do CALIEMT foram submetidos ao processamento do MAT Tagger (Nini, 2015). A Tabela 3 apresenta as médias dos Scores Z agregados para as cinco primeiras macrodimensões da língua, cancelando a polarização de registros entre as modalidades oral e escrita dos aprendizes.

Tabela 3 - Distribuição dos Scores Z por Dimensão no CALIEMT

Dimensão de Biber (AMD)	Subcorpus Escrito (Média Score Z)	Subcorpus Oral (Média Score Z)	Tendência Funcional Predominante
Dimensão 1 (Informatividade vs. Envolvimento)	-12,45	+23,12	Escrito: Altamente Informativo / Oral: Interativo
Dimensão 2 (Narratividade)	-1,84	+4,15	Escrito: Não-narrativo / Oral: Estrutura Narrativa
Dimensão 3 (Elaboração Sintática)	+4,72	-2,10	Escrito: Alta Complexidade / Oral: Fragmentada
Dimensão 4 (Persuasão / Expressão Subjetiva)	+1,05	+5,88	Escrito: Argumentação Contida / Oral: Alta Subjetividade
Dimensão 5 (Referência Abstrata vs. Situacional)	+3,27	-1,65	Escrito: Altamente Abstrato / Oral: Situacional/Concreto

Fonte: Dados do estudo gerados a partir do MAT Tagger

Os dados consolidados na Tabela 3 evidenciam uma polarização funcional nítida entre as modalidades do CALIEMT, validando estatisticamente as análises qualitativas precedentes. Na Dimensão 1, a discrepância entre o subcorpus escrito (Score Z = -12,45) e o oral (Score Z = +23,12) traduz o distanciamento entre a alta densidade informativa da escrita planejada e a natureza interativa e engajada da fala espontânea. Enquanto as redações acumulam substantivos e preposições para adensar o conteúdo, as produções orais apoiam-se em pronomes e verbos de ação para sustentar a interação imediata.

Esse planejamento formal da escrita reflete-se diretamente na Dimensão 3 (Score Z = +4,72) e na Dimensão 5 (Score Z = +3,27), eixos nos quais o subcorpus escrito pontua de forma positiva. Esses índices atestam o emprego de uma estrutura sintática elaborada e caracterizada por sentenças dependentes e modificadores complexos, além de uma orientação para referências abstratas e generalizações conceituais, distantes do contexto imediato de produção. Em contrapartida, o subcorpus oral apresenta Scores negativos em ambas as dimensões (Dim3 = -2,10; Dim5 = -1,65), indicando uma sintaxe fragmentada e uma ancoragem referencial eminentemente situacional e concreta.

Por fim, os comportamentos nas Dimensões 2 e 4 revelam que o subcorpus oral é substancialmente mais dinâmico e subjetivo. O Score positivo na Dimensão 2 (Score Z = +4,15) confirma a presença de estruturas narrativas estruturadas no passado, enquanto a Dimensão 4 (Score Z = +5,88) expõe a saturação de marcas de persuasão, adjetivos avaliativos e

posicionamentos epistêmicos de primeira pessoa. Na escrita, a argumentação assume um tom mais contido e formal (Dim4 = +1,05), com baixa recorrência de sequências cronológicas temporais (Dim2 = -1,84). Em suma, a síntese estatística dos *Scores Z* marca o CALIEMT como um corpus cuja variação de registro é fortemente condicionada pelas restrições de produção e objetivos comunicativos de cada modalidade.

Depois de analisar a estrutura gramatical geral das cinco primeiras dimensões, o próximo passo é entender como as palavras se combinam na prática. Para isso, investigam-se a seguir os blocos de palavras (N-gramas) extraídos com o AntConc. Essas sequências repetitivas nos ajudam a identificar os padrões de coesão lexical (Dimensão 6) e, principalmente, o uso de expressões fixas e clichês que definem a formulaicidade (Dimensão 7) no subcorpus oral e no escrito.

4.2 Análise dos N-gramas no CALIEMT

A análise abaixo apresenta os N-gramas mais frequentes dos dois subcorpora e os pacotes lexicais foram usados para ilustrar as dimensões ao longo dos subcorpora. O levantamento de N-gramas priorizou a extração parametrizada de quadrigramas e sequências contínuas de 3 a 5 palavras devido ao alto potencial funcional desses blocos (*chunks*) para ilustrar comportamentos de formulaicidade estável. A Tabela 4 apresenta os N-gramas mais frequentes do subcorpus escrito, que revelam padrões distintos das dimensões propostas por Biber (1988) na Análise Multidimensional: informação acadêmica (informatividade) vs foco interpessoal e na interação (ou no envolvimento).

Tabela 4 - N-Gramas do subcorpus escrito

Type	Rank	Freq	Range	NormFreq	NormRange
I'm (x) years old	1	43	43	241.351	0.039
the power of art	2	41	23	230.125	0.021
is a form of	3	39	35	218.899	0.032
best form of art	4	38	18	213.287	0.016
the art can be	6	35	30	196.448	0.027
has the power to	7	34	24	190.835	0.022
live in ouro preto	7	34	34	190.835	0.031
art is very important	9	31	28	173.997	0.026
need to learn English	9	31	31	173.997	0.028
the art is important	11	29	29	162.771	0.027
and racial prejudices that	12	28	28	157.159	0.026
art and culture are	12	28	18	157.159	0.016
commonplace in brazilian society	12	28	28	157.159	0.026
prejudices that were once	12	28	28	157.159	0.026
the best form of	12	28	15	157.159	0.014

Fonte: dados do estudo.

A análise sugere que esse corpus pode se aproximar da dimensão de informatividade, caracterizada pelo uso de sintagmas nominais complexos. N-gramas como *the power of art*, *is a form of*, *as the power to* e *art is very important* indicam uma prevalência de construções nominais e a expressão de ideias abstratas e conceituais, afastando-se da informalidade e da pessoalidade, conforme mostra o exemplo 3.

Exemplo 3:

The power of art has been a significant force in shaping societies and expressing human emotions throughout history. This argumentative text explores the impact of art on health, violence, and imagination, illuminating its true transformative power in enriching our lives. Art's therapeutic quality has been acknowledged for centuries, benefiting mental health. Creative activities like painting and crafting offer an outlet for self-expression and reduce stress, anxiety, and depression. art therapy addresses emotional (CALIEMT_TA_2023_0199_1_ENG3_MIN_B1)

A presença de substantivos como *art*, *power* e *form* em pacotes lexicais organizadores - a exemplo de (*is a form of*), cuja função é introduzir definições - indica um direcionamento voltado para a transmissão de conteúdos dissertativos, introdução aos tópicos, e na construção da argumentação ou elaboração dos tópicos (Simpson-Vlach, 2010). Esses pacotes

lexicais se aproximam do registro acadêmico por sua formalidade, sua ênfase em conceitos e ideias, e contribuem para a construção de um discurso objetivo e informativo. Os N-gramas (*and racial prejudices that, commonplace in Brazilian society e prejudices that were once*) indicam a discussão de temas sociais relevantes, que são frequentes no discurso acadêmico nas áreas de Humanas, conforme mostra o exemplo 4.

Exemplo 4:

Some cartoons may show some ethnic and racial prejudices that **were once commonplace in Brazilian society**. Since the children are supple¹⁵, they repeat the actions that the characters do, and this influences them to become violent people when adults. Tom and Jerry, South Park and Wood woody Pecker **are examples that bring** these characteristics. (CALIEMT_TA_2012_0058_2 EDI_A1)

No Exemplo 4, observa-se a proposta teórica de Simpson-Vlach (2010), explicitada por meio do emprego de N-gramas complexos e pacotes lexicais organizadores de base abstrata. O fragmento “ethnic and racial prejudices that were once commonplace in Brazilian Society” funciona como um pacote de densidade conceitual, característico do registro acadêmico da área de Ciências Humanas. A inserção desse N-grama projeta no texto um direcionamento dissertativo e voltado ao debate de problemáticas sociais de relevância contemporânea. Ademais, a articulação argumentativa do estudante consolida-se na última sentença através do pacote organizador “are examples that bring”. Essa construção cumpre a função metadiscursiva de estruturar e encerrar o tópico por meio da exemplificação e da classificação de conceitos anteriores (no caso, a influência dos desenhos animados). Essa reiteração de molduras formulaicas e substantivos abstratos confere ao trecho o distanciamento enunciativo, a objetividade e a formalidade técnico-científica exigidas pelo letramento acadêmico

Por outro lado, alguns N-gramas presentes na Tabela 3 se distanciam do registro acadêmico formal, embora possam ocorrer em contextos acadêmicos mais informais ou em gêneros específicos. O N-grama *I'm (x) years old* é um marcador de linguagem pessoal, típico de uma carta, redação de apresentação pessoal ou relato de experiências. Da mesma forma, (*live in Ouro Preto*) sugere um contexto local e possivelmente mais descritivo desses registros e (*need to learn English*) também se afasta do registro formal acadêmico típico porque

¹⁵ Impressionable

expressa uma necessidade de ordem estritamente prática, utilitária e pessoal do sujeito. Textos puramente informativos e científicos priorizam processos conceituais amplos e universais em detrimento de demandas particulares de agência imediata. O exemplo 5 revela uma perspectiva subjetiva da carta de apresentação pessoal, o uso da primeira pessoa do singular e uma preocupação com o eu e suas experiências ou necessidades.

Exemplo 5:

I am Isa. **I'm 16 years old. I live in Ouro Preto.** I study Business in IFMG, because I believe that this school can teach me really important things that would help me in my professional life. I speak Portuguese and English **and I need to learn English** because it will help me understand foreign people and of course, complete my skills. My favourite school subject is Math, because I know every step that are made, but my favourite course subject is Accounting, it's the one that I love the most.
(CALIEMT_TAP_2022_0108_2_ENG_ADM_B1)

O uso da primeira pessoa do singular (*I, my*) sinaliza uma perspectiva subjetiva porque ancora o discurso no próprio enunciador, o que contrasta com a estratégia de despersonalização e apagamento do sujeito comumente requerida para a validação do distanciamento dissertativo formal. Biber (1988) aponta que traços como alta dependência do contexto situacional imediato, referências egocêntricas de primeira pessoa e menor densidade de vocabulário de conteúdo abstrato constituem eixos centrais da fala que se opõem diametralmente à generalização e ao planejamento informacional da escrita. Os N-gramas que tratam de informações muito específicas e contextuais, como *live in ouro preto*, indicam uma relação mais direta com a situação imediata do autor ou com um contexto compartilhado que não é típico da generalização e abstração da escrita acadêmica. A carta pessoal tende a ser mais situacionalmente dependente e menos densa em informação abstrata. O CALIEMT escrito apresenta registros diversos, e esta breve análise de N-gramas sugere que, apesar de ser um corpus escrito, ele incorpora características de registros que valorizam a expressão pessoal, a interação direta e a referência a experiências ou fatos do cotidiano, em contraste com a impessoalidade e a densidade informacional esperadas em textos acadêmicos formais.

A Tabela 5 abaixo apresenta os N-gramas (3 a 4 palavras) mais frequentes do subcorpus oral do CALIEMT e revela padrões linguísticos típicos da interação falada com uma proeminência de estruturas como pacotes lexicais de opinião epistêmica (Simpson-Vlach, 2010): *I don't know, I think it is e what do you think*.

Tabela 5 - N-gramas do Corpus Oral (3-4 palavras)

Type	Rank	Freq	Range	NormFreq	NormRange
I don' t know	1	496	41	3.456.181	0.804
don' t know what	2	148	17	1.031.280	0.333
I think it is	3	147	34	1.024.312	0.667
I know what to	4	133	6	926.758	0.118
know what to do	5	130	7	905.854	0.137
what to do is	8	121	2	843.141	0.039
I' m going to	12	73	20	508.672	0.392
I believe I believe	13	72	1	501.704	0.020
what do you think	13	72	26	501.704	0.510
don' t know how	17	53	22	369.310	0.431
we re going to	19	48	22	334.469	0.431
I would like to	20	47	14	327.501	0.275
I don't think	21	44	20	306.597	0.392

Fonte: dados do estudo.

Essas ocorrências são consistentes com as características da análise multidimensional de Biber (1988), que identifica a dimensão de foco interpessoal e interação como característica marcante da linguagem falada. Os N-gramas predominantes refletem o uso de pronomes de primeira pessoa em suas formas puras e clíticas (*I, I'm, we*) associados a marcadores de opinião, modulação epistêmica - *don't know* - e projeção futura - *going to* -, que são elementos chave na construção da interação e na expressão de subjetividade (Garcez et al., 2012). O exemplo 6¹⁶ abaixo, extraído do corpus oral, é de um encontro em que os participantes discutem como identificar *fake News* e como se proteger na internet.

Exemplo 6:

Ana: Have you got some suggestions on how to spot fake news?

Shelley: Yeah, we have started talking about that.

Ana: So, Jefferson, Lidia, Carolina, **would you guys like** to contribute?

Carolina: Yes (...), in our group, I have a confident, a real person, and **I believe (...)** **I believe** in her. She's extremely, **I don't know** if I may use the correct word, but she's aggressive with the fake news. And then she helps me a lot about when I, for example, I sent to her the message and she wrote me, no, it's not true, come on, let's

¹⁶ Nomes fictícios foram acrescentados aos turnos para resguardar a identidade dos participantes e pelo fato de o *TurboScribe* nomear todos os participantes como palestrantes.

go, for this, for that. And then I check the correct search, I do this, but sometimes I ask her.

Expressões como *I don't know* e *I think it is* atuam como marcadores de hesitação ou atenuação, que permitem que o falante gerencie a progressão do turno e evite imposições diretas (Garcez, 2006). O *chunk* da pergunta *What do you think* pode servir como um movimento de abertura de turno ou de busca de engajamento, convidando os interlocutores a participarem. A alta frequência de *I'm going to* pode indicar uma projeção de ações futuras, comum em narrativas pessoais e planejamento colaborativo em conversas.

5 Considerações finais

O CALIEMT - corpus de aprendizes da língua inglesa do Ensino Médio Técnico - configura-se como uma contribuição à Linguística de Corpus e à Linguística Aplicada ao reunir produções escritas e orais de jovens aprendizes do ensino médio técnico profissional, uma área pouco explorada. A estrutura do corpus, apoiada por metadados pedagógicos e perfis linguísticos, oferece suporte para pesquisas futuras sobre aquisição linguística e interlíngua. A análise preliminar das WordLists e dos N-gramas dos subcorpora revelam padrões que se alinham às dimensões propostas por Biber (1988). O subcorpus escrito destaca-se pelas altas ocorrências e complexidade de substantivos, preposições e conectores informativos, apontando forte densidade informacional e baixa subjetividade, características típicas de registros acadêmicos. O subcorpus oral revela presença de pronomes pessoais, verbos modais e marcadores conversacionais, demonstrando maior envolvimento interpessoal, subjetividade e traços narrativos.

A partir da breve análise dos dados pela AMD, nota-se que a escrita apresenta maior coesão lexical (Dimensão 6), com uso de nominalizações e termos abstratos, enquanto a oralidade evidencia elevada formulaicidade (Dimensão 7), por meio de expressões interacionais como *I think, you know, kind of*. Essa oposição entre envolvimento e informatividade sublinha os estilos e funções distintas da produção linguística dos aprendizes.

Este estudo mapeou a variação de registro e o uso de linguagem nos subcorpora oral e escrito do corpus CALIEMT, utilizando como base as dimensões de Biber (1988). A integração entre a análise qualitativa de exemplos textuais e a validação estatística por meio de Scores Z

permitiu alcançar conclusões sobre a produção linguística dos aprendizes. Os resultados do MAT Tagger comprovaram que as duas modalidades operam em polos funcionais opostos. O subcorpus escrito caracterizou-se por uma alta informatividade, grande complexidade sintática (Dimensões 1 e 3) e um foco em conceitos abstratos (Dimensão 5). Já o subcorpus oral destacou-se pelo forte envolvimento interpessoal (Dimensão 1), presença de estruturas narrativas (Dimensão 2) e alta expressão subjetiva (Dimensão 4). Essas diferenças mostram que os alunos ajustam sua linguagem dependendo do canal de comunicação utilizado. No que diz respeito à formulaicidade (Dimensão 7) e à coesão lexical (Dimensão 6), a extração de N-gramas via AntConc revelou que os estudantes recorrem a blocos de palavras pré-moldados em ambas as modalidades. Enquanto na oralidade predominam marcadores conversacionais e interativos, na escrita há um uso sistemático de pacotes lexicais acadêmicos e colocações fixas para organizar o texto e debater problemas sociais. Em suma, a abordagem multidimensional aplicada nesta pesquisa demonstrou que o desenvolvimento da interlíngua dos aprendizes não é homogêneo.

Espera-se que este trabalho contribua para os estudos de Linguística de Corpus e ofereça subsídios para a criação de materiais pedagógicos para o ensino de línguas, de forma a expor os alunos do ensino médio técnico profissional tanto a estruturas gramaticais complexas quanto a padrões fraseológicos fixos em contextos reais de uso.

A capacidade do CALIEMT de captar essas variações reforça seu potencial para iluminar a relação entre o engajamento oral e escrito na língua-alvo e contribui para reflexões pedagógicas mais contextualizadas sobre o ensino de inglês no Ensino Médio Técnico.

Agradecimentos

Agradecimentos ao Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais pelo financiamento da pesquisa e aos professores, alunos bolsistas e alunos participantes do Projeto.

Referências

ALUÍSIO, S. M.; ALMEIDA, G. M. de B. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. **Calidoscópio**, [S. l.], v. 4, n. 3, p. 156–178, 2021. Disponível em:

<https://revistas.unisinos.br/index.php/calidoscopio/article/view/6002>. Acesso em: 22 abr. 2025.

ALVES, J. H. L. **CorpuScript**. Software em desenvolvimento. 2024.

ARBOR, A. **Michigan Corpus of Upper-level Student Papers**. MI: The Regents of the University of Michigan, 2009.

BAKER, P. **Using Corpora in Discourse Analysis**. London: Continuum. 2006.
<https://doi.org/10.5040/9781350933996>.

BIBER, D. S. **Variation across speech and writing**. Cambridge: Cambridge University Press, 1988. <https://doi.org/10.1017/CBO9780511621024>.

BIBER, D. S.; CONRAD; REPPEN, R. **Corpus Linguistics: Investigating language structure and use**. Cambridge: Cambridge University Press, 1998.
<https://doi.org/10.1017/CBO9780511804489>.

BIBER, D.; CONRAD, S. **Register, genre, and style**. Cambridge: Cambridge University Press, 2009. <https://doi.org/10.1017/CBO9780511814358>.

BOTELHO, J. M. **Oralidade e escrita sob a perspectiva do letramento**. Jundiaí: Paco Editorial, 2012.

CRISTAL, D. **English as a global language**. Cambridge: Cambridge University Press, 2003.

CONSELHO DA EUROPA. **Quadro comum europeu de referência para as línguas: aprendizagem, ensino, avaliação**. Edição portuguesa. Porto: Edições Asa, 2001. Disponível em:
https://www.dge.mec.pt/sites/default/files/Basico/Documentos/quadro_europeu_comum_referencia.pdf. Acesso em 12 dez. 2025.

DUTRA, D. P., ORFANÒ, B. M., GUEDES, A. de S., ALVES, J. C., FEKETE, J. G. The learner corpus path: a worthwhile methodological challenge. **DELTA: Documentação De Estudos Em Lingüística Teórica E Aplicada**, v. 38, n. 2, 2022, DOI: <https://doi.org/10.1590/1678-460X202238249731>

FERNANDES, L.; SILVA, C. E. COPA-TRAD Version 2.0 (Translational Parallel Corpus). Universidade Federal de Santa Catarina, Florianópolis, 2016. Disponível em: <<http://copa-trad.ufsc.br>>. Acesso em: 10 Dec. 2025.

GARCEZ, P. M. A organização da fala-em-interação na sala de aula: controle social, reprodução de conhecimento, construção conjunta de conhecimento. **Calidoscópio**, v. 4, n. 1, p. 66-80, 2006.

GARCEZ, P. M.; FRANK, I.; KANITZ, A. Interação social e etnografia: sistematização do conceito de construção conjunta de conhecimento na fala-em-interação de sala de aula.

Calidoscópio, São Leopoldo, v. 10, n. 2, p. 211-204, 2012.

<https://doi.org/10.4013/cld.2012.102.08>

GILLINGS, M.; MAUTNER, G.; BAKER, P. **Corpus-Assisted Discourse Studies: elements in Corpus Linguistics**. Cambridge: Cambridge University Press. 2023.

<https://doi.org/10.1017/9781009168144>.

GILQUIN, G.; DE COCK, S.; GRANGER, S. **The Louvain International Database of Spoken English Interlanguage**. Louvain-la-Neuve: Presses universitaires de Louvain, 2010.

GRANGER, S.; DUPONT, M.; MEUNIER, F.; NAETS, H.; PAQUOT, M. **The International Corpus of Learner English**. Version 3. Louvain-la-Neuve: Presses universitaires de Louvain, 2020.

<https://dial.uclouvain.be/pr/boreal/object/boreal:229877>

GRANGER, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (ed.) **Learner English on Computer**. Addison Wesley Longman: London & New York, 1998, p. 3-18. <https://doi.org/10.4324/9781315841342-1>.

GONZÁLEZ, Z. M. G. **Linguística de Corpus na análise do Internetês**. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) - Faculdade de Letras. Pontifícia Universidade Católica de São Paulo, São Paulo, 2007.

KAATARI, H., WANG, Y., & LARSSON, T. (2024). Introducing the Swedish Learner English Corpus: a corpus that enables investigations of the impact of extramural activities on L2 writing. **Corpora**, v. 19, n. 1, 2024, p. 17–30. <https://doi.org/10.3366/cor.2024.0296>.

MARQUES, C. G. F. CorAgrarian: um corpus de artigos científicos de ciências agrárias. In: **Semana de Eventos da Faculdade de Letras (SEvFALE)**, Universidade Federal de Minas Gerais, Belo Horizonte, 2024.

MATTOS, A. M. A. Novos Letramentos: perspectivas atuais para o ensino de inglês como língua estrangeira. **SIGNUM: Estud. Ling.**, Londrina, n. 17/1, p. 102-129, jun. 2014. <https://doi.org/10.5433/2237-4876.2014v17n1p102>

MICROSOFT COPILOT. Interações com assistente de IA em ambiente digital para revisão textual. Disponível em: <<https://copilot.microsoft.com/>>. Acesso em: 22 abr. 2025.

NESI, H.; GARDNER, S.; THOMPSON, P. WICKENS, P. **BAWE CORPUS. The British Academic Written English Corpus**. [S.l.]: University of Warwick, 2008-2009. Disponível em: <https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/> . Acesso em: 03 dec. 2025.

NINI, A. **Multidimensional Analysis Tagger** (Version 1.3). [S. l.]: [s. n.], 2015. Disponível em: <http://sites.google.com/site/multidimensionaltagger>. Acesso em: 20 Jun. 2025.

OLIVEIRA, S. B. Abordando temas polêmicos: representações sociais construídas em aulas de conversação de língua inglesa. In: XXVI Jornada do Grupo de Estudos Linguísticos do

Nordeste, 2016, Recife. **Anais da XXVI Jornada do Grupo de Estudos Linguísticos do Nordeste**. Recife: Pipa Comunicação, 2017. v. 8. p. 719- 734.

OLIVEIRA, S. B.; ANICETO, A. B. V.; ALCANTARA, B. G.; SILVA, C. A. L.; AZEVEDO, L. T.; MELO, L. O.; FONSECA, M. M. S. Escrita colaborativa no ensino e aprendizagem da Língua Inglesa. In: GONÇALVES, G. E.; OLIVEIRA, S. B. (Org.). **Pesquisa e extensão na escola pública: possibilidades e desafios**. 1ed. Goiânia: Editora Espaço Acadêmico, 2018, v. 1, p. 140-154.

OLIVEIRA, E. B. SOUZA, W. F. Educação Profissional Tecnológica: implicações do movimento CTSA para o currículo. In: OLIVEIRA, A. R.; XAVIER, H. C. SILVA, J. F.; OLIVERIA S. B. (Orgs.). **Educação Profissional no Brasil: da história à teoria, da teoria à práxis**. Coleção Educação Profissional no Brasil. v. 1. Curitiba: Editora CRV, 2020, p. 137-152.
<https://doi.org/10.24824/978652510121.7>.

RAJAGOPALAN, K. **Por uma linguística crítica: linguagem, identidade e a questão ética**. São Paulo: Parábola Editorial, 2003.

RESENDE, S. V.; MAVERICK, R. Planejamento, Compilação e Organização de Corpora. **Blucher Social Sciences Proceedings**, v.2, n. 3, março de 2016. Disponível em:
http://pdf.blucher.com.br.s3-sa-east-1.amazonaws.com/socialsciencesproceedings/viii-ebic-xiii-elc/06_artigo_03.pdf Acesso em 28 jan. 2018.

RISSANEN, M.; KYTÖ, M.; KAHLAS-TARKKA, L. KILPIÖ, M.; NEVANLINNA, S. TAAVITSAINEN, I.; NEVALAINEN, T.; RAUMOLIN-BRUNBERG, H. **The Helsinki Corpus of English Texts**. Department of Modern Languages, University of Helsinki. 1991.

SARDINHA, T. B. Linguística de Corpus: Histórico e Problemática. In: **DELTA: Documentação e Estudos em Linguística Teórica e Aplicada**. v.16, n. 2, 2000, p. 323-367.
<https://doi.org/10.1590/S0102-44502000000200005>

SARDINHA, T. B. O Corpus de Aprendizes BR-ICLE. **Intercâmbio**. v. x., 2001. p. 227-239.

SARDINHA, T. B. **Linguística de Corpus**. Barueri-SP. Manole, 2004.

SARMENTO, S. Linguística de Corpus: histórico, metodologia, campos de aplicação. **Revista Trama**. v. 6, n. 12, 2º Semestre de 2010, p. 87 - 107.

SELINKER, L., Interlanguage, **IRAL; International Review of Applied Linguistics in Language Teaching**, v. 10, n. 3, 1972. <https://doi.org/10.1515/iral.1972.10.1-4.209>.

SIMPSON-VLACH, R.; ELLIS, N. An academic formulas list: new methods in phraseology research. **Applied Linguistics**. Advance Access Published. January 12, 2010, p. 1-26.

SINCLAIR, J. **Corpus, Concordance and Collocation**. Oxford: Oxford University Press, 1991.

SINCLAIR, J.; KARLSSON, F. **The Bank of English (COBUILD Corpus)**. Harper Collins Publishers, Glasgow, 1993. Disponível em: https://www.lt-world.org/kb/resources-and-tools/language-data/ltw_x3alanguage_x5fdata_.2010-09-23.5678579368 Acesso em: 12 dec. 2025.

SON, R. V. Amsterdam Centre for Language and Communication, University of Amsterdam, 2007, **IFA Dialog Video corpus**. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), 2007. <http://hdl.handle.net/11372/LRT-735> .

STORCH, N. Collaborative writing: Product, process, and students' reflections. **Journal of Second Language Writing**, 14, p. 153-173, 2005. <https://doi.org/10.1016/j.jslw.2005.05.002>.

STREET, B. **Letramentos Sociais. abordagens críticas do letramento no desenvolvimento, na etnografia e na educação**. Trad. Marcos Bagno. São Paulo: Parábola Editorial, 2014.

STUBBS, M. **Words and Phrases: Corpus Studies of Lexical Semantics**. London: Blackwell. 2001.

SULIS, G. Engagement in the foreign language classroom: Micro and macro perspectives. **System**, v. 110, n. 102902, 2022, p. 1-13. <https://doi.org/10.1016/j.system.2022.102902>.

TILLMANN, C.; DRAXLER, K.; KOTTEN, F. S. The Phonetic goals of the new Bavarian Archive for Speech Signals. **Proceedings of the ICPHS 1995**. Stockholm, 1995, v. 4, p. 550-553.

VIANA, V. Linguística de corpus: conceitos, técnicas e análises. In: VIANA, V.; TAGNIN S. E. O. (Org.) **Corpora no ensino de línguas estrangeiras**. São Paulo. Hub Editorial, 2010. p. 25-95.

VINOKUROVA, V.; CENTANIN-BERTHO, M.; STAPLES, S. **Modelling learner corpus-based instruction: examples from less commonly taught languages**. Workshop. University of Arizona. 2023.

VYGOTSKY, L. S. **Mind in society: The development of higher psychological processes**. Cambridge: Harvard University Press, 1978.

VYGOTSKY, L. S. **A formação social da mente**. São Paulo: Martins Fontes, 1987.

WANG, Y. KAATARI, H. LARSSON, T. Introducing the Chinese Learner English Corpus. **Book of abstracts**. National Forum for English Studies 2025, Lund 9-11 April. 2025. Disponível em: https://konferens.ht.lu.se/fileadmin/user_upload/conference/NFES-25/Book_of_abstracts.pdf#page=10 Acesso em 20/04/2025.

WRAY, A. **Formulaic language: pushing the boundaries**. Oxford: Oxford University Press, 2012.

ZAVAGLIA, C.; FERRARESI, M. L. Construção de um corpus paralelo e alinhado português-italiano-português para o domínio literário. **Estudos Linguísticos XXXV**, p. 191-200, 2006. [191 / 10]

Recebido em: 20.08.2025

Aprovado em: 01.07.2026

Anexo – Procedimento de desidentificação dos arquivos

- Deletar o email do participante
- Deletar o nome da pessoa
- Salvar arquivo com o código: CALIEMT_TAP_2022_0001_1_ENG2_MIN_A1

Nome do projeto: CALIEMT

Registro

texto de apresentação pessoal TAP

texto argumentativo TA

texto expositivo de opinião TO

poesia PO

narrativa de aprendizagem NAP

transcrição de aulas ou de discussão: TRA

Ano

2022

Número do texto de acordo com o ano

sexo / gênero

1 masculino

2 feminino

3 outro

Nível

ENG1 Inglês 1 primeiro ano

ENG2 Inglês 2 segundo ano

ENG3 Inglês 3 terceiro ano

ENG4 proficientes

curso

MIN Mineração

MET Metalurgia

ADM Administração

EDI Edificações

AUT Automação

Identificação do nível de proficiência do aluno de acordo com o Quadro Comum Europeu de Referência para as Línguas (CEFR, na sigla em inglês) que define seis níveis de proficiência linguística, de A1 a C2, que são divididos em três níveis amplos: Usuário Básico (A1 e A2), Usuário Independente (B1 e B2) e Usuário Proficiente (C1 e C2).