DOI: 10.12957/palimpsesto.2024.79135

Análise multidimensional de um *corpus* multinível de aprendizes de língua inglesa

Multidimensional analysis of a multilevel learner corpus of English language

Cicero Soares da Silva
Pontifícia Universidade Católica de São Paulo (PUC-SP)
pardonmester@gmail.com
https://orcid.org/0000-0003-2815-7977

RESUMO

Este artigo tem como objetivo identificar as dimensões de variação nos padrões de elementos lexicogramaticais da linguagem utilizada por aprendizes de inglês, utilizando o *Corpus* Multinível de Aprendizes Brasileiros de Inglês (COBRA-7), compilado por W. M. Dantas em 2012. O COBRA-7 consiste em 2516 redações escolares, divididas em seis níveis de proficiência, totalizando 571.564 palavras. A pesquisa se insere na área da Linguística de *Corpus* de Aprendiz (LCA), que investiga textos autênticos produzidos por aprendizes de línguas estrangeiras. A metodologia empregada é a Análise Multidimensional Funcional (AMD) no contexto da Linguística de *Corpus*, utilizando ferramentas computacionais. A análise buscou identificar as dimensões de variação gramático-funcionais no corpus e, através do uso do procedimento estatístico ANOVA, examinou a variação multidimensional funcional em relação a variáveis independentes como "escola", "nível" e "gênero biológico". Os resultados indicaram a existência de variação estatística entre os diferentes níveis de ensino.

Palavras-chave: Linguística de *corpus*; Linguística de *corpus* de aprendiz; níveis de proficiência; Análise Multidimensional.

ABSTRACT

This article aims to identify the dimensions of variation in the patterns of lexicogrammatical elements used by English learners, using the Multilevel *Corpus* of Brazilian English Learners (COBRA-7), compiled by W. M. Dantas in 2012. COBRA-7 consists of 2516 school essays, divided into six proficiency levels, totaling 571,564 words. The research falls within the field of Learner *Corpus* Linguistics (LCL), which investigates authentic texts produced by foreign language learners. The methodology employed is Functional Multidimensional Analysis (AMD) within the context of *Corpus* Linguistics, using computational tools. The analysis aimed to identify the grammatical-functional dimensions of variation in the *corpus* and, through the use of the ANOVA

statistical procedure, examined functional multidimensional variation in relation to independent variables such as "school," "level," and "gender." The results indicated the existence of statistical variation among different proficiency levels.

Keywords: *Corpus* Linguistics; Learner *Corpus* Research; Proficiency Levels; Multidimensional Analysis.

INTRODUÇÃO

Atuando como professor de inglês por mais de 25 anos, sempre priorizei trabalhar com uma abordagem comunicativa apesar de haver poucos recursos disponíveis nas escolas públicas e cursos de língua em que atuei. E para isso, sempre procurei fazer adaptações no material didático, pois seu conteúdo era muitas vezes descontextualizado e com 100% do foco na estrutura gramatical da língua. Percebi que grande parte do conteúdo proposto no livro didático era adaptado para aquela temática e que, na maioria das vezes, não atendia à necessidade dos meus alunos.

Algo desafiador para professores e aprendizes de Inglês como Língua Estrangeira (ILE) é atingir domínio para conseguir se expressar natural e fluentemente tanto na escrita quanto na fala. Muitas teorias já foram elaboradas para tentar explicar o porquê de a linguagem de aprendizes ser artificial e "desfluente" (do inglês *dysfluent*), enfocando pressupostos da Gramática Gerativa e da Aquisição de Segunda Língua (Chomsky, 1995; Krashen, 1981), priorizando a aplicação de regras gramaticais formais e abstrata na produção da linguagem. Segundo Biber e Conrad (2009, p. 4) "[...] alguns falantes não-nativos são criticados por soarem muito 'como um livro' quando falam". (tradução nossa)[1]

Vale ressaltar que a proposta da Linguística de *Corpus* (LC) utilizada nesta pesquisa (Sinclair, 1991, Biber, Conrad, Reppen, 1988; Partington, 1988; Hunston, 2002, Berber-Sardinha, 2004) vê a produção da linguagem não do ponto de vista da aplicação de regras, mas sim do uso de sequências linguísticas, por exemplo, sequências de palavras típicas de um dado contexto ou situação de uso de língua (registros). "Registro[2] é uma

[1] [...]some non-native speakers are criticized for sounding too much "like a book" when they speak. Thus, proficiency with spoken registers for conversations and meetings is also important. (Biber, Conrad, 2009, p. 4)

[2] No original: "A cover term for any language variety defined by its situational characteristics, including the speaker's purpose, the relationship between speaker and hearer, and the production circumstances" (Biber, 2009 p. 823).

variedade linguística, definido por aspectos situacionais, que inclui o propósito do falante, a relação entre o falante e o ouvinte e o contexto de produção" (Biber, 2009, p. 823). Para Biber e Conrad (2009, p. 4) "[...] ter um conhecimento dos registros falados como conversações e reunião [...] é importante".

A linguagem não é estruturada pelo princípio da escolha aberta como um "preenchimento de lacunas" (*slot-and-filler model)* (Lewis, 2000b; Sinclair, 1991). Vale enfatizar que a linguagem é altamente padronizada (*patterned*) e os estudos com base na LC têm revelado empiricamente essas descobertas, pois os traços linguísticos coocorrem de modo estatisticamente significativo (Biber, Conrad & Reppen, 1998; Sinclair, 1991; Berber-Sardinha, 2010).

Berber-Sardinha (2000, p. 51) ainda ressalta que o estudo de padronização encontra amparo teórico na noção do princípio idiomático (*idiom principle*), segundo o qual "o usuário da língua tem à sua disposição um grande número de frases pré- ou semiconstruídas que se constituem em escolhas únicas, muito embora pareçam analisáveis em segmentos" (Sinclair, 1987, p. 320).

O princípio idiomático, estabelecido por Sinclair (1991), pressupõe que a linguagem deve ser estudada a partir de seu uso em situações cotidianas, que definem quais palavras sofrerão maior ou menor grau de atração. É o uso que determina qual agrupamento e significado virá desse processo. Assim, esse princípio defende que o falante tem à sua disposição um conjunto de combinações lexicais e gramaticais pré-construídas dentre as quais usará aquela que melhor convier à situação.

A pesquisa descrita aqui tem como objetivo investigar empiricamente os padrões de coocorrência das características linguísticas subjacentes às variedades textuais dos aprendizes de Inglês como Língua Estrangeira (ILE), desde o nível básico ao avançado, presentes no *corpus* COBRA-7, com base no Quadro Europeu Comum de Referência para Línguas – (CEFR)[3].

[3] O Quadro Europeu Comum de Referência para Línguas (QECL ou CEFR) categoriza a proficiência em seis níveis. Começando pelo A1 (Iniciante) com expressões familiares, segue para o A2 (Básico) com compreensão de frases comuns. No nível B1 (Intermediário), aborda tópicos familiares, enquanto o B2 (Usuário Independente) lida com textos complexos. O C1 (Proficiência Operativa Eficaz) compreende textos longos e significados implícitos, e o C2 (Domínio Pleno) engloba a compreensão quase completa e a síntese de informações de várias fontes. (British Council Brasil. https://www.britishcouncil.org.br › quadro-comum-eu. Acesso em 04.12.2021)

Este estudo oferece uma contribuição para a área da Linguística de *Corpus*, especificamente a Linguística de *Corpora* de Aprendizes no Brasil, uma área de pesquisa promissora que já dá seus primeiros passos no contexto acadêmico nacional. Por exemplo, Dantas (2012) investigou os erros de escrita em inglês por brasileiros: identificação, classificação e variação entre níveis. Silveira (2012) analisou a produção escrita de aprendizes brasileiros e Gil (2014) pesquisou a incidência do princípio idiomático e do princípio da escolha aberta na produção escrita de alunos brasileiros de inglês como língua estrangeira.

O presente estudo utiliza uma metodologia da Linguística de *Corpus* (Berber-Sardinha, 2004), na vertente da Análise Multidimensional – AMD (Biber, 1988, 2009, Berber-Sardinha, Veirano-Pinto, 2014, 2019), que "usa procedimentos estatísticos (principalmente análise fatorial), visando ao mapeamento das associações entre um conjunto variado de características linguísticas dentro do *corpus* de estudo" (Berber-Sardinha, 2004, p. 300).

O *corpus* usado neste estudo, o COBRA-7 (*Corpus* Multinível de Aprendizes Brasileiros de Inglês como Língua Estrangeira – Seven Idiomas (Dantas, 2012), contém redações produzidas por aprendizes de inglês como língua estrangeira acerca de 6 registros: *Personal letter, Essay, Internet post, Formal letter, Narrative* e *Review* sobre diversos tópicos. Dessa forma, o COBRA-7 se caracteriza por ser um c*orpus* de aprendiz, uma vez que apresenta uma coleção "de textos eletrônicos produzidos por aprendizes de uma língua..." (Granger, 2002, p. 7).

Na investigação de *corpora* de aprendizes, segundo Berber-Sardinha (2004, p. 269), "a LC privilegia, em suma, a linguagem dos alunos tal qual manifestada em seus ambientes naturais, ou seja, no emprego, no entretenimento, na escola e no mundo real em geral".

Diante disso, os objetivos específicos da pesquisa são: identificar as dimensões de variação relativas aos textos do COBRA-7; identificar até que ponto as dimensões de variação explicam a variação existente entre nível de ensino, local de ensino e o gênero biológico masculino e feminino.

Dessa forma, este estudo almeja responder às questões elencadas abaixo:

1. Quais são as dimensões do *corpus* de Aprendiz COBRA 7?

2. Qual a influência de variáveis independentes como o nível de ensino, local de ensino, gênero biológico masculino e feminino na variação entre os textos do COBRA 7?

Este artigo está organizado em cinco seções. Na primeira seção, a introdução e, em seguida, serão apresentados os principais Fundamentos Teóricos desta pesquisa ao detalhar a Linguística de *Corpus*, a Linguística de *Corpus* de Aprendiz e a Análise Multidimensional.

O texto descreve a metodologia da pesquisa em duas seções. Na primeira, explica o desenho do *corpus* de estudo e os procedimentos de coleta. Na segunda, detalha a análise funcional completa, incluindo a extração das dimensões funcionais do COBRA-7. Também são apresentados os procedimentos de análise dos resultados das interpretações das duas dimensões e das análises de variância (ANOVA).

Por último, apresentaremos as considerações finais que concluem o estudo.

FUNDAMENTAÇÃO TEÓRICA

O trabalho aqui proposto tem como fundamentação teórica principal a Linguística de *Corpus* (LC) que se ocupa da coleta e da exploração de *corpora*, ou conjunto de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador (Berber-Sardinha, 2004). Assim, o objeto de estudo da LC é o que pode ser definido como um corpo de linguagem natural (autêntica) que pode ser usado como base para pesquisa linguística (Berber-Sardinha, 2004; Sinclair, 1995). De acordo com Biber, Conrad & Reppen (1994), as pesquisas baseadas em LC apresentam como características principais:

a) Ser empírica e analisar padrões reais de uso de textos naturais;
b) Utilizar uma coleção grande de textos naturais e criteriosamente selecionados, conhecida como *corpus*;
c) Fazer análises por meio de técnicas tanto automáticas, como interativas;
d) Empregar técnicas de análise ambas quantitativas e qualitativas (interpretativas).

Segundo Berber-Sardinha (2004), o objeto central do estudo da LC é o *corpus*, que pode ser definido como um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos) sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para descrição e análise (2004, pp.18-19). Esta definição é mais completa porque menciona vários pontos importantes:

- **A origem**: os dados devem ser autênticos.
- **O propósito**: o *corpus* deve ter a finalidade de ser um objetivo de estudo linguístico.
- **A composição**: o conteúdo do *corpus* deve ser criteriosamente escolhido.
- **A formatação**: os dados do *corpus* devem ser legíveis por computador.
- **A representatividade**: o *corpus* deve ser representativo de uma língua ou variedade.
- **A extensão**: o *corpus* deve ser vasto para ser representativo.

LINGUÍSTICA DE *CORPUS*

A coleta de dados linguísticos para fins de pesquisa já existia bem antes da invenção do computador. Existem dados históricos de *corpora* de citações bíblicas compiladas desde a Antiguidade e a Idade Média. Já no século XX, pode-se encontrar vários exemplos de trabalhos de pesquisadores dedicados à compilação linguística, tais como: Thorndike, Boas e Frias. No entanto, de acordo com Berber-Sardinha (2004), há duas diferenças significativas ao comparar os estudos daquela época com os estudos da LC atuais: os *corpora* daquela época não eram eletrônicos e a ênfase e destaque dados a tais estudos era, de modo geral, o ensino de línguas.

Embora os primeiros *corpora* sejam considerados limitados comparados aos *corpora* atuais, faz-se necessário ressaltar que um *corpus* não-computadorizado serviu de base para a criação de *corpora* como os existentes na atualidade. O *SEU (Survey of*

English Usage), compilado em Londres, em 1959 pela equipe de Randolf Quirk, serviu como referência para outros *corpora*, inclusive o primeiro *corpus* linguístico eletrônico *Brown University Standard Corpus of Present-Day American English*, contendo 1 milhão de palavras, considerado uma revolução para o estudo de LC na época.

Com a chegada dos computadores pessoais na década de 80, a pesquisa de LC começou a se popularizar. Foi também nessa época que teve o início o projeto COBUILD, uma parceria da Universidade de Birmingham e a editora Collins, sob o comando de John Sinclair, uns dos maiores pesquisadores em LC, sendo ainda considerado um marco para a história da LC. Juntos publicaram dicionários, gramáticas, e livros didáticos baseados na pesquisa com *corpora* para o ensino de inglês, sendo de grande valia para pesquisas subsequentes.

LINGUÍSTICA DE *CORPUS DE* APRENDIZ

Na pesquisa, a Linguística de *Corpus* (LC) desempenha um papel crucial ao descrever a linguagem de aprendizes, especialmente de inglês como língua estrangeira, usando *corpora* de textos de falantes não-nativos. De acordo com Granger (1998, 2002) apenas anos 80 e 90 com a coleta de corpora de inglês não-nativo, chamados de corpora de aprendizes (CLC), seguindo os princípios da LC. O projeto ICLE (*International Corpus Learner English*) é um exemplo notável na Linguística de Corpus de Aprendiz (LCA), liderado por Sylviane Granger, resultando em descrições mais precisas da linguagem utilizada por aprendizes.

É importante salientar que a linguagem aqui pode ser representada tanto por alunos de cursos de língua ou falantes experientes que deixaram de ter aulas de língua. Berber-Sardinha ressalta que, historicamente, "o *corpus* de aprendiz redefine o conceito original de *corpus*, que previa (na prática, não na teoria) que a linguagem permitida no *corpus* tinha de pertencer à variedade nativa" (2004, p. 265). Ao permitir a compilação de *corpora* de falantes não-nativos, várias questões de pesquisa surgem:

> Quais características linguísticas da língua-alvo são empregadas com mais (sobreuso) ou menos (subuso) frequência em comparação com falantes nativos? Qual é a extensão da influência da língua nativa (transferência) na produção de aprendizes? Em que áreas eles tendem a usar estratégias de evitação deixando de explorar a fundo o potencial da língua-alvo? Em que

> áreas eles tendem a demonstrar desempenho nativo ou não-nativo? Quais são as áreas nas quais os aprendizes de um dado país parecem necessitar de mais ajuda para desenvolver sua produção na língua-alvo? (Berber- Sardinha, 2004, p. 266).

Para Berber-Sardinha, os conceitos de subuso e sobreuso na Linguística de *Corpus* de Aprendizagem (doravante LCA) são descritivos e diagnósticos, não normativos, visando compreender a interlíngua e orientar o ensino. Esses conceitos contrastam com a ideia de erro na aprendizagem de uma segunda língua, pois a LCA reconhece erros como inevitáveis, não considerando anomalias na aquisição. Em vez disso, a LCA descreve esses desvios de forma neutra, com base na produção dos aprendizes, privilegiando a linguagem dos alunos em ambientes naturais, como trabalho, entretenimento, escola e contexto real.

Nesta pesquisa, as produções dos aprendizes são consideradas como autênticas, pois segundo as afirmações de Sinclair (1996) e Granger (2002) supracitadas, a produção escrita dos aprendizes possui um caráter verdadeiro, pois se destina a um leitor. Logo, os *corpora* de aprendizes são legítimos no sentido de que não são apenas uma coletânea de textos avulsos.

Para Granger (2002), a maioria dos *corpora* de aprendizes pertencem às seguintes categorias: (I) são monolíngues; (II) são escritos; (III) possuem amostras de linguagens produzidas por não especialistas; (IV) tendem a ser sincrônicos, ou seja, "descrevem o uso feito pelos aprendizes em um ponto no tempo" (p. 266).

CORPUS DE APRENDIZ

Ao definir um *corpus* de aprendiz[4], Granger (2002) reconhece que apesar de poder ser definido como uma coleção de dados produzidos por aprendizes, esse tipo de definição deve ser evitado porque pode fazer com que o termo seja usado para tipos de dados que de fato não são *corpora*. Segundo a autora (2002, p. 4), uma definição de *corpora* mais completa, seria baseada na definição adotada por Sinclair:

[4] Computer learner *corpora* are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardized and homogeneous way and documented as to their origin and provenance (Granger, 2002, p. 4).

> *Corpora* de aprendiz computadorizados são coleções eletrônicas de dados textuais autênticos de Língua Estrangeira ou Segunda Língua reunidos de acordo com critérios de desenho explícitos para um determinado propósito para Aquisição de Língua Estrangeira/Ensino de Língua Estrangeira. São codificados de forma padrão e homogênea assim como são documentados suas origens e procedência (Granger, 2002, p. 7).

Com base nessa definição, podemos compreender os critérios a serem utilizados na compilação de um *corpus* de aprendiz: o meio (escrito ou falado), o registro (variedade textual), tópico, tecnicismo e condições da atividade.

A COMPILAÇÃO DE UM *CORPUS* APRENDIZ

A autenticidade do material coletado é um dos critérios fundamentais na compilação de *corpora*. O que significa que o material reunido deve resultar de situações autênticas e genuínas de comunicação entre as pessoas e não de condições experimentais ou em quaisquer outras condições artificiais (Sinclair, 1996 *apud* Granger, 2002, p. 5). Em se tratando de *corpora* de aprendizes, esse termo autêntico tem vários parâmetros que variam desde dados coletados de contextos genuínos de comunicação até aos que resultam de legítimas situações de sala de aula. Uma vez que a produção de escrita de textos é uma legítima atividade de sala de aula, os *corpora* de aprendizes de produção escrita podem ser considerados autênticos (2002, p. 5).

METADADOS

De acordo com Granger (1998), na compilação de um *corpus* de aprendiz também se devem considerar as características que dizem respeito à língua e as que dizem respeito ao aprendiz. Os critérios relativos à língua são bastante similares aos usados na compilação de um *corpus* nativo: se o meio é falado ou escrito; qual a variedade textual - por exemplo, se é um texto argumentativo ou uma narrativa - ou se é uma conversa espontânea ou uma entrevista informal.

As variáveis de um *corpus* de aprendiz comumente são: idade, sexo, língua materna, nível de proficiência, tópico, contexto de aprendizagem, região, outra língua estrangeira, experiência prática. O contexto de aprendizagem distingue os alunos que

estão aprendendo inglês num país de Língua Inglesa (Inglês como Segunda Língua – ISL) ou não (Inglês como Língua Estrangeira – ILE). Assim, a experiência prática refere-se ao tempo que o aprendiz estuda inglês, ao material utilizado, e ao número de horas- aulas por semana; também, se o aluno já esteve em um país de língua inglesa (Granger, 1998, p. 6; Gil, 2017, p. 24). A tabela abaixo aponta resumidamente as principais variáveis para o desenho de um *corpus* de aprendiz, segundo Atkins & Clear (1992, p.1-16):

Tabela 1 - Critérios para compilação de um *corpus* de aprendizes

Língua	**Aprendiz**
Meio	Idade
Registro: variedade Textual	Sexo
Tópico	Língua Materna
Tecnicismo	Região
Condições da Atividade	Outra Língua Estrangeira
	Nível de proficiência
	Contexto de Aprendizagem
	Experiência Prática

Fonte: o autor, 2022. Adaptado de Atkins & Clear (1992, pp. 1-16)

ANÁLISE MULTIDIMENSIONAL

De acordo com Berber-Sardinha (2004, p. 300), a Análise Multidimensional (doravante AMD) deriva do conceito de dimensão de variação que pode ser definida como "um conjunto de traços que subjazem a um *corpus*". A AMD é definida como:

> Uma abordagem para análise de *corpus* que usa procedimentos estatísticos (principalmente análise fatorial), 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 (*part of speech tagging*) (Berber-Sardinha, 2004, p. 300).

A Abordagem Multidimensional (AMD), desenvolvida por Douglas Biber em 1988, é uma metodologia que se enquadra na vertente americana da Linguística de *Corpus* (LC) e é considerada uma contribuição significativa para o estudo da linguagem baseado em corpora eletrônicos. A AMD tem como princípio central a análise de textos, registros e tipos de textos, em vez de focar em traços linguísticos individuais.

Através da AMD é possível identificar textos e registros que estão relacionados a parâmetros situacionais ou funcionais, como formalidade/informalidade e

interatividade/não-interatividade. Por exemplo, Biber (1988, p. 115) sugere que uma das dimensões de variação é a "Oralidade versus Letramento", onde textos interativos tendem a ser menos informativos e vice-versa. Isso demonstra como a AMD oferece uma abordagem abrangente para a análise da linguagem, considerando múltiplas dimensões de variação em vez de apenas polos opostos.

DIMENSÃO DE VARIAÇÃO DA LÍNGUA INGLESA

Com essa abordagem, Biber (1988) investigou a variação da Língua Inglesa (LI) e identificou cinco dimensões a partir da análise de 67 variáveis linguísticas presentes em 481 textos que representam 23 registros da LI, disponibilizando um modelo da variação existente entre os registros que ainda não foi superado (Delfino, 2016). A Tabela 2 a seguir apresenta as dimensões identificadas em Biber (1988):

Tabela 2- Dimensões de variação da Língua Inglesa

Dimensão 1	Oralidade versus Letramento
Dimensão 2	Preocupações Narrativas versus Não Narrativas
Dimensão 3	Expressões Explícitas versus Dependentes do Contexto
Dimensão 4	Persuasão Explícita
Dimensão 5	Informação Abstrata versus Não Abstrata

Fonte: o autor, 2022. Adaptado de Biber (1988)

De acordo com Biber (1988, p. 113), uma "dimensão linguística é determinada a partir de uma correlação consistente de características linguísticas, ou seja, quando um grupo de características ocorre com frequência em textos, essas características definem uma dimensão linguística".[5]

Na AMD, conjuntos de variáveis são formados por traços linguísticos nos textos (características lexicogramaticais). Estas são cruciais para classificar textos em dimensões de variação, refletindo sua função comunicativa e situacional. Os falantes

[5] No original: "A linguistic dimension is determined on the basis of a consistent co-occurrence pattern among features. That is, when a group of features consistently co-occurs in texts, those features define a linguistic dimension".

fazem escolhas de características linguísticas para expressar ideias conforme seu propósito e função comunicativa.

Vários estudos baseados em *corpora* com a metodologia AMD já foram feitos em diversas línguas, tais como: na Língua Inglesa (Biber; 1988; Crossley; Louwerse, 2007; De Mönnik; Brom; Oostdijk, 2003; Lee, 2000), coreano (Kim; Biber, 1994), somali (Biber; Hared, 1994), nuklaelae tuvalan (Bressnier, 1988), gaélico (Lamb, 2008), espanhol (Biber; Davis; Jones; Tracy-Ventura, 2006, Parodi, 2007) e o português (Berber-Sardinha; Kauffmann; Acunzo, 2012; 2014a; 2014b).

No Brasil, a pesquisa em AMD está em crescimento. Inclui estudos sobre gêneros jornalísticos por Kauffmann (2005), variações em reportagens da revista Time por Condi de Souza (2012), linguagem em filmes americanos ao longo de décadas por Veirano-Pinto (2013), dimensões de variação de metáforas por Berber-Sardinha (2015), análise aditiva nas dimensões de Biber (1988) em letras de música de bandas de Rock por Delfino (2016) para fins didáticos de ensino de inglês, e investigação de Reality TV shows por Fonseca de Araújo (2017), também usando a AMD.

ANÁLISE DE VARIÂNCIA – ANOVA

A Análise de Variância é um procedimento estatístico que permite avaliar afirmações sobre as médias de populações. A análise verifica se há uma diferença significativa entre as médias e se os fatores exercem influência em alguma variável dependente. A análise de variância é utilizada quando se quer decidir se as diferenças das amostras são reais, ou seja, causadas por diferenças significativas nas populações observadas se as causas são decorrentes da mera variabilidade da amostra. Logo, essa análise parte do pressuposto que o acaso só produz pequenos desvios, sendo as grandes diferenças geradas por causas reais.

A ANOVA é usada para avaliar como variáveis extralinguísticas, como registro/sub-registro, impactam as variáveis linguísticas dependentes, usando índices estatísticos como F, p e R^2 gerados pelo programa SAS. A magnitude da diferença entre grupos é determinada pela razão F, sendo maior quando F é mais alto, indicando maior probabilidade de significância representada por p. Veirano-Pinto (2013) estabelece um valor de F superior a 3,35 como necessário para uma amostra significativa nessa análise.

O valor de p deve ser inferior a 0,05 para que as médias sejam significativas, indicando uma diferença real e não aleatória de até 5%. Além disso, o R^2 é importante, pois quanto mais próximo de 1, maior é a capacidade das variáveis independentes em prever a variação entre as características lexicogramaticais neste estudo.

CARACTERIZAÇÃO DAS VARIÁVEIS LINGUÍSTICAS DO COBRA-7

No estudo baseado em Biber (1988; 2009), são analisadas 128 variáveis linguísticas divididas em categorias gramaticais, semânticas e marcadores de posicionamento, que refletem opiniões e atitudes dos autores. Essas variáveis incluem verbos descritivos de ações, tanto transitivos quanto intransitivos, e substantivos abstratos relacionados a cognição e processo.

A análise gramatical-funcional deste estudo, baseada em Biber (1988; 2009), inclui variáveis como adjetivos, advérbios, verbos, conjunções, preposições, pronomes, e orações complementares e subordinadas com *"that"* e *"to"*. As tabelas que exibem estas variáveis e exemplos são adaptadas de Condi de Souza (2012) e Veirano-Pinto (2016), com dados do COBRA-7.

Os adjetivos possuem características morfológicas como flexão em grau comparativo, semânticas na descrição de qualidades e sintáticas atuando atributivamente antes de substantivos ou predicativamente após verbos de ligação, sendo os adjetivos atributivos mais comuns na linguagem oral e escrita, conforme indicado na Tabela 5.

Tabela 5 – Variáveis Linguísticas: adjetivos

Tipo	Etiqueta	Classe	Exemplos Extraídos do COBRA-7
Adjetivos atributivos	<adj_attr>	Gramatical	*The MBA is a **strong fator**.*
Adjetivos predicativo	<pre_dj>	Gramatical	*Even a change is **important.***
Adjetivos todos	<alladj>	Semântica	*I believe you will be **better.***

Fonte: o autor, 2022, adaptado de Veirano-Pinto (2013)

Advérbios, segundo Biber et al. (1999, p. 536), modificam adjetivos, verbos e outros advérbios, integrando-se tanto a elementos da oração quanto da frase. Por exemplo, *"almost"* modifica o adjetivo *"positive"* em *"I am almost [positive] she borrowed that off Barbie!"*, enquanto *"reasonably"* modifica o advérbio *"accurately"* em *"First, health*

service managers must be able to price their services reasonably [accurately] for trading purposes". A Tabela 6 do estudo detalha as subcategorizações desses advérbios.

Tabela 6 – Variáveis Linguísticas: Advérbios

Tipo	Etiqueta	Classe	Exemplos Extraídos do COBRA-7
Todos os advérbios de posicionamento	<Alladv>	marcador de posicionamento (stance)	*But, **unfortunately**, even with those solids arguments, we can't prove anything*
Advérbio enfatizador	<advs>	Gramatical	*It is **really** a difficult decision*

Fonte: o autor, 2022, adaptado de Veirano-Pinto (2013)

Biber *et al.* (1999, p. 358) explicam que os verbos têm duas funções principais: como verbos principais ou auxiliares. Verbos principais podem ser únicos em uma frase, enquanto os auxiliares, como *"be"* e *"can"*, aparecem com outro verbo principal. Verbos modais, incluindo *'can'*, *'could'*, e *'must'*, expressam significados como possibilidade ou obrigação. Além disso, os verbos podem denotar ações ou processos, como indicado na Tabela 7.

Tabela 7 – Variáveis linguísticas: Verbos

Tipo	Etiqueta	classe	Exemplos Extraídos do COBRA-7
Verbos modais de possibilidade, permissão e habilidade	<prd_mod>	gramatical	*Laughter **may** be a form of courage*
Verbos modais de necessidade e obrigação	<Pos_mod>	gramatical	*We **must** take care of our environment*
Verbo (não inclusão dos verbos auxiliares)	<Nec_mod>	gramatical	***solve** problems and **do** business.*
Contração	<contrac>	gramatical	***She's** average height*
Todas passivas	<Allpasv>	gramatical	*the way things **are done***

Fonte: o autor, 2022, adaptado de Veirano-Pinto (2013)

As conjunções, de acordo com Biber (1988), têm o propósito de indicar relações lógicas, particularmente em textos informativos. Elas se dividem em 'coordenadas', que ligam elementos com funções sintáticas e hierárquicas semelhantes em duas orações, e 'subordinadas', que introduzem orações dependentes da principal, esclarecendo relações de condição, causa, entre outras, como demonstrado na Tabela 8.

Tabela 8 – variáveis Linguísticas: Conjunções

Tipo	Etiqueta	classe	Exemplos Extraídos do COBRA-7
Todas as conjunções	<allcon>	gramatical	***but also*** *be grateful for what we took for granted*

Fonte: o autor, 2022, adaptado de Veirano-Pinto (2013)

Conforme Sinclair (1991), enquanto as conjunções conectam orações, as preposições têm a função de unir substantivos. Biber (1988) observa que as preposições fornecem informações que frequentemente se associam a nominalizações e verbos na voz passiva, especialmente em registros formais como documentos oficiais, cartas profissionais e artigos acadêmicos. Existem dois tipos de preposições: aquelas que dependem de verbos, substantivos e adjetivos, marcando tempo, lugar e movimento; e as que aparecem isoladas na frase. A Tabela 9 exemplifica o uso das preposições dependentes de verbos.

Tabela 9 – Variáveis Linguísticas: Preposições

Tipo	Etiqueta	Classe	Exemplos Extraídos do COBRA-7
Preposições	<prep>	gramatical	*I disagree* ***with*** *you*

Fonte: o autor, 2022, adaptado de Veirano-Pinto (2013)

Os pronomes de primeira e segunda pessoa, conforme Biber (1988, p. 131), indicam forte envolvimento entre os interlocutores, enquanto os de terceira pessoa estão associados a narrativas com verbos no passado e aspecto perfeito. Os pronomes indefinidos, segundo Biber *et al.* (1999), referem-se vagamente à terceira pessoa de forma genérica, como mostrado na Tabela 10.

Tabela 10 – Variáveis Linguísticas: Pronomes

Tipo	Etiqueta	classe	Exemplos Extraídos do COBRA-7
Pronome indefinido	<pany>	gramatical	*When comes the weekend* ***everyone*** *is happy*
Pronomes Possessivos de segunda pessoa e pessoais	<Pro2>	gramatical	*books that make* ***you*** *feel comfortable and happy.*
Pronome de primeira pessoa/possessivo	<Pro1>	gramatical	***I*** *wish* ***I*** *could buy a confortable car,*

Fonte: o autor, 2022, adaptado de Veirano-Pinto (2013)

As orações complementares e subordinadas introduzidas por *'that'* e *'to'* expressam opiniões e atitudes, utilizando verbos, substantivos, adjetivos e advérbios. Biber et al.

(1999, pp. 660-661) destacam que essas orações relatam pensamentos, atitudes ou emoções, como ilustrado pelo exemplo *"I was quite confident that it would say in very well"*. Estas construções estão associadas a orações de pensamento, evidenciando sua função na expressão de subjetividade. A Tabela 11, no COBRA-7, fornece exemplos dessas orações.

Tabela 11 – Variáveis Linguísticas: Orações complementares e subordinadas com *THAT* e *TO*.

Tipo	Etiqueta	Classe	Exemplos Extraídos do COBRA-7
That usado em oração complementar controlada pelo verbo	<Vcmp>	marcador de posicionamento (stance)	*I **know that all answers would be the same***
That usado em oração complementar controlada por verbo de probabilidade	<lkly_vth>	marcador de posicionamento (stance)	***[...] shows that you can't forget someone that you really love,***
Apagamento do *that*	<that_del>	marcador de posicionamento (stance)	*And I **think** everbody should watch because it*
That usado em oração complementar controlada por verbo factivo	<fact_vth>	marcador de posicionamento (stance)	*but every time **we really decide that the moment arrived** something happens*
That usado em oração complementar controlada por verbo não factivo	<nonf_vth>	marcador de posicionamento (stance)	*Using that argument would **suggest that the prison should be abolish,***
Todas as orações complementares com *to*	<all_to>	marcador de posicionamento (stance)	*I **need to go out from here!***

Fonte: o autor, 2022, adaptado de Veirano-Pinto (2013)

Três variáveis na Tabela 12, não enquadradas em categorias linguísticas, são destacadas por Biber (1988) e Veirano-Pinto (2013) como relevantes para a descrição linguística. Elas ajudam a analisar a variação em diferentes registros da língua. Especificamente, a variável 'razão entre item e ocorrência' é útil para medir a variação vocabular de textos, indicando maior variação quando seu valor se aproxima de 1.

Tabela 12 – Variáveis Linguísticas: Extras

VARIÁVEIS LINGUÍSTICAS EXTRAS
Razão entre item e ocorrência
Quantidade de palavras
Tamanho de palavras

Fonte: o autor, 2022, adaptado de Veirano-Pinto (2013)

METODOLOGIA

São apresentadas aqui, as etapas utilizadas para a realização desta pesquisa. Na primeira seção, incluem-se a descrição do *corpus*, bem como a especificação dos procedimentos de coleta do *corpus*. Na segunda seção, a análise fatorial completa, com a extração das dimensões funcionais de variação do COBRA-7.

LOCAL DE COLETA DO *CORPUS* COBRA-7

Segundo Dantas (2012), a escola analisada neste estudo tem 14 unidades, sendo 12 delas franquias localizadas em várias partes da capital de São Paulo e algumas cidades do interior do estado. Em dezembro de 2011, tinha um total de 4.708 alunos de diferentes faixas etárias. Fundada em 1987, a metodologia de ensino utilizada pela escola se baseia na teoria das inteligências múltiplas de Gardner (1985) e oferece aulas de inglês e espanhol para pessoas de diversas idades, com diversos modelos de cursos, como mostrado na Figura 2.

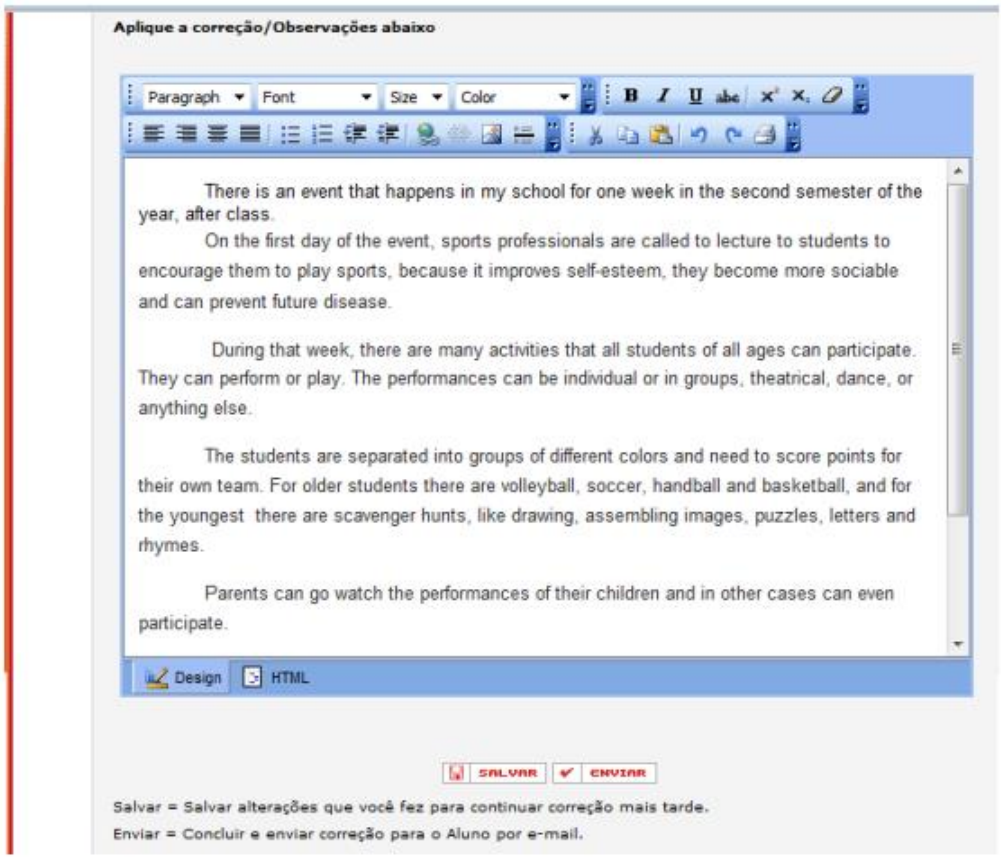
Figura 2 -Tipos de cursos oferecidos pela escola de idiomas

Adolescentes:	Aulas sempre às tardes (horários fixos[48]); a) Duas aulas semanais com duração de 1h e 15 minutos cada até o nível intermediário (*Teen*); b) Duas aulas semanais com duração de 1h e 30 minutos cada a partir do nível intermediário superior (*Teen*).
Adultos:	Aulas de manhã, à noite ou aos sábados (horários fixos); a) Duas aulas semanais com duração de 1h e 30 minutos cada (*Twice*); b) Duas aulas semanais com duração de 3 horas cada (*Vertical*); c) Aulas aos sábados com duração de 3 horas e 30 minutos (*Saturday*); d) Aula individual de 1h e 30 minutos às sextas-feiras (*Flexi*).

Fonte: o autor, 2022. As redações do corpus COBRA-7, adaptado de Dantas (2012)

Desde 2009, a escola adotou uma ferramenta de composição e armazenamento de redações semelhante ao Microsoft Word. Os professores foram instruídos a não mais exigir redações escritas em papel durante os exames, mas a postar os temas online com uma semana de antecedência. Os alunos, com nomes de usuário e senhas, acessavam o sistema chamado de e-portfólio para realizar a redação, que compunha parte da nota. As redações eram armazenadas indefinidamente no servidor, com autorização dos responsáveis legais no contrato de matrícula, conforme mostrado na Figura 3.

Figura 3 - Recorte tela: a área de composição de texto



Fonte: o autor 2022, adaptado de Dantas (2012)

Em cada nível de curso nessa escola de Idiomas os aprendizes faziam de três a quatro redações com mediação dos professores. Logo, considerando que a escola tem seis níveis de curso (Básico I, Básico II, Pré-intermediário, Intermediário, Pós-intermediário e Avançado), é possível que um aprendiz que estude nesta franquia desde o nível básico, provavelmente faça entre 18 e 24 redações.

COMPOSIÇÃO DO *CORPUS* COBRA-7

O COBRA-7 foi compilado por Dantas em 2012 e as redações de aprendizes postadas entre 2009 e 2010 foram buscadas no sistema online (e-portfolio) de acesso restrito da escola de Idiomas. Primeiramente, acessando o servidor online de acesso restrito da escola por meio de nome de usuário e senha funcional.

Em seguida, foi acessado o espaço virtual de cada uma das unidades da instituição, demonstrando o recorte da tela, a página na qual são escolhidos o ano de produção, o idioma, a situação da turma e o período, conforme a Figura 4, abaixo.

Figura 4 - ano de produção das redações do COBRA-7.

Fonte: o autor, 2022. Adaptado de Dantas (2012).

Dantas (2012) destaca a importância de explicar que a busca, seleção e cópia das redações foram necessárias porque elas estavam armazenadas em um banco de dados da escola de idiomas em que o autor trabalhava. A partir dessas redações coletadas, o autor criou o *Corpus* Brasileiro Multinível de Aprendizes de Inglês como Língua Estrangeira - Seven Idiomas (COBRA-7), que consiste em 2.516 redações escolares de diferentes tópicos, totalizando 571.564 palavras (tokens) e 19.800 tipos, distribuídas em seis níveis de proficiência: básico I (Bs1), básico II (Bs2), pré-intermediário (Pre), intermediário (Int), pós-intermediário (Hig) e avançado (Adv), conforme a Tabela 13.

Tabela 13 – Distribuição dos níveis de proficiência do COBRA-7

NÍVEL	TEXTOS
AVANÇADO (Adv)	326 (12,9%)
PÓS – INTERMEDIÁRIO (Hig)	370 (14,7%)
INTERMEDIÁRIO (Int)	471 (18,7%)
PRÉ-INTERMEDIÁRIO (Pre)	359 (14,2 %)
BÁSICO II (Bs2)	526 (20,9%)
BÁSICO I (Bs1)	464 (18,4%)
TOTAL	2516

Fonte:o autor 2022.

Além de ser um corpus de aprendizes, de acordo com os critérios tipológicos de Berber- Sardinha (2004) e Granger (1998, 2002 e 2008), pode-se definir o COBRA-7 como sendo: de extensão média, monolíngue, sincrônico, estático e contemporâneo (Dantas, 2012).

VARIÁVEIS INDEPENDENTES DA ANÁLISE DO COBRA-7

Na AMD, estimamos a quantidade de variação explicada pelas variáveis independentes por meio do procedimento de Análise de Variação (no pacote SAS, e realizado por PROC ANOVA ou PROC GLM). Identificamos as variáveis independentes relevantes e usamos uma variável de cada vez com um *fixed effect* no SAS *on demand* (PROC ANOVA ou PROC GLM. Quando usamos uma variável independente como *fixed effe*ct, estamos interessados em saber qual a variável capturada pela variável em cada dimensão. A Tabela 14 mostra as variáveis independentes empregadas na pesquisa.

Tabela 14 - Variáveis independentes da análise do COBRA-7

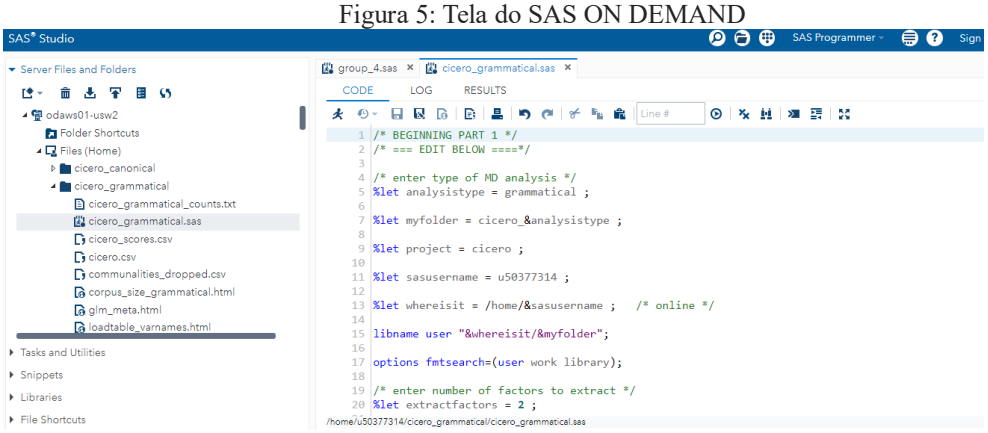
Variáveis independentes	Nome das Variáveis	Níveis das Variáveis	Número dos níveis	PROC GLM
Nível	Nível	adv bs1 bs2 hig int pre	6	*Fixed*
Escola	Escola	aclim alpha augus guaru rebou santa santo socae tatua vilal vilam vinhe	12	*Fixed*
Gênero Biológico	Gênero Biológico	f, m	2	*Fixed*

Fonte: o autor, 2022.

ANÁLISE FATORIAL DOS DADOS

Com os dados etiquetados via Biber Tagger, o passo seguinte foi a utilização de um *software* de análise fatorial para a execução dos dados, o *software* usado nesta pesquisa foi o *SAS on Demand.* Para a realização da análise fatorial, foi utilizado um programa SAS (*script* contendo uma sequência de comandos do SAS) desenvolvido por Berber-Sardinha, conforme Figura 5 abaixo.

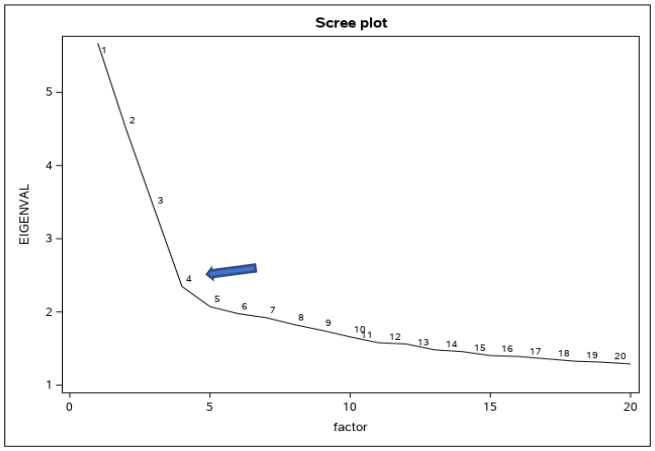
Figura 5: Tela do SAS ON DEMAND



Fonte: o autor, 2022. SAS *University Edition (2022).*

Em seguida, foi rodada a primeira análise fatorial não rotacionada, que gerou o gráfico de sedimentação, ou *screeplot* que ordena os eigenvalores (representam a quantidade de variância das variáveis contidas em cada fator; quanto maior seu valor, mas variância é revelada pelo fator). Interpretamos que a melhor solução seria a de quatro fatores conforme o gráfico de segmentação 6 abaixo:

Gráfico 6: *Scree plot*



Fonte: o autor, 2022. SAS *University Edition*

O gráfico sugere a extração de quatro fatores, com o cotovelo mais marcado na quarta iteração, indicando estabilidade após esse ponto e a ausência de diferenças significativas entre os fatores. Uma análise fatorial rotacionada foi realizada, mas apenas os dois primeiros fatores foram considerados devido à falta de diferenças analisáveis nos fatores subsequentes. Os escores de cada texto foram calculados para esses dois fatores, permitindo a interpretação das dimensões e a análise dos textos com escores mais altos em cada uma delas. A estrutura final dos fatores está detalhada nas Tabelas 15 e 16.

Tabela 15 – Variáveis do Fator 1 – Polo Positivo

Fator 1		
Polo Positivo		
Característica linguística	**Etiqueta**	**Carga**
That usado em oração complementar controlada pelo verbo	Vcmp	.59
That usado em oração complementar controlada por verbo de probabilidade	lkly_vth	.46
Todas as orações complementares com *to*	all_to	.43
Apagamento do *that*	that_del	.42
That usado em oração complementar controlada por verbo factivo	fact_vth	.41
Pronome indefinido	pany	.40
Verbo modal preditivo	Prd_mod	.40
That usado em oração complementar controlada por verbo não factivo	nonf_vth	.46
Todas as conjunções	allcon	.40
Pronomes Possessivos de segunda pessoa e pessoais	Pro2	.38
Verbos modais de possibilidade, permissão e habilidade	Pos_mod	.37
Todos os advérbios de posicionamento	All_adv	.36
Advérbio (excluindo outros tipos)	advs	.35
Verbos modais de necessidade e obrigação	Nec_mod	.33
Polo Negativo		
Sem traços negativos	-	-

Fonte: o autor, 2022.

Tabela 16 –Variáveis do fator 2. Polo Positivo e Polo Negativo

Fator 2		
Polo Positivo		
Característica linguística	**Etiqueta**	**Carga**
Tamanho da palavra	wrlength	.79
Razão forma-ocorrência	Ttr	.58
Adjetivos em posição atributiva	Adj_attr	.51
Quantidade de palavras	Wcount	.48
Todas passivas	Allpasv	.40
Preposições	Prep.	.40

Fator 2		
Polo Positivo		
Todos os adjetivos	Alladj	.32
Fator 2		
Polo Negativo		
Característica linguística	**Etiqueta**	**Carga**
Pronome de primeira pessoa/possessivo	Pro1	-.71
Verbo (não inclusão dos verbos auxiliares)	allverb	-.55
Contração	contrac	-.42

Fonte: o autor, 2022.

Desse modo, encerramos aqui a explanação da metodologia de pesquisa e partiremos para a apresentação e discussão dos resultados.

APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Serão apresentados e discutidos resultados da análise descrita na Metodologia de pesquisa, a Análise Multidimensional gramático-funcional**.** As etapas da Análise Fatorial apresentadas anteriormente, na seção Metodologia apontaram dois fatores, nos quais 24 variáveis carregaram. Nesta seção, a interpretação das duas dimensões resultantes desta análise será apresentada, assim como os resultados das análises de variância (ANOVA). A Tabela 17 ilustra as nomeações destas dimensões.

Tabela 17 - As duas dimensões do *corpus* COBRA-7

DIMENSÃO 1. Posicionamento do falante
DIMENSÃO 2. Letramento *versus* Oralidade

Fonte: o autor, 2022.

DIMENSÃO 1 – POSICIONAMENTO DO FALANTE

A primeira dimensão do COBRA-7, chamada de "Posicionamento do falante," incluiu 14 variáveis (ver Tabela 17). Esta dimensão engloba orações subordinadas *"that"* como complementos controlados por verbos não-factivos (como *"believe", "think", "say"*), incluindo a omissão do *"that"*. Os verbos não-factivos refletem a atitude do falante, como opiniões ou posicionamentos. Além disso, os verbos modais (como

"would", "can", "must", "will") indicam posicionamento, como sugestão, obrigação, predição e possibilidade. Os pronomes pessoais de segunda pessoa (como *"you", "your"*) indicam alto envolvimento entre os falantes. Um exemplo marcante dessa dimensão apresenta essas características linguísticas.

Texto 1
If God did not exist, it ***would****<Prd_mod> be necessary to invent him. This question is very important and* ***can****<Pos_mod> be difficult t to* ***answer that****< Vcmp>. I* ***believe that*** *<lkly_vth> god exists but I* ***must*** *<Nec_mod> be honest because sometimes I do not believe it. I* ***think that****<lkly_vth> believe or not depends on the situation* ***you****< Pro2> are and how is* ***your****< Pro2> feeling. If* ***you****< Pro2>* ***believe that****<lkly_vth> the situation* ***can****<pos_mod> be better and someday* ***will****< Prd_mod> be ok* ***you****< Pro2>* ***will****< prd_mod>* ***say that*** *<nonf_vth> god is good and take care about* ***you****<pro2> here.*
(**Metadados**: **número do texto**: 01369, **sequência da redação**: 1, **sexo**: m, **nível de proficiência**: avançado e **escola**: Vila Mariana).

Características Linguísticas:
That (vcmp) usado em oração complementar controlada pelo verbo (.59); ***That*** **(lkly vth)** usado em oração complementar controlada por verbo de probabilidade (.46); verbo modal preditivo **(prd mod)** (.40); verbos modais de possibilidade, permissão e habilidade **(pos_mod)** (.37); pronomes Possessivos de segunda pessoa e pessoais (Pro2) (.38); verbos modais de necessidade e obrigação **(nec_mod)** (.33).

RESULTADO DA ANOVA DA DIMENSÃO 1 – POSICIONAMENTO DO FALANTE

Para cada dimensão, apresentaremos os resultados da Análise de Variância (ANOVA). Os resultados da ANOVA (Tabela 18) para a dimensão 1 mostram que F= 6,17 e p <0, = 0001, significativo estatisticamente, porém o R^2 indica que apenas cerca de 2,6 % da variação é explicada pela variável independente "escola" (R^2 =.026408). A variável independente "nível" tem F=147,30 e p= 0001, o que é significativo, com R^2 indicando que cerca de 22,7% da variação é explicada pelo nível (R^2 =.226906). A variável "gênero" não apresentou valores de F e de R^2 notáveis.

Tabela 18 – Resultado da ANOVA da Dimensão 1

Variáveis	**F**	**p**	**R^2**
Unidade Escolar	6,17	<.0001	0,026408

Nível	147,34	<.0001	0,226906
Gênero	0,000332	0.3609	0.000332

Fonte: o autor, 2022.

DIMENSÃO 2 – LETRAMENTO VERSUS ORALIDADE

A segunda dimensão é composta por 10 variáveis, com 7 no polo positivo e 3 no polo negativo (ver Tabela 18). Essa dimensão é chamada de "Letramento versus Oralidade". No polo positivo "Letramento" há características relacionadas à expressão de informações, incluindo palavras longas como substantivos, adjetivos atributivos e preposições, indicando uma abordagem mais escrita. No polo negativo "Oralidade", apenas três variáveis se destacam, incluindo pronomes de primeira pessoa e verbos principais, sugerindo uma abordagem mais falada. Para exemplificar a dimensão 2 no polo positivo, há o "Texto 1: Letramento", e no polo negativo, o "Texto 2: Oralidade".

Polo Positivo: Letramento
Texto 1
*A 20% increase, **in**< Prep > salary may be something **important**<Alladj >that can untie **at**<Prep> the time **of**<Prep> the **final**<Adj_attr> choice. **About**<Prep> your **current**<Adj_attr> work, xxx, which does not pay well, is a **great**<Adj_attr> company **with**<Prep>potential **for**<Prep> growth. The MBA is a **Strong**<Adj_attr> factor to not get **out of**<Prep> there. It is something that many people want and try hard to get this **opportunity**<wrlength> **in**<Prep> life. Well, my opinion **about** <Prep> what you should do is: stay where you are! I believe you will be **better**<Alladj> **in**<Prep> an **environmen**t<wrlength> that you already know. A bird **in**<Prep> hand is worth more than two birds flying. Even a change is **important**<Alladj> **for**<Prep> everyone, do not risk yourself to not regret later! **Of**<Prep> course, this is only my **opinion**<wrlength>! I want the **best**<Alladj> **for**<Prep> you! But if you want changes, do it and I hope you achieve success wherever you choose to go!*
(**Metadados**: **número do texto**: 01123, **sequência da redação**: 1, **sexo**: m, **nível de proficiência**: pós-intermediário e **escola**: Tatuapé)
Características Linguísticas:
Tamanho da palavra **(wrlength**) (.79); Adjetivos em posição atributiva **(Adj_attr)** (.51); Preposições **(Prep)** (.40); Todos adjetivos **(Alladj)** (.32).

Polo Negativo: Oralidade
Texto 2
*Dreams and accomplishments **We**<Pro1> can **imagine**<allverb> and **dream**< allverb> lot of things and how many times **we want**<allverb>, because it **is**<allverb> something natural in **our**<Pro1> life. Everybody **has**<allverb> a dream and **wants**<allverb> to realize these dreams, but how can **we**<Pro1> **do**<allverb> it? This **is**<allverb> a question that everybody **wants**<allverb> to know, but to realize the dreams*

depends<*allverb*> *on you, because everything* ***is***<*allverb*> *possible when* ***we***<*Pro1*> ***believe***<*allverb*> *and* ***want***<*allverb*> *to conquer them.* ***I***<*Pro1*> ***have***<*allverb*> *a lot of dreams everyday,* ***I***<*Pro1*> *already* ***dreamed***<*allverb*> *to live in Canada, but* ***I***<*Pro1*> ***got***<*allverb*>*married, then* ***I***<*Pro1*> ***gave up***<*allverb*>. ***I***<*Pro1*> ***dreamed***<*allverb*> *to get a surgery of myopia and* ***I***<*Pro1*> ***realized***<*allverb*> *it, but* ***I***<*Pro1*> ***had***<*allverb*> *to go ahead.*
(**Metadados**: **número do texto**: 00995, **sequência da redação**: 4, **sexo**: f, **nível de proficiência**: avançado e **escola**: Santo André).

Características Linguísticas:
Pronome de primeira pessoa/possessivo **(Pro1)** (-.71); Verbo (não inclusão dos verbos auxiliares **(allverb)** (-.55)

RESULTADO DA ANOVA DA DIMENSÃO 2 – LETRAMENTO *VERSUS* ORALIDADE

Para a dimensão 2, os resultados da ANOVA (Tabela 19) indicam variância relativamente significativa, com F= 5.01 e p < 0, = 0001. Contudo, o R^2 indica que apenas cerca de 2% da variação é explicada pela variável independente "escola" (R^2=.021534). Já a variável independente "nível" também é significativa, com F= 3.33 e p<0, = 0001 e R^2 indica que cerca de 19,9% da variação é explicada pelo nível do aluno (R^2=.198677). A variável "gênero" não apresentou valores de F e R^2 notáveis.

Tabela 19 - Resultado da ANOVA da Dimensão 2

Variáveis	F	p	R^2
Unidade Escolar	5,01	<.0001	0,021534
Nível	3,33	<.0001	0,198677
Gênero	0,0680	0,3609	0,001324

Fonte: o autor, 2022.

RELAÇÃO ENTRE AS DIMENSÕES

Tabela 20 – Médias do Fator 1

Nível	Média	Desvio padrão
AVANÇADO (Adv)	2,70	5,76
PÓS – INTERMEDIÁRIO (Hig)	2,95	5,74
INTERMEDIÁRIO (Int)	1,49	6,63
PRÉ-INTERMEDIÁRIO (Pre)	-0,57	5,37
BÁSICO II (Bs2)	-0,14	6,54
BÁSICO I (Bs1)	-5,16	4,22

Fonte: o autor, 2022.

Tabela 21 – Médias do Fator 2

Nível	Média	Desvio padrão
AVANÇADO (Adv)	3,92	4,49
PÓS – INTERMEDIÁRIO (Hig)	1,67	5,06
INTERMEDIÁRIO (Int)	1,51	4,94
PRÉ-INTERMEDIÁRIO (Pre)	-0,30	4,64
BÁSICO II (Bs2)	-2,34	4,72
BÁSICO I (Bs1)	-3,20	5,26

Fonte: o autor, 2022.

Na Análise Multidimensional Gramático-Funcional, os resultados destacam que a variável "nível" foi a principal responsável pela variação explicada. As médias dos níveis de proficiência nos fatores 1 e 2 (Tabelas 20 e 21) indicam que o Fator 1 tem a média mais alta no nível Pós-intermediário (Hig) (2,95), seguido pelo nível Avançado (Adv) (2,70) e, por último, o nível Intermediário (Int) (1,49). Vale observar que alguns alunos são falsos iniciantes, enquanto outros possuem mais conhecimento de mundo e linguístico. Além disso, variáveis como tópico, contexto de aprendizado e experiência prática também podem influenciar as opiniões dos falantes. Por outro lado, as médias dos níveis no Fator 2 (Tabela 21) revelam diferenças mais notáveis no nível intermediário, indicando que à medida que os aprendizes progridem nos níveis, eles utilizam mais recursos linguísticos associados às dimensões.

CONSIDERAÇÕES FINAIS

A pesquisa apresentada neste artigo teve como objetivo principal identificar as dimensões de variação subjacentes nas escritas dos aprendizes de Inglês como Língua Estrangeira por meio do *corpus* COBRA-7. Foram elencadas duas perguntas de pesquisa: 1- Quais são as dimensões do *corpus* de Aprendiz COBRA 7? 2- Qual a influência de variáveis independentes como nível de ensino, local de ensino, gênero biológico masculino e feminino na variação entre os do COBRA-7?

Os resultados mostraram duas dimensões de variação do COBRA-7 nomeadamente “Posicionamento do falante” (dimensão 1), expressando uma atitude por parte do falante, ou seja, uma opinião ou um posicionamento em relação ao tópico da redação. Em contraponto, a dimensão 2 nomeada de “Letramento *versus* oralidade” revelou que a expressão de informação e a expressão de interação são usadas pelos alunos.

A dimensão 1 reflete que os falantes que expressam interação utilizam pronomes de segunda pessoa, verbos modais e orações complementares, indicando presença de posicionamento em seus discursos. Já a dimensão 2 mostra que os textos apresentam traços linguísticos frequentes, com registros informativos no polo positivo e características mais orais no polo negativo.

As médias de cada nível revelaram que na Dimensão 1, a média mais alta ocorre no nível Pós-intermediário (Hig) (2,95), em seguida, no nível Avançado (Adv) (2,70) e, por último, no nível Intermediário (Int) (1,49). Isso mostra que depois do nível básico, os alunos passam a usar linguagem de cunho informacional, com características marcantes da escrita letrada, enquanto os alunos de nível básico dão preferência à linguagem de cunho interacional, com características marcantes de oralidade. As médias dos níveis do Fator 2 mostraram também que há uma distinção entre os níveis básicos e os demais, já que a partir do nível intermediário, os alunos passam a se valer da expressão marcada pelo posicionamento frente às questões enfocadas nas redações.

A pesquisa revelou que há uma diferenciação estatisticamente significativa entre os níveis de ensino em relação a duas principais funções comunicativas: "letramento versus oralidade" e "posicionamento do falante". No entanto, a variação atribuída ao nível de ensino, embora seja a mais influente, ainda é relativamente baixa, cerca de 20%, indicando que os níveis de ensino são permeáveis, abrangendo alunos com diferentes habilidades linguísticas. Isso pode ser resultado da progressão das habilidades linguísticas dos alunos ou das exigências específicas das tarefas de redação. Além disso, a interação entre o progresso da aprendizagem e o efeito da tarefa pode desempenhar um papel. Pesquisas futuras devem se concentrar nessas hipóteses levantadas.

Dessa forma, espera-se que, professores inquietos e incomodados com a intuição presente nos materiais didáticos, como é o caso do pesquisador, não priorizem apenas as regras gramaticais da língua que soam tão *"bookish*" e *"pedantic"* (Granger,1988, p. 49). Além do mais, espera-se que esta pesquisa estimule o desenvolvimento de outros estudos de Análise Multidimensional (AMD), principalmente na Lingúística de *Corpus* de Aprendiz no que se refere à variação linguística e aos diferentes registros da linguagem em uso.

REFERÊNCIAS

ACUNZO, Cristina Mayer. *O que e como escrevemos na web:* um estudo multidimensional de variação de registro em Língua Inglesa. 2018, 129 f. Tese de Doutorado em Linguística Aplicada e Estudos da Linguagem – LAEL - PUC. São Paulo, 2018.

ATKINS, Sue.; CLEAR, Jeremy. *Corpus design criteria*. Literary and Linguistic Computing 7(I): 1-16, 1992.

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

BERBER-SARDINHA, Tony. Como usar a Linguística de Corpus no ensino de língua estrangeira. Por uma Linguística de Corpus educacional brasileira. *In*: VIANA, Vander; TAGNIN, Stella. Corpora *no ensino de línguas estrangeiras.* São Paulo: Editora Contexto. 2013, pp. 293-348.
.

BERBER-SARDINHA, Tony. *Linguística de Corpus:* histórico e problemática *D.E.L.T.A.*, Vol. 16, N º 2, pp. 323 – 367, 2000.

BERBER-SARDINHA, Tony. A abordagem Metodológica da Análise Multidimensional, *Gragoatá*, Niterói, n. 29, pp. 107-125, 2. Sem. 2010.

BERBER-SARDINHA, Tony. Register Variation in metaphor: a multi-dimensional perspective. *In*: BERBER-SARDINHA, T.; HERRMANN, B. (Eds) metaphor specialist discourse. Amsterdam/Philadephia: John Benjamins, 2015, pp. 17-52.

BERBER-SARDINHA, Tony; KAUFFMAN, Carlos; ACUNZO, Cristina Mayer. "A Multi-dimensional analysis of register variation in Brazilian Portuguese". *Corpora* (2014): 239-271.

BESNIER, Nico. *The linguistics relationship of spoken and written Nukulaelae register*. Language, v. 64, n. 4, 1988, p. 707-736.

BIBER, Douglas *et al*. *Grammar of Spoken and Written English*. London, Pearson. 1999.

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi. *Corpus* Linguistics: investigating language structure and use. Cambridge University Press. 1994.

BIBER, Douglas. *The Cambridge handbook of English Corpus Linguistics*. Cambridge University Press. 2005.

BIBER, Douglas. *Variation across Speech and Writing. Cambridge University Press*, 1988.

BIBER, Douglas.; CONRAD, Susan. *Register, Genre and Style*. Cambridge; New York: Cambridge University Press, 2009.

BIBER, Douglas.; TRACY-VENTURA, Nicole. Dimensions of register variation in Spanish. *In*: PARODI, Giovanni. (Org.). *Working with Spanish Corpora*. London: Continuum, 2007. p. 54-89.

BRITISH COUNCIL. *Quadro Comum Europeu de Referência para Línguas (CEFR)*. Disponível em: <https://www.britishcouncil.org.br>. Acesso em 4 dez. 2021.

CHOMSKY, Noam. *The Logical Structure of Linguistic Theory*. Massachusetts: M.I.T. Library, 1955.

SOUZA, Renata Condi de. *A revista Time em uma perspectiva multidimensional*. 2012. 330 f. Tese de Doutorado em Linguística Aplicada e Estudos da Linguagem – LAEL-PUC. São Paulo, 2012.

CROSSLEY, Scott; LOUWERSE, Max. Multi-dimensional register classification using bi-grams. *International Journal of Corpus Linguistics*, v. 12, n. 4, p. 453-478, 2007.

DANTAS, Wendel Mendes. *Erros de escrita em Inglês por brasileiros:* Identificação, classificação e variação entre níveis. 2012. 164 f. Dissertação de Mestrado em Linguística Aplicada e Estudos da Linguagem- LAEL, PUC, São Paulo, 2012.

DELFINO, Maria Claudia Nunes. *Uso de música para o Ensino de Inglês como Língua Estrangeira em um ambiente baseado em* Corpus. 2016. 159 f. Dissertação de Mestrado em Linguística Aplicada e Estudos da Linguagem - LAEL, PUC. São Paulo, 2016.

DE MÖNNINK, Inge *et al.* Using the MF/MD method for automatic text classification. *In:* GRANGER, SYLVIANE.; PETCH TYSON, STEPNANIE. (Org.). *Extending the scope of corpus based research*: new applications new challenges. Amsterdam: Rodopi, 2003. p. 15-25.

OLIVEIRA, Marcos Roberto de. *Aquisição de terceira pessoa do singular na língua inglesa*: Uma perspectiva da linguística de *corpus*. 2020. 82 f. Dissertação de Mestrado em Letras – Universidade Federal de São Paulo, Escola de Filosofia, Letras e Ciências Humanas, 2020.

ARAÚJO, Rafael Fonseca de. *A linguagem dos reality TV shows norte-americanos*: análise e classificação. 2017. 242 f. Dissertação de Mestrado em Linguística Aplicada e Estudos da Linguagem – LAEL, PUC. São Paulo, 2017.

GIL, Cristina Borges. *A incidência do princípio idiomático e do princípio da escolha aberta na produção escrita de alunos brasileiros de inglês como língua estrangeira*. 2014. 102 f. Dissertação de Mestrado em Linguística Aplicada e Estudos da Linguagem - LAEL. PUC, São Paulo, 2014.

GRANGER, Sylviane *et al*. *International Corpus of Learner English*, version 3. Presses Universitaires de Louvain, Belgium, 2020.

GRANGER, Sylviane. *Learner English on Computer.* Longman London and New York, 1998.

GRANGER, Sylviane; GILQUIN, Gaetanelle; MEUNIER, Fanny. *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press, United Kingdom, 2015.

GRANGER, Sylviane; HUNG, Joseph; PETCH-TYSON, Stephanie. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. John Benjamins B. V., 2002.

HALLIDAY, Michael Alexander Kirkwood. *An introduction to Functional Grammar*, 3rd ed. Hodder Arnold. 2004.

HUNSTON, Susan. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002.

KAUFFMANN, Carlos. *O* corpus *do jornal*: variação lingüística, gêneros e dimensões da imprensa diária escrita. 2005. 203 f. Dissertação de Mestrado em Linguística Aplicada e Estudos da Linguagem - LAEL, PUC, São Paulo, SP, 2005.

KIM, Young Jin.; BIBER, Douglas. A corpus-based analysis of register variation in Korean. *In*: BIBER, Douglas.; FINEGAN, Edward. (Org.). *Sociolinguistic perspectives on register*. Oxford: Oxford University Press, 1994. pp. 157-181.

KRASHEN, Stephen, *Second Language Acquisition and second Language learning*. New York: Pergamon Press, 1981.

LAMB, William. *Scottish Gaelic speech and writing: register variation in an endangered language.* Belfast: Cló Ollscoil na Banríona, 2008.

LEE, David Yue-Wei. *Modelling variation in spoken and written language:* the Multi-Dimensional Approach revisited. 2000. Unpublished Doctoral Dissertation. – Department of Linguistics and Modern English Language- Lancaster University, Reino Unido, 1999.

LEWIS, Michael (Ed.). T*eaching Collocation: Further Developments in the Lexical Approach*. Boston: Thomson Heinle, 2000.

PARODI, Giovanni. Variation across registers in Spanish: Exploring the El-Grial PUCV Corpus. In: PARODI, GIOVANNI. (Org.). *Working with spanish corpora.* London: Continuum, 2007. pp. 11-53.

PARTINGTON, Alan. *Patterns and Meanings:* Using Corpora for English Language Research and Teaching. Amsterdam: John Benjamins, 1998.

SILVEIRA, Gustavo Estef Lino da. A linguística de corpus e o ensino de inglês*:* um estudo de produção escrita de aprendizes brasileiros. *Palimpsesto – A revista do corpo*

discente do Programa de Pós-graduação em Letras da UERJ, ano 11, n. 15, p. 2-30, 2012.

SINCLAIR, John Mchardy. *Corpus*, Concordance, Collocation. Oxford University Press. 1991.

SINCLAIR, John Mchardy. How to use *corpora* in Language Teaching. John Benjamins B.V. 2004.

SINCLAIR, John Mchardy. Trust the Text. Language, *corpus*, and discourse. Routledge. 2004.

TARONE, Elaine. Interlanguage. Elsevier ldt., v.4, p. 1715–1719, 1994.

VEIRANO-PINTO, Marcia. *A linguagem dos filmes norte-americanos ao longo dos anos*: uma Abordagem Multidimensional. 2013. 489 f. Tese de Doutorado em Linguística Aplicada e Estudos da Linguagem – LAEL, PUC. São Paulo, 2013.

Recebido em: 28/10/2023
Aceito em: 04/01/2024

Cícero Soares da Silva: Doutorando em Linguística Aplicada e Estudos da Linguagem pela PUC/SP, Mestre em Linguística Aplicada e Estudos da Linguagem pela PUC/SP (Pontifícia Universidade Católica de São Paulo), Licenciado em Letras - Língua Portuguesa pela Universidade Anhembi Morumbi (2019). Possui o CELTA (Certificate in Teaching English to Speakers of Other Languages) da Universidade de Cambridge. Leciona Inglês há mais de 25 anos no Ensino Fundamental, Médio, e, principalmente, em empresas e cursos particulares com ênfase em Inglês para fins Específicos para todos os níveis de proficiência: A1, A2, B1, B2, C1 e C2, com base no Quadro Comum Europeu de Referência para Línguas (CEFR), habilitado em cursos preparatórios para os Exames da Universidade de Cambridge (KEY, PET, FCE, CAE, CPE, IELTS E LINGUASKILL), TOEIC E TOEFL IBT. Membro do grupo de pesquisa em Linguística de Corpus -GELC.