CRISTINA BORGES GIL

O USO DE N-GRAMAS DE CLASSE SEMÂNTICA EM UM *CORPUS* DE APRENDIZ

DOUTORADO EM LINGUÍSTICA APLICADA E ESTUDOS DA LINGUAGEM

PUC – SP
2024

CRISTINA BORGES GIL

O USO DE N-GRAMAS DE CLASSE SEMÂNTICA EM UM CORPUS DE APRENDIZ

Doutorado em Linguística Aplicada e Estudos da Linguagem

Tese apresentada à banca examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de Doutor em Linguística Aplicada e Estudos da Linguagem sob orientação do Professor Doutor Antonio Paulo Berber Sardinha.

SÃO PAULO
2024

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Número do Processo: 88887.286116/2018-00

Para Edith e Tomás
Para Celila (*in memoriam*) e Izabel (*in memoriam*)

AGRADECIMENTOS

Em primeiro lugar, ao meu orientador, Prof. Dr. Tony Berber Sardinha, pela confiança e paciência, pelo seu profissionalismo e seriedade, pelas suas orientações.

À Capes, pela concessão da bolsa.

Às professoras Deise Dutra, UFMG, e Ana Bocorny, UFRGS, pelas observações, comentários e sugestões.

Aos professores e funcionários do LAEL, em especial à querida Malu, pela prestimosa ajuda, sempre.

Ao GELC e grupo de orientandos do Prof. Dr. Tony Berber Sardinha, pelo apoio e pelas contribuições que muito acrescentaram a este trabalho, em especial os queridos Aline Zamboni, Andressa Costa, Cícero Soares da Silva, Cláudia Delfino, Carlos Kauffmann, Katherine Oliva Ortholani e André Luiz Siqueira Alencar (*in memoriam*). E também a querida colega da matemática, Ana Luiza Ozores.

“I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way.”

Charles J. Fillmore

RESUMO

O objetivo deste trabalho é analisar o uso de n-gramas de classe semântica (BERBER SARDINHA, 2023) na produção escrita e oral de aprendizes de inglês como língua estrangeira. Com esta pesquisa, avaliamos se variação no uso dos n-gramas de classe semântica pode ser explicada pelo fato de o texto ser escrito ou falado, pela tarefa atribuída ao aprendiz, pelo seu nível de proficiência, pela sua língua materna, pela sua idade ou pelos anos de estudo do idioma inglês. O *corpus* empregado neste estudo foi o COREFL, cujo acrônimo significa Corpus de Inglês como Língua Estrangeira (*Corpus of English as a Foreign Language*). Primeiramente, o *corpus* foi etiquetado com o USAS, um etiquetador semântico. Em seguida, foram extraídos e selecionados os n-gramas de classe semântica e calculada a sua chavicidade. Com essas variáveis foi feita uma análise fatorial, procedimento padrão da Análise Multidimensional, e os fatores interpretados. Observamos que a tarefa e o modo desempenham um papel importante na variação dos n-gramas de classe semântica utilizados pelos aprendizes.

Palavras-chave: Linguística de *Corpus,* Linguística de *Corpus* de Aprendiz; Chavicidade; Análise Multidimensional; N-Gramas de Classe Semântica

ABSTRACT

The aim of this paper is to analyze the use of semantic class n-grams (BERBER SARDINHA, 2023) in the written and oral production of learners of English as a foreign language. With this research, we evaluated whether variation in the use of semantic class n-grams can be explained by the fact that the text is written or spoken, by the task assigned to the learner, by their level of proficiency, by their mother tongue, by their age or by the years studying the English language. The corpus used in this study was COREFL, whose acronym stands for Corpus of English as a Foreign Language. First, the corpus was tagged with USAS, a semantic tagger. Next, the semantic class n-grams were extracted and selected and their keyness calculated. A Factor Analysis was carried out on these variables, a standard procedure for Multidimensional Analysis, and the factors were interpreted. We observed that the task and the mode play an important role in the variation of the semantic class n-grams used by the learners.

Keywords: Corpus Linguistics; Learner Corpus Research; Keyness; Multidimensional Analysis; Semantic Class N-Grams.

SUMÁRIO

INTRODUÇÃO **1**

1. FUNDAMENTAÇÃO TEÓRICA **4**

1.1 LINGUÍSTICA DE *CORPUS* 4
1.2 ANÁLISE MULTIDIMENSIONAL 9
1.3 LINGUÍSTICA DE CORPUS DE APRENDIZ 15
1.3.1 Corpus de aprendiz: definição *18*
1.3.2 A compilação de um corpus de aprendiz *19*
1.4 PACOTES LEXICAIS E NGCS 20
1.5 CATEGORIAS SEMÂNTICAS E O ETIQUETADOR USAS 25

2. METODOLOGIA **27**

2. 1 COREFL – *CORPUS OF ENGLISH AS A FOREIGN LANGUAGE* 27
2.2 ANOTAÇÃO SEMÂNTICA, EXTRAÇÃO E SELEÇÃO DOS NGCS 31
2.3 CÁLCULO DA CHAVICIDADE DOS NGCSS 40
2.4 ANÁLISE FATORIAL 42

3. RESULTADOS E DISCUSSÃO **48**

3.1 FATOR 1 48
3.1.1 ANOVA da Dimensão 1 ..... *58*
3.2 FATOR 2 61
3.2.1 ANOVA da Dimensão 2 ..... *70*
3.3 FATOR 3 73
3.3.1 ANOVA da Dimensão 3 ..... *82*
3.4 DISCUSSÃO 86

4. CONSIDERAÇÕES FINAIS **88**

REFERÊNCIAS **90**

ANEXOS **98**

LISTA DE TABELAS

TABELA 1: SEQUÊNCIAS DE PALAVRAS QUE COMPARTILHAM OS SIGNIFICADOS DAS CATEGORIAS SEMÂNTICAS 2

TABELA 2: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS QUE COMPARTILHAM OS MESMOS NGCS 2

TABELA 3: CARACTERÍSTICAS DO CORPUS COREFL 27

TABELA 4: CARACTERÍSTICAS DO CORPUS PILOTO DO COREFL (L1 ESPANHOL) ... 29

TABELA 5: TAREFAS INCLUÍDAS NO CORPUS COREFL 30

TABELA 6A: RESULTADOS EM DOIS TESTES DIFERENTES DE LÍNGUA..... 46

TABELA 6B: COMPARAÇÃO DOS ESCORES Z PARA OS TESTES A E B..... 46

TABELA 7: EXEMPLO DE ESCORES DE FATORES EM CADA TEXTO 47

TABELA 8: PADRÃO FATORIAL DO FATOR 1..... 48

TABELA 9: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 1..... 52

TABELA 10: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 1..... 52

TABELA 11: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 1..... 53

TABELA 12: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 1..... 54

TABELA 13: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 1..... 54

TABELA 14: DISTRIBUIÇÃO POR PROFICIÊNCIA, MODO E LÍNGUA MATERNA DOS 50 TEXTOS QUE MAIS CARREGARAM NO FATOR 1..... 55

TABELA 15: DISTRIBUIÇÃO POR IDADE E ANOS DE ESTUDO DA LÍNGUA INGLESA DOS 50 TEXTOS QUE MAIS CARREGARAM NO FATOR 1 55

TABELA 16: RESULTADO DA ANOVA DA DIMENSÃO 1..... 58

TABELA 17: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL NÍVEL DE PROFICIÊNCIA NA DIMENSÃO 1..... 59

TABELA 18: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL ANOS DE ESTUDO DA LÍNGUA INGLESA NA DIMENSÃO 1 59

TABELA 19: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL IDADE NA DIMENSÃO 1 60

TABELA 20: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL TAREFA NA DIMENSÃO 1..... 60

TABELA 21: PADRÃO FATORIAL DO FATOR 2 61

TABELA 22: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 2 64

TABELA 23: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 2 64

TABELA 24: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3 65

TABELA 25: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3 65

TABELA 26: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3 66

TABELA 27: DISTRIBUIÇÃO POR PROFICIÊNCIA, MODO E LÍNGUA MATERNA DOS 50 TEXTOS QUE MAIS CARREGARAM NO FATOR 2 67

TABELA 28: DISTRIBUIÇÃO POR IDADE E ANOS DE ESTUDO DA LÍNGUA INGLESA DOS 50 TEXTOS QUE MAIS CARREGARAM NO FATOR 2 67

TABELA 29: RESULTADO DA ANOVA DA DIMENSÃO 2 70

TABELA 30: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL ANOS DE ESTUDO DE INGLÊS NA DIMENSÃO 2 71

TABELA 31: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL IDADE NA DIMENSÃO 2 71

TABELA 32: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL L1 NA DIMENSÃO 2 . 72

TABELA 33: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL NÍVEL DE PROFICIÊNCIA NA DIMENSÃO 2 72

TABELA 34: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL TAREFA NA DIMENSÃO 2 72

TABELA 35: PADRÃO FATORIAL DO FATOR 3 73

TABELA 36: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3 75

TABELA 37: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3 76

TABELA 38: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3 76

TABELA 39: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3 77

TABELA 40: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3 78

TABELA 41: DISTRIBUIÇÃO POR PROFICIÊNCIA, MODO DE PRODUÇÃO E L1 DOS 50 TEXTOS QUE MAIS CARREGARAM NO FATOR 3 79

TABELA 42: DISTRIBUIÇÃO POR IDADE E ANOS DE ESTUDO DA LÍNGUA INGLESA DOS TEXTOS QUE MAIS CARREGARAM NO FATOR 3 79

TABELA 43: DISTRIBUIÇÃO POR TAREFA E NÍVEL DE PROFICIÊNCIA DOS TEXTOS QUE MAIS CARREGARAM NO FATOR 3 80

TABELA 44: RESULTADO DA ANOVA DA DIMENSÃO 3..... 83

TABELA 45: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL L1 NA DIMENSÃO 3 . 83

TABELA 46: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL MODO NA DIMENSÃO 3 83

TABELA 47: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL NÍVEL DE PROFICIÊNCIA NA DIMENSÃO 3..... 84

TABELA 48: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL TAREFA NA DIMENSÃO 3..... 84

TABELA 49: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL IDADE NA DIMENSÃO 3 85

TABELA 50: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL ANOS DE ESTUDO DE INGLÊS NA DIMENSÃO 3..... 85

LISTA DE FIGURAS

FIGURA 1: CAMPOS SEMÂNTICOS DO ETIQUETADOR USAS 25

FIGURA 2: CATEGORIAS E SUBCATEGORIAS DO CAMPO SEMÂNTICO *TIME* NO USAS 26

FIGURA 3: LISTA DOS SUBCORPORA 31

FIGURA 4: CATEGORIAS E SUBCATEGORIAS DO ETIQUETADOR USAS 33

FIGURA 5: EXEMPLO DOS VALORES OBTIDOS PARA O CÁLCULO DA CHAVICIDADE DE UM NGCS 41

FIGURA 6: GRÁFICO DE SEDIMENTAÇÃO DA SOLUÇÃO NÃO ROTACIONADA COM AUTOVALORES 44

FIGURA 7: NUVEM DE PALAVRAS COM AS ETIQUETAS SEMÂNTICAS QUE CARREGARAM NO FATOR 1 50

FIGURA 8: NUVEM DE PALAVRAS COM AS PALAVRAS PRESENTES NAS ETIQUETAS SEMÂNTICAS DO FATOR 1 51

FIGURA 9: NUVEM DE PALAVRAS COM AS ETIQUETAS SEMÂNTICAS QUE CARREGARAM NO FATOR 2 62

FIGURA 10: NUVEM DE PALAVRAS COM AS PALAVRAS PRESENTES NAS ETIQUETAS SEMÂNTICAS DO FATOR 2 63

FIGURA 11: NUVEM DE PALAVRAS COM AS ETIQUETAS SEMÂNTICAS QUE CARREGARAM NO FATOR 3 74

FIGURA 12: NUVEM DE PALAVRAS COM AS PALAVRAS PRESENTES NAS ETIQUETAS SEMÂNTICAS DO FATOR 3 75

INTRODUÇÃO

Desde fins da década de 1980, estuda-se a linguagem produzida por aprendizes utilizando os princípios, ferramentas e métodos da Linguística de *Corpus* (doravante LC) (GRANGER, 2002, p. 1). Granger (2002) destaca que essa nova área de pesquisa linguística, denominada Linguística de *Corpus* de Aprendiz, estabeleceu uma relevante conexão entre a LC e a pesquisa sobre a aquisição e a aprendizagem de uma língua estrangeira/ segunda língua.

Berber Sardinha (2004) salienta que o estudo de *corpus* de aprendizes redefiniu o "conceito original de *corpus*, que previa (na prática, não na teoria) que a linguagem permitida no *corpus* tinha de pertencer à variedade nativa" (idem, p. 265).

Dentre as pesquisas em Linguística de *Corpus* de Aprendiz, podemos destacar a de Paquot e Granger (2012). As autoras fazem uma síntese das pesquisas de fraseologia em *corpora* de aprendiz e salientam que "o tipo de sequência formulaica frequentemente mais estudada em pesquisa de *corpus* de aprendiz são as colocações verbo – substantivo"[1] (2012, p.7)[2]. Na conclusão do seu artigo, chamam a atenção para importantes áreas de pesquisa que são pouco exploradas, dentre elas a comparação entre a fraseologia da produção escrita e a da produção oral de aprendizes. Em outro artigo, Granger e Bestgen (2014) discutem o uso de colocações em textos de alunos não-nativos intermediários versus avançados analisando o uso de bigramas (pares de palavras diretamente adjacentes[3]). Observaram nesse estudo que o uso de colocações de aprendizes intermediários ou avançados é caracterizado por uma mistura de colocações de alta e baixa frequência, sendo que a primeira é mais característica de alunos intermediários e a segunda de alunos avançados. Shin, Cortes e Yoo (2018) analisaram o uso de pacotes lexicais (*lexical bundles*) que "incluem artigos definidos na sua estrutura" [4] (2018, p.1) na produção escrita acadêmica de aprendizes coreanos da língua inglesa, tais como, *on the other hand* (por outro lado), *for the purpose of* (para fins de). Pacotes lexicais foram usados "como uma ferramenta para identificar e analisar o uso do artigo definido por aprendizes"[5] (2018, p. 36). As autoras destacam que o papel dos erros no uso de artigos em pacotes lexicais deve ser mais bem explorado uma vez que dão informações relevantes a respeito de "áreas problemáticas na formação de pacotes lexicais, com implicações práticas e pedagógicas"[6] (idem, ibidem). Concluem apontando para a necessidade de mais pesquisa no uso de artigos em pacotes lexicais e observam que um dos fatores que talvez tenha limitado e influenciado o uso de pacotes lexicais pelos

[1] "(…) *the most frequently studied type of formulaic sequence in learner corpus research is verb-noun collocations* (…)"
[2] Tradução minha. Doravante todas as traduções serão feitas pela autora.
[3] "(…) *directly adjacent word pairs."*
[4] "(…) *include definite articles in their internal structure* (…)"
[5] "(…) *as a tool to identify and analyze definite article usage by L2 learners."*
[6] "(…) *problematic areas in the formation of LBs, with practical and pedagogical implications."*

aprendizes possa ter sido os tópicos dos ensaios.

A pesquisa aqui descrita tem como objetivo principal fazer uma investigação descritiva da linguagem de aprendiz baseada em *corpora.* Para tal, analisaremos o uso de n-gramas de classe semântica (doravante NGCSs) na produção escrita e oral de aprendizes de inglês como língua estrangeira. Um NGCS é um agrupamento de classes semânticas diretamente adjacentes. Uma classe semântica é uma etiqueta que se refere à classificação do sentido das palavras. Será explicada mais detalhadamente na metodologia. Com esta análise pretendemos avaliar o quanto da variação desses NGCSs pode ser explicada pelo fato de o texto ser escrito ou falado, pela tarefa designada ao aluno, pelo seu nível de proficiência, pela sua língua materna, pela sua idade, ou pelos anos estudando a língua inglesa.

Esta investigação encontra suporte teórico na Linguística de *Corpus* (doravante LC), na Análise Multidimensional e na Linguística de *Corpus* de Aprendiz. A LC é uma área de pesquisa da linguagem que coleta e analisa quantitativa e qualitativamente *corpora*, que são cuidadosamente compilados de modo a responder a uma pergunta de pesquisa a respeito de uma língua ou variedade linguística, por meio de evidências empíricas extraídas por computador (McENERY; HARDIE, 2012; TOGNINI-BONELLI, 2010; BIBER; CONRAD; REPPEN, 1998; BERBER SARDINHA, 2004).

A pesquisa de NGCSs foi introduzida por Berber Sardinha (2023). Essa metodologia agrupa sequências de palavras que compartilham as mesmas categorias semânticas, definidas pela tipologia semântica do USAS, UCREL *Semantic Analisys System* (ARCHER; WILSON; RAYSON, 2002). Cada categoria semântica representa um conceito ou um significado. Por exemplo, ‘Z5 Q2 Z5’ e ‘Z5 N3 E6’ são dois n-gramas de classe semântica. Z5, Q2, A9, N3 e E6 são categorias semânticas. Abaixo, na Tabela 1, são mostrados dois exemplos de sequências de palavras que compartilham os significados dessas categorias.

TABELA 1: SEQUÊNCIAS DE PALAVRAS QUE COMPARTILHAM OS SIGNIFICADOS DAS CATEGORIAS SEMÂNTICAS

Z5 Q2 Z5: *to talk about* (falar sobre)	**Z5 N3 E6: *a growing concern* (uma preocupação crescente)**
Z5: *Grammatical bin* (Caixa grammatical)	Z5: *Grammatical bin*
Q2: *Speech acts* (Atos de fala)	N3: *Measurement* (Medições)
A9: *Grammatical bin*	E6: *Worry, concern* (Preocupação, inquietação)

Fonte: Berber Sardinha, 2023b

Dito de outra forma, apesar de cada sequência de palavras ter um significado particular, o sentido, a ideia não são únicos e tendem a ser compartilhados com outras sequências de palavras que possuem uma estrutura semelhante, como pode ser observado na Tabela 2, abaixo (BERBER SARDINHA, 2023b).

TABELA 2: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS QUE COMPARTILHAM OS MESMOS NGCSs

Z5 Q2 Z5 (*Grammar* + *Speech* + *Grammar*) (Gramática + Falas + Gramática)	**A3 Q2 Z5 (Being + Speech + Grammar) (Sendo + Falas + Grammática)**

to talk about (falar sobre)	*be explained from* (ser explicado a partir de)
to predict the (prever o/a/os/as)	*been referred to* (foi/ foram referidos a)
to describe the (descrever o/a/os/as)	*were asked whether* (foram perguntados se)
to convey a (transmitir um/ uma)	*was applied to* (foi aplicado a)
to report a (reportar um/ uma)	*was asked to* (foi-lhe pedido que)

Fonte: Berber Sardinha, 2023b

Uma vez que NGCSs reúnem sequências de palavras com um significado semelhante, isso permite que se possa explicar uma grande proporção de sequências de palavras organizadas conceitualmente em um corpus. Além disso, também é possível entender melhor o significado desses NGCSs quando são vistos como grupos correlacionados, uma vez que se pode observar como esses diferentes NGCSs interagem nos textos de modo a melhor expressar a mensagem desejada pelo autor (BERBER SARDINHA, 2023b).

A relevância do nosso estudo se deve ao fato de que realizaremos uma investigação descritiva da linguagem de aprendiz baseada em corpora, analisando a variação do uso de NGCSs em cada um dos textos que compõem o corpus de estudo, um corpus de aprendiz, e também ao fato de que, pela primeira vez, serão analisados os NGCSs usados em corpora da produção oral e escrita de aprendizes de modo que se avalie a variação no que diz respeito ao uso desses NGCSs entre os textos, assim como em cada texto.

A fim de atingir os objetivos desta pesquisa, formulamos as seguintes questões:

1. Quais são as dimensões de variação de uso de NGCSs na fala e na escrita de alunos de inglês como língua estrangeira?
2. Quanto da variação no uso desses NGCSs pode ser explicada:
   a) pelo modo, escrito ou falado?
   b) pela língua materna dos aprendizes?
   c) pelo nível de proficiência dos alunos?
   d) pela idade dos alunos
   e) pela tarefa designada ao aluno?
   f) pelos anos estudando inglês?

A presente pesquisa está dividida em cinco partes. Após a introdução, tratamos no capítulo 1 da fundamentação teórica e no capítulo 2 da metodologia empregada. No capítulo 3, apresentamos e discutimos os resultados. Por último, no capítulo 4, expomos as considerações finais, e, na sequência, as referências bibliográficas.

1. FUNDAMENTAÇÃO TEÓRICA

Nesta seção é apresentada a fundamentação teórica desta pesquisa. Inicialmente trataremos da LC e de sua vertente estadunidense, a Análise Multidimensional. A seguir abordaremos a LC de Aprendiz. Na sequência, apresentaremos os pacotes lexicais (*lexical bundles*) e os NGCSs. Por último, versaremos sobre categorias semânticas e o etiquetador USAS.

1.1 LINGUÍSTICA DE *CORPUS*

O trabalho aqui proposto tem como fundamentação teórica principal a LC, que se ocupa da "coleta e da exploração de *corpora*, ou conjuntos de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística" (BERBER SARDINHA, 2004, p. 3). Dito de outra forma, o objetivo da LC é a pesquisa de uma língua ou variedade linguística e, para tal, ela utiliza indícios empíricos extraídos por meio de processamento computacional. Seu desenvolvimento está estreitamente associado ao uso do computador, que tornou possível a análise de grandes quantidades de texto (SINCLAIR, 1991; BERBER SARDINHA, 2004; McNERY; HARDIE, 2011).

Biber, Conrad e Reppen (1998) referem-se à LC como uma abordagem baseada em *corpus* (*corpus-based approach*). Esses autores enumeram o que consideram ser as características essenciais da LC para diferenciá-la de outras abordagens analíticas na linguística: é empírica e analisa padrões da língua em uso em textos; utiliza um *corpus*, definido como sendo uma grande e criteriosa coletânea de textos naturais, como base para a pesquisa; emprega extensivamente computadores nas suas análises e faz uso tanto de técnicas qualitativas quanto de técnicas quantitativas (1998, p. 4).

McEnery e Hardie (2012) ao discutirem sobre o que seria afinal a LC, ressaltam que ela não se concentra diretamente na investigação de um aspecto linguístico específico. Para os autores, a LC é uma área que se dedica a uma "variedade de procedimentos, ou métodos, para o estudo da língua"[7] (2012, p. 1), procedimentos esses que podem ser usados em diferentes áreas da linguística. Além disso, reforçam a ideia de que a LC não é um grupo monolítico de metodologias. Contudo, é possível fazer algumas generalizações para caracterizá-la. A primeira é que a LC trata de coletâneas de textos que podem ser processados por computadores (*machine-readable texts*), considerados apropriados para abordar um conjunto específico de questões de pesquisa (idem, ibidem). Em segundo lugar, os *corpora* são explorados qualitativa e quantitativamente por meio de ferramentas que permitem aos

[7]"[…] *a set of procedures, or methods, for studying language*."

pesquisadores analisá-los de forma rápida e precisa, como os concordanciadores, que facilitam a análise das palavras em seus contextos, e as listas de frequência de palavras, que catalogam todas as palavras do *corpus* e indicam quantas vezes cada uma ocorreu. A terceira generalização, que se origina das anteriores, é que o corpus deve ser criteriosamente selecionado para se adequar às questões de pesquisa (2012, p. 2).

McEnery e Hardie (2012) destacam que "o desenvolvimento da LC gerou, ou pelo menos facilitou, a exploração de novas teorias da linguagem – teorias que se inspiram no uso atestado da língua e nos resultados obtidos a partir deles"[8] (2012, p.1).

A partir dessas diferentes caracterizações, podemos dizer que a LC é uma área de pesquisa da linguagem que analisa qualitativa e quantitativamente os *corpora*, que são textos legíveis por computador e que foram coletados com o intuito de responderem a perguntas de pesquisa a respeito de uma variedade linguística, ou de uma língua, por intermédio da extração de dados empíricos mediante processamento computacional (BIBER, CONRAD e REPPEN, 1998; BERBER SARDINHA, 2004; McNERY; HARDIE, 2012).

Além de empírica, outra distinção da LC é a sua visão probabilística da linguagem, que pressupõe que "embora muitos traços linguísticos sejam possíveis teoricamente, eles não ocorrem com a mesma frequência" (BERBER SARDINHA, 2000b, p.350; HALLIDAY, 1991). O mais importante nessa variação de frequências entre as características linguísticas é o fato delas não serem aleatórias: "[...] há um mapeamento regular entre a frequência maior ou menor de um traço e um contexto de ocorrência" (BERBER SARDINHA, 2000b, p. 351).

Essa não aleatoriedade implica na padronização da linguagem, que se revela pela recorrência de traços linguísticos (STUBBS, 1995; PARTINGTON, 1998; SINCLAIR, 1991, 2004; BERBER SARDINHA, 2000, 2004, 2014a, 2023a, 2023). Em outras palavras, "a linguagem forma padrões que apresentam regularidade", ou seja, são "estáveis em momento distintos, isto é, tem frequência comparável em corpora distintos", e apresentam "variação sistemática, isto é, correlacionam-se com variedades textuais, dialetais, etc" (BERBER SARDINHA, 2004, p. 31).

Sinclair (1991) ao expor como concebe a coocorrência de palavras desenvolve dois modelos de interpretação para 'explicar de que forma o significado surge no texto"[9] (1991, p.109). Um deles é o princípio idiomático (*the idiom principle*) que "consiste em o usuário da língua dispor de um vasto número de frases semiconstruídas que constituem escolhas únicas, ainda que possam ser analisadas

[8] "[…] *the development of corpus linguistics has also spawned, or at least facilitated the exploration of, new theories of language – theories which draw their inspiration from attested language use and the findings drawn from it.*"
[9] "[…] *to explain the way in which meaning arises from language text*."

em segmentos” [10]**,** como por exemplo, *of course* (claro/ certamente/ naturalmente) (SINCLAIR, 1991, p. 110).

Sinclair (1991) expõe que o exemplo mencionado anteriormente desempenha um papel semelhante ao de uma palavra. Ele explica que o *of* de *of course* não é a preposição dos livros de gramática, precedida por um substantivo, núcleo de um grupo nominal, como, por exemplo, *a member of*, ou por um quantificador como em *much of.* Nem o substantivo *course* é o substantivo contável encontrado em dicionários; se fosse, teria de ter sido antecedido por um determinante: ‘o seu significado não é uma propriedade da palavra, mas da frase”[11] (1991, p.111).

O outro princípio de organização geral da língua em uso desenvolvido por Sinclair (1991) para explicar como o significado emerge do texto é o princípio da escolha aberta (*the open-choice principle*), que significa que “a cada ponto em que uma unidade é completada – uma palavra, uma frase, uma oração – uma grande amplitude de escolhas se abre, e a única restrição é a gramaticalidade [12] (SINCLAIR, 1991, p. 109). Segundo Sinclair (1991), esse é o princípio no qual as gramáticas se baseiam.

Os dois princípios coexistem nos textos e são necessários para a sua interpretação, apesar de serem essencialmente opostos: “não há gradação de um para o outro; a mudança de um modelo para o outro será brusca”[13] (1991, p. 114). A coexistência dos dois foi demonstrada graficamente por Berber Sardinha (2014a) em produções textuais de falantes competentes do português, e também por Gil (2023) em textos de aprendizes brasileiros de inglês.

Segundo Berber Sardinha (2004), o pioneirismo de Sinclair inspirou um extenso número de trabalhos assim como foi importante para o desenvolvimento da fraseologia baseada em *corpus.*

Granger e Paquot (2008) destacam que fraseologia tem um amplo escopo e que há uma vasta e confusa terminologia a ela associada. Segundo as autoras, pode-se dizer que fraseologia se refere ao “estudo da estrutura, significado e uso de combinações de palavras” (COWIE apud GRANGER; PAQUOT, 2008, p. 27).

Segundo Granger e Paquot (2008), há duas principais abordagens na pesquisa de padrões fraseológicos. Uma delas, ligada à tradição da ex-União Soviética e da Europa Oriental, é denominada como abordagem fraseológica (*phraseological approach*). As autoras destacam que essa é uma abordagem mais tradicional e que merece o crédito de ter estabelecido a fraseologia como uma disciplina. Nela pesquisam-se combinações fixas, “cujo significado não deriva do significado dos seus

[8] *The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments.*
[11] “[…] *its meaning is not a property of the word, but of the phrase*.”
[12] “*At each point where a unit is completed (a word, a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness.*”
[13] “*There is no shading of one into another; the switch from one model to the other will be sharp.*”

constituintes”[14], tais como, provérbios ou expressões idiomáticas. Unidades múltiplas de palavras, mais variáveis e provenientes de combinações livre, não eram consideradas objeto dessa abordagem (2008, p. 28). Ebeling e Hasselgard (2015) salientam que essa análise usa o método descendente (*top-down*).

A outra é denominada como abordagem baseada na frequência (*frequency-based approach*). É uma abordagem indutiva (GRANGER; PAQUOT, 2008, p. 28), que usa o método ascendente (*bottom-up*) e “adota uma definição mais ampla para o seu objeto de estudo”[15], não necessariamente constituindo uma unidade semântica (EBELING; HASSELGARD, 2015, p. 208): esse enfoque identifica coocorrências lexicais (GRANGER; PAQUOT, 2008, p. 28). Granger e Paquot (2008) ressaltam que essa análise se originou a partir do trabalho lexicográfico de John Sinclair e ilustra o princípio idiomático, gerando “uma ampla gama de combinações de palavras que não se encaixavam em categorias linguísticas pré-definidas”[16] (GRANGER; PAQUOT, 2008, p. 29), como, por exemplo, estruturas (*frames*), estruturas colocacionais (*collocational frames*) e coligações (*colligations*).

Gray e Biber (2015), ao tratarem da fraseologia no inglês ressaltam que uma boa porção da língua usada rotineiramente é ‘composta de expressões pré-fabricadas”[17].

Segundo Gray e Biber (2015), os estudos de padrões fraseológicos na língua inglesa utilizando *corpus* datam do início da década de 1990, e tratavam basicamente de duas questões que ainda motivam os estudos no presente: a definição e a identificação de unidades fixas de multipalavras e a análise da função discursiva que essas unidades de palavras têm. Os autores também salientam que sequências de palavras foram investigadas sob diversas definições, tais como, ‘frases lexicais’ (*lexical phrases*) ‘expressões fixas’ (*fixed expressions*), ‘padrões pré-fabricados’ (*prefabricated patterns*), n-gramas (*n-grams*), pacotes lexicais (*lexical bundles*). O que esses estudos tinham em comum, independentemente da definição usada, era o ‘foco em como as palavras se associam em combinações mais ou menos fixas’[18] (GRAY; BIBER, 2015, p. 125).

Independentemente da definição usada na investigação de sequências de palavras, Gray e Biber (2015) apontam para uma clara diferença entre a pesquisa baseada (*corpus-based*) ou guiada por *corpus* (*corpus-driven*), ressaltando que ambas são bastante profícuas. Os autores, contudo, destacam que tal distinção não é absoluta, e que há estudos de estruturas lexicais descontínuas (*discontinuous lexical frames*) que utilizam as duas abordagens. Os trabalhos dados como exemplos (BIBER, 2009; ROMER, 2010) iniciam com uma investigação guiada por *corpus* para a identificação de estruturas lexicais contínuas, e em uma segunda etapa, baseada

[14] “[…] *whose meanings cannot be derived from the meanings of the constituents*”
[15] “[…] *adopts a broader definition for its object of study* […]”
[16] *“*[…] *generates a wide range of word combinations, which do not all fit predefined linguistic categories.*”
[17] […] *composed of prefabricated expressions.”*
[18] “[…] *a focus on how words combine into more or less fixed combinations.”*

em *corpus,* averiguam se as estruturas que foram identificadas podem igualmente ocorrer em estruturas descontínuas (GRAY; BIBER, 2015). Todavia, Gray e Biber (2015) sublinham que uma análise exploratória, guiada por *corpus,* é capaz de identificar padrões de ocorrência que poderiam deixar de ser detectados numa análise baseada em *corpus*, a despeito de sua alta frequência.

O uso de padrões fraseológicos é um dos aspectos distintivos entre os aprendizes de uma língua e os seus falantes nativos (GRANGER; PAQUOT, 2008; EBELING; HASSELGARD, 2015; ELLIS et al., 2015).

A história da LC não inicia, contudo, no século passado (BERBER SARDINHA, 2004; BARNBROOK; MASON; KRISHNAMURTHY, 2013). A existência de *corpora* data de épocas anteriores, como o Corpus Helenístico na Grécia Antiga e os *corpora* de citações bíblicas durante a Idade Média. Ao longo do século XX, houve um notável interesse por parte de vários pesquisadores na descrição da linguagem através de *corpora*, por exemplo, linguistas, como Boas e Fries, e o educador Thorndike (BERBER SARDINHA, 2004). Berber Sardinha (2004) destaca que, diferentemente da atualidade, esses *corpora* eram coletados e analisados manualmente. Outra distinção é que eram destinados ao ensino de línguas.

Um dos eventos de destaque na história da LC foi a criação do *Survey of English Usage* (SEU), iniciado em 1959 e "compilado por Randolf Quirk e sua equipe em Londres" (BERBER SARDINHA, 2004, p. 3). Coletado manualmente, esse *corpus* foi planejado para conter um milhão de palavras. Segundo Berber Sardinha (2004), o SEU foi uma referência para outros *corpora* subsequentes. Para a sua compilação foram estabelecidos critérios precisos. Havia um número fixo de duzentos textos e uma quantidade fixa de cinco mil palavras para cada texto. A organização do corpus foi feita em fichas de papel. Cada ficha continha uma palavra contextualizada em um texto de dezessete linhas. Além disso, as palavras foram classificadas gramaticalmente, atribuindo-se a cada ficha uma categoria específica. Esse conjunto de categorias serviu como "base para o desenvolvimento dos etiquetadores computadorizados contemporâneos", que realizam a identificação automática de traços gramaticais (BERBER SARDINHA, 2004, p. 4).

O surgimento dos computadores *mainframe* na década de 60 marcaram uma mudança significativa nesse cenário: começaram a ser adotados por centros universitários de pesquisa, tornando-se instrumentos importantes para investigações na área da linguagem. Um maior número de pesquisadores passou a ter acesso ao processamento de linguagem natural, ao mesmo tempo em que o avanço da tecnologia computacional permitiu a realização "de tarefas mais complexas de maneira mais eficiente" (BERBER SARDINHA, 2004, p. 4).

Outro marco significativo na história da LC foi o surgimento do *Brown University Standard Corpus of Present-day American English* (1964), que foi o primeiro *corpus* linguístico eletrônico a conter um milhão de palavras. Berber Sardinha (2004) salienta que esse *corpus* apresentava uma quantidade considerável de palavras, especialmente em uma época em que as dificuldades relacionadas à informatização

de textos eram significativas, além de ter sido coletado num período histórico no qual investir na compilação de registros linguísticos era visto com desconfiança.

As transformações continuaram com a popularização dos microcomputadores pessoais durante os anos 1980, não apenas difundindo o uso de *corpora*, mas também proporcionando o acesso a ferramentas para o seu processamento, o que concorreu para o ressurgimento e o revigoramento da "pesquisa linguística baseada em *corpus*", como observado por Berber Sardinha (2004, p. 5).

Kauffmann (2020) salienta que, quaisquer que sejam os caminhos nos próximos anos, é muito pouco provável que a LC vá abandonar "a visão indutiva e empírica", que tem como fundamento "a natureza probabilística da linguagem" (2020, p. 12).

1.2 ANÁLISE MULTIDIMENSIONAL

Desenvolvida por Douglas Biber (1984, 1985, 1986, 1988), a Análise Multidimensional (doravante AMD) é considerada a escola americana da LC. É hoje considerada a principal contribuição "de Biber para o estudo da linguagem baseado em corpus eletrônico" (BERBER SARDINHA, 2004, p. 299).

Egbert e Staples (2019) traçam um breve percurso que antecedeu o desenvolvimento da AMD por Biber. Os autores relatam que, na década de 1970, linguistas começaram a observar padrões salientes de variação e uso da linguagem.

> Alguns pesquisadores observaram que variedades linguísticas diferiam ao longo de múltiplas dimensões funcionais (BERNSTEIN, 1970; IRVINE, 1979; OCHS, 1979). Outros notaram que características linguísticas frequentemente coocorrem em variedades linguísticas e que esses padrões de coocorrência revelam dimensões funcionais subjacentes (ERVIN-TRIPP, 1972; HYMES, 1974; BROWN e FRASER, 1979) (2019, p. 125)[19].

Apesar de relativamente simples teoricamente, tais diagnósticos a respeito da variação linguística eram penosos de quantificar, "requerendo sofisticadas técnicas estatísticas multivariadas"[20] (idem, ibidem). Em 1960, John Carroll realizou "a primeira tentativa de medir a variação textual levando em consideração padrões de

[19] "*Some researchers noted that linguistic varieties differed along multiple functional dimensions (Bernstein 1970; Irvine 1979; Ochs 1979). Others observed that linguistic features often co-occur in linguistic varieties and those co-occurrence patterns reveal underlying functional dimensions (Ervin-Tripp 1972; Hymes 1974; Brown and Fraser 1979).*"

[20] "[…] *requiring sophisticated multivariate statistical techniques.*"

coocorrência entre variáveis linguísticas"[21] usando a análise fatorial, mas, segundo Egbert e Staples (idem, ibidem), seus métodos ainda não eram robustos o suficiente para dar conta da explicação multivariada de dados.

As constatações acima, juntamente com o trabalho de Carroll, serviram de inspiração para Biber desenvolver a AMD (EGBERT; STAPLES, 2019). Como no trabalho de 1960, a análise fatorial é uma das etapas essenciais na AMD. É essa técnica estatística que permite identificar os padrões MD (multidimensionais) de coocorrência linguística.

Em outras palavras, a partir dos resultados da análise fatorial[22], que é um procedimento estatístico multivariado, empregado para "investigar as correlações subjacentes entre um grupo de variáveis observadas"[23] (LOEWEN; GONULAL, 2015, p. 182), identifica-se o grupo de variáveis que coocorrem em cada fator. Ao ser interpretada a função comunicativa dessas variáveis, o fator assume o estatuto de dimensão (BERBER SARDINHA, 2004, p. 304). Ou seja, cada dimensão é formada "por um grupo independente de características linguísticas que coocorrem [no fator], e cada padrão de coocorrência pode ser interpretado em termos funcionais"[24] (BIBER, 1988, p. 14).

Biber (1988) ressalta que a coocorrência dos padrões por si só não é muito interessante. O que se busca é entender "[...] por que que esse grupo particular de características coocorre nos textos [...] [25], quais os parâmetros funcionais ou situacionais que se relacionam com esse grupo de características, causando o seu uso sistemático (1988, p.16).

Um dos pressupostos do trabalho de Biber (1988) é que "a variação linguística em qualquer língua é muito complexa para ser analisada em termos de uma única dimensão"[26] (1998, p. 22), tais como, formal/ informal, envolvido/ distante. Portanto, do seu ponto de vista teórico, "a descrição da variação linguística em uma língua será multidimensional"[27] (idem, ibidem).

Biber (1988) também destaca que, no seu trabalho, a variação linguística não é entendida dicotomicamente, mas sim "como parâmetros de variação *contínuos* e, quantificáveis, isto é, como escalas *contínuas*" [28] (idem, ibidem). Então, o mesmo texto pode ser mais ou menos formal, mais ou menos elaborado porque cada texto terá uma caracterização quantitativa precisa em relação a cada dimensão.

---

[21] "*The earliest attempt to measure textual variation by accounting for co-occurrence patterns among linguistic variables* […]"

[22] Esse procedimento estatístico multivariado será explicado na Metodologia.

[23] "[...] *to investigate the underlying correlations among a set of observed variables."*

[24] "*Each dimension comprises an independent group of co-occurring linguistic features, and each co-occurrence pattern can be interpreted in functional terms."*

[25] "[…] *why these particular set of features co-occur in texts.*"

[26] "[…] *linguistic variation in any language is too complex to be analyzed in terms of any single dimension."*

[27] "[…] *the description of linguistic variation in a given language will be multi-dimensional* […]"

[28] "[…] *as continuous quantifiable parameters of variation, i.e., as continuous scales.*"

Segundo Conrad e Biber (2001), a AMD é uma abordagem metodológica que foi desenvolvida com dois objetivos. O primeiro foi "identificar os padrões salientes de coocorrência linguística em uma língua em termos empíricos e quantitativos"; o segundo foi "comparar registros escritos e falados no espaço linguístico definido pelos padrões de coocorrência"[29] (2001, p. 5).

Biber (1988) diferencia duas abordagens usadas na investigação da variação textual: a macroscópica, que se propõe a definir as dimensões de variação em uma língua, a identificar "os parâmetros gerais de variação textual em um dado 'domínio'"[30], por exemplo, falado ou escrito; e a microscópica, que "proporciona uma descrição detalhada das funções comunicativas de características linguísticas particulares"[31], por exemplo, o uso de pronomes da primeira pessoa do singular como "marcadores de envolvimento pessoal"[32] (1988, p. 61).

As duas abordagens se complementam. Biber (1988) afirma que a microanálise possibilita "a identificação e interpretação funcional de características linguísticas potencialmente significativas"[33], enquanto a macro análise proporciona a "estrutura teórica geral", que é o conhecimento das dimensões (1988, p. 62). A análise feita em seu trabalho de 1988 baseia-se nessas duas abordagens. Ao analisar a coocorrência de 67 características linguísticas em 481 textos, identificando 7 dimensões, é a abordagem macroscópica que está sendo usada. Na interpretação dessas características em termos funcionais, é a abordagem microscópica (idem, p. 63).

Biber (1988) usou dois *corpora,* um escrito: o *Lancaster-Oslo-Bergen Corpus of British English* (LOB *Corpus*), com 500 textos, publicados in 1961, com aproximadamente 2000 palavras cada, e um total de cerca de um milhão de palavras e 15 registros[34], por exemplo, editoriais, ficção, humor, religião, biografias, ensaio, dentre outros (BIBER, 1988, p. 66). O outro *corpus* era falado: o *London- Lund Corpus of Spoken English*, com 87 textos de inglês britânico, com aproximadamente 5000 palavras cada um. O *London_Lund* tinha por volta de 500.000 palavras, com 6 tipos de registros orais, como por exemplo, conversas privadas, conversas pelo telefone, programas de rádio etc. (idem, p. 66).

---

[29] "[...] *to (1) identify the salient linguistic co-occurrence patterns in a language, in empirical/quantitative terms; and (2) compare spoken and written registers in the linguistic space defined by those co-occurrence patterns.*"

[30] "[...] *identify the overall parameters of linguistic variation within a given 'domain'* […]"

[31] "[...] *provides a detailed description of the communicative functions of particular linguistic features* […]"

[32] "[…] *markers of personal involvement."*

[33] "[...] *the identification and functional interpretation of potentially important linguistic features* […]"

[34] Berber Sardinha (2004) destaca que Biber inicialmente usou 'gênero, mas, a partir de 1995, tem utilizado registro. Por registro, ou gênero, entenda-se "uma variedade definida por variáveis situacionais, isto é, não linguísticas, cujos rótulos são empregados por falantes nativos da língua no dia a dia. [...] é um termo impreciso [...] prosa acadêmica, conversação espontânea e editoriais jornalísticos seriam três registros [...]" (2004, p 303).

Antes da comparação dos textos, decidiu-se quais seriam as características linguísticas que seriam usadas, tendo como objetivo a inclusão do máximo possível de traços linguísticos avaliados como potencialmente importantes (idem, p.72).

Inicialmente havia 7 fatores, mas Biber (1988) considerou que o sétimo não era robusto o suficiente para uma interpretação (1988, p.114). Ficaram então seis (BIBER, 1998), mas também o sexto apresentava poucas variáveis com carregamento significativo, não sendo incluído em estudos subsequentes (CONRAD; BIBER, 2013, p.39). Abaixo temos o resultado da interpretação dos fatores – as dimensões, na tradução de Veirano Pinto (2013).

- Dimensão 1: Produção interacional *versus* produção informacional;
- Dimensão 2: Propósitos narrativos *versus* não narrativos;
- Dimensão 3: Referência explícita *versus* dependente de situação;
- Dimensão 4: Persuasão explícita;
- Dimensão 5: Informação abstrata *versus* não abstrata.

A dimensão 1 tem no polo negativo traços marcadamente informacionais, tais como substantivos e adjetivos. Os registros que mais carregaram nesse polo são documentos oficiais, prosa acadêmica e reportagens. O polo positivo, por sua vez, tem contrações, pronomes da 2ª e da 1ª pessoa, que caracterizam uma interação mais espontânea. Os registros que mais carregam nesse polo são conversas por telefone e presenciais. A dimensão 2 está associada à narrativa, com verbos no passado pronomes da 3ª pessoa no polo positivo; o registro que mais carregou foi ficção romântica. No polo negativo, temos verbos no presente e adjetivos; os registros que mais carregaram foram transmissão de rádio e passatempo (*hobbies*). Na dimensão 3 carregaram no polo positivo orações relativas como objeto e como sujeito, associadas a nominalizações, opondo-se a formas adverbiais no polo negativo. O registro que mais carregou polo negativo foi transmissão de rádio; no positivo foram documentos oficiais e cartas profissionais, marcando nessa dimensão um contraste entre referências endofóricas, exemplificadas pelos últimos registros, e exofóricas, ilustrada pelo primeiro. A dimensão 4 só apresentou polo positivo, com infinitivos, verbos modais e persuasivos (*suasive*). Os textos que mais carregaram foram cartas profissionais e editoriais. Por último, a dimensão 5 apresenta no polo positivo conjunções, formas passivas, entre outras, com apenas uma variável no polo negativo, de razão forma/ ocorrência, indicando densidade lexical.

Conrad e Biber (2001) afirmam que com a AMD a noção de coocorrência linguística adquire "um estatuto formal, em que diferentes padrões de coocorrência são analisados como dimensões subjacentes de variação"[35] (2001, p. 6). Após serem determinados quantitativamente, os padrões de coocorrência linguística que compõem cada dimensão são interpretados qualitativamente em suas bases

[35] "[...] *formal status in the MD approach, in that different co-occurrence patterns are analyzed as underlying dimensions of variation*."

funcionais. Isso se dá porque características linguísticas que coocorrem em textos "refletem funções partilhadas" [36], como por exemplo, os primeiros e segundos pronomes pessoais e imperativos denotando interação (idem, p.6).

Segundo Conrad e Biber (2001), as características essenciais da abordagem MD podem ser resumidas como:

1. o foco da pesquisa está em textos, registros e tipos de texto, e não em "construções linguísticas individuais"[37];
2. baseia-se no pressuposto de que tipos diferentes de texto também são linguística e funcionalmente distintos, portanto, a análise de uma ou duas variedades textuais não é suficiente para tecer conclusões a respeito de um campo discursivo. Por exemplo, a análise de cartas pessoais e prosa acadêmica, não seria suficiente para uma concepção adequada da escrita. Outras variedades deveriam ser incluídas, tais como, editoriais, ficção, etc.
3. sendo multidimensional, o que significa dizer que pressupõe "múltiplos parâmetros de variação" operando "em qualquer campo discursivo"[38] (2001, p. 7)
4. sendo uma abordagem quantitativa e empírica;
5. sintetiza metodologias qualitativas e quantitativas, uma vez que "análises estatísticas são interpretadas em termos funcionais para determinar as funções comunicativas subjacentes associadas a cada padrão de distribuição"[39] (idem, ibidem).

Outras línguas foram analisadas empregando os mesmos procedimentos da AMD utilizados para a descrição da língua inglesa, tais como o como o português brasileiro (BERBER SARDINHA; KAUFFMANN; ACUNZO, 2014c), o espanhol (BIBER; DAVIES; JONES; TRACY-VENTURA, 2006; PARODI, 2007), o coreano (KIM; BIBER, 1994), o gaélico (LAMB, 2008), o somali (BIBER; HARED, 1994) e o tuvaluano de Nukulaelae (BESNIER, 1988). Também foi utilizada para a investigação da variação de registro (BERBER SARDINHA; VEIRANO PINTO, 2014, 2019) e em registros específicos, como a literatura (EGBERT, 2012; KAUFFMANN, 2020).

Kauffmann (2020) afirma que a AMD, tal qual desenvolvida por Biber (1988), consolidou-se como um parâmetro fundamental na análise da variação linguística. Trata-se de uma análise funcional, uma vez que se concentra nas funções desempenhadas por categorias gramaticais coocorrentes, que são as variáveis da análise (2020, p. 15).

Uma vertente inspirada na AMD funcional de Biber (1988 et seq.) é a AMD lexical (CROSSLEY; LOUWERSE, 2007; KAUFFMANN, BERBER SARDINHA, 2021; BERBER SARDINHA, 2014, 2017, 2019, 2021, 2023a). A AMD lexical tem por objetivo

[36] "[...] *reflect shared functions*."
[37] "[...] *individual linguistic constructions."*
[38] "[...] *multiple parameters of variation will operate in any discourse domain*."
[39] "[...] *the statistical analyses are interpreted in functional terms, to determine the underlying communicative functions associated with each distributional pattern*."

“capturar os parâmetros lexicais que orientam a variação”[40] linguística (BERBER SARDINHA, 2023a, p. 71). Na AMD funcional, as variáveis são unidades estruturais, tais como classes de palavras e tipos de oração. O objetivo é identificar a função comunicativa das variáveis coocorrentes nos textos. Na AMD lexical, por sua vez, as variáveis são unidades lexicais, como palavras, lemas, n-gramas, colocações presentes nos textos (BERBER SARDINHA, 2019). Outra diferença entre a AMD funcional e a lexical é que, na primeira, a listagem de variáveis é delimitada pelo número de classes gramaticais identificadas pelos programas etiquetadores. Na segunda, em contrapartida, a lista de variáveis é aberta, dependendo do *corpus* em análise e das unidades lexicais que estão sendo investigadas (DELFINO; ARAÚJO; BERBER SARDINHA, 2018; KAUFFMANN, 2020).

A AMD Lexical já foi realizada com múltiplos objetivos, como, por exemplo, a classificação de registros usando bigramas (CROSSLEY; LOUWERSE, 2007), a identificação da representação de dois países, Brasil e EUA (BERBER SARDINHA, 2014), a análise da variação de registro (BERBER SARDINHA, 2017) e a identificação de formas linguísticas de representação relacionadas à nacionalidade (BERBER SARDINHA, 2019). Podemos ainda mencionar os trabalhos de Mayer (2018), que conduziu uma análise sobre a variação lexical presente em comentários de postagens em diferentes redes sociais de língua inglesa, evidenciando seu perfil temático ou discursivo nesse registro linguístico; Romeiro (2020), que explorou a obra da fotógrafa Sally Mann, utilizando textos produzidos pela crítica especializada sobre a artista e sua obra; Veiga (2021), que compilou e analisou textos sagrados de diversas religiões, traduzidos ou escritos em inglês, identificando 6 dimensões de variação lexical e 4 clusters no contexto religioso; Souza (2020), que empreendeu uma análise sobre viés étnico-racial em letras de músicas brasileiras, contribuindo para o entendimento das representações sociais presentes nesse tipo de produção cultural.

Abaixo, no Quadro 1, um resumo com as principais semelhanças e diferenças entre a AMD funcional e a lexical.

QUADRO 1 – ANÁLISE MULTIDIMENSIONAL FUNCIONAL E LEXICAL

Tipo de AMD	Funcional	Lexical
Objetivo	Identificar parâmetros subjacentes de variação funcional	Identificar parâmetros subjacentes de variação lexical
Método	Procedimentos estatísticos multivariados	
Características Linguísticas	Principalmente gramaticais	Lexicais
Interpretação	Baseada na função	Baseada no léxico

Fonte: Adaptado de Berber Sardinha, 2023b.

A AMD funcional, historicamente a primeira vertente, foi introduzida e

[40] “[...] *capture the lexical parameters driving the variation.*”

divulgada no Brasil com trabalhos a respeito da produção escrita de aprendizes no final da década de 90 e início do presente século (Berber Sardinha, 2004). O primeiro deles, Pacheco (1997), contrastou a produção textual de alunos brasileiros com a de americanos. Eram 270 textos, 90 de falantes nativos do português brasileiro; 90 de falantes nativos de inglês, e 90 textos em inglês, produzidos pelos estudantes brasileiros. O tema do texto foi controlado. Berber Sardinha (2004) destaca que os resultados desse trabalho indicam que possa haver influência cultural na produção escrita dos alunos.

Outro trabalho foi o de Shimazumi (1998), contrastando a escrita de 20 falantes nativos, 10 alunos britânicos e 10 jornalistas britânicos, e 10 não nativos. A produção textual foi controlada por tópicos, e a análise de Shimazumi baseou-se em traços sistêmicos funcionais (Berber Sardinha, 2004). Segundo Berber Sardinha (2004), os resultados atestam uma propensão para o uso de expressões encontradas em materiais didáticos de ensino de inglês, como *personally, I think*, ao passo que os falantes nativos expressam a interpessoalidade com um repertório mais amplo, que habitualmente não faz parte desses materiais.

Um último exemplo é o de Conde (2002) que analisou a produção escrita de alunos brasileiros. Seu estudo utilizou dois corpora, um procedente de uma escola bilíngue e outro de alunos universitários, o Br-ICLE. Sua pesquisa revelou que o tipo de escola, seja ela ou não bilíngue, influencia a qualidade da escrita dos alunos, mas também que não há diferenças salientes no que diz respeito a importantes características da escrita entre os dois *corpora*, evidenciando que estudar numa escola bilíngue não faz muita diferença na habilidade de escrever um bom texto (BERBER SARDINHA, 2004, p. 324).

Na sequência, versaremos sobre a Linguística de *Corpus* de Aprendiz.

1.3 LINGUÍSTICA DE CORPUS DE APRENDIZ[41]

No estudo de Aquisição de Segunda Língua (*Second Language Acquisition* - SLA), a produção escrita e falada de aprendizes sempre foi um recurso fundamental. Contudo, Granger, Gilquin e Meunier (2015) destacam que os dados usados eram provenientes de tarefas linguísticas extremamente controladas, não retratando, portanto, o que “os aprendizes fazem em contextos de comunicação mais naturais” (2015, p. 1). Outro problema apontado pelas autoras é que as evidências empíricas costumavam ser oriundas de um pequeno grupo de indivíduos, o que gerava apreensão em relação à representatividade. Tais questões, juntamente com o desejo

[41] Neste trabalho optou-se por usar a expressão Linguística de *Corpus* de Aprendiz já usada há mais de duas décadas em dissertações e teses na PUC - SP (CONDE, 2002; DELEGÁ-LUCIO, 2006. 2013; GIL, 2017; DA SILVA, 2022) como equivalente a *Learner Corpus Research.*

de elaborar ferramentas pedagógicas mais direcionadas para os aprendizes, ensejou o surgimento dos *corpora* de aprendizes, que as autoras definem como "uma coleção eletrônica de dados naturais, ou quase naturais, produzidos por aprendizes de uma segunda língua ou de uma língua estrangeira e reunidos de acordo com critérios explícitos na sua elaboração"[42] (idem, ibidem). Segundo as autoras, os *corpora* de aprendizes deram origem a uma profusão de trabalhos que foram agrupados sob denominação de "*learner corpus research*" (LCR).

O desenvolvimento desse campo de pesquisa ocorreu no final da década de 1980 e início da década de 1990, quando acadêmicos e editoras começaram a compilar *corpora* de inglês não nativo, os *corpora* de aprendizes (*learner corpora*), reconhecendo seu potencial tanto prático quanto teórico (GRANGER, 1998, 2002). Tal acontecimento foi historicamente relevante porque redefiniu o conceito de *corpus*, "que previa (na prática, não na teoria) que a linguagem permitida no *corpus* tinha de pertencer à variedade nativa" (BERBER SARDINHA, 2004, p. 265).

A LC de Aprendiz foi inicialmente nomeada CLC (sigla para *computer learner corpora* - GRANGER, 1998); atualmente é denominada *Learner Corpus Research* (GRANGER, GILQUIN; MEUNIER, 2013; GRANGER, GILQUIN; MEUNIER, 2015). Essa vertente de pesquisa adota os principais fundamentos, ferramentas e métodos da LC, visando proporcionar descrições mais precisas da língua dos aprendizes (GRANGER, 2002, p. 1). De acordo com a mesma autora, essa área de investigação linguística estabeleceu uma importante conexão entre a LC e as pesquisas sobre Aquisição de Segunda Língua, que visam compreender os mecanismos envolvidos na aquisição de uma segunda língua, e as pesquisas sobre Ensino de Língua Estrangeira, cujo objetivo é aprimorar o processo de ensino e aprendizagem de línguas estrangeiras (2002, p. 2).

Tanto a LC de Aprendiz quanto a Aquisição de Segunda Língua têm o mesmo objeto de estudo, qual seja, a linguagem de aprendizes. Porém, os estudos de Aquisição de Segunda Língua focam na competência, ao passo que os da LC de Aprendiz enfatizam o desempenho (*performance*): "o seu principal objetivo é descrever a língua em uso pelos aprendizes em produção real" [43] (GILQUIN; GRANGER, 2015, p. 419). Outra diferença é que a Aquisição de Segunda Língua historicamente privilegiou a morfologia e a gramática, ao passo que a LC de Aprendiz se distingue por seu "marcante enfoque no léxico, na léxico-gramática, e em uma variedade de fenômenos discursivos"[44] (idem, p. 420). Mais uma distinção destacada pelas autoras é que, diferentemente dos estudos de Aquisição de Segunda Língua não baseados em *corpus* que tradicionalmente testam hipóteses, os estudos de LC de Aprendiz são frequentemente descritivos ou exploratórios (GILQUIN; GRANGER, 2015, p. 420). A título de ilustração podemos mencionar a pesquisa de Gil (2023), na

[42] […] *as electronic collections of natural or near-natural data produced by foreign or second language (L2) learners and assembled according to explicit design criteria*."
[43] "[…] *their main objective is to describe the use of language by learners in actual production*."
[44] "[…] *a strong focus on lexis, lexico-grammar, and a range of discourse phenomena*."

qual foi investigada a incidência do princípio idiomático e o do princípio da escolha na produção escrita de aprendizes brasileiros; Gilquin e Granger (2015), que averiguaram o uso de marcadores discursivos de duas palavras na produção oral de falantes nativos e não nativos do inglês; Biber et al. (2020), cujo trabalho descreveu as estruturas gramaticais e os usos característicos de textos de pesquisadores não nativos de inglês com o objetivo de examinar a complexidade gramatical da produção escrita desses aprendizes; Da Silva (2024), cujo propósito foi identificar as dimensões de variação nos padrões léxico-gramaticais da linguagem empregada por aprendizes de inglês.

Um projeto importante na LC de Aprendizes é o *International Corpus of Learner English* - ICLE (*Corpus* Internacional de Aprendizes de Inglês), na Universidade Católica da Louvain, que tem por objetivo mapear o inglês da produção escrita de falantes não nativos. Inicialmente eram catorze nacionalidades: franceses, alemães, holandeses, espanhóis, suecos, finlandeses, poloneses, tchecos, búlgaros, russos, italianos, hebreus, japoneses e chineses; atualmente totalizam 25, entre elas a brasileira. Na mesma instituição, em parceria com outras universidades em diversos países, entre eles o Brasil, há outra iniciativa, o *Louvain International Database of Spoken English Interlanguage* - LINDSEI (Banco de Dados Internacional de Louvain de Interlíngua do Inglês Falado). O propósito deste projeto foi disponibilizar um equivalente falado ao ICLE, com a produção oral de aprendizes avançados de inglês provenientes de diversos países [45].

Berber Sardinha salienta que, além de permitir a descrição da interlíngua, “os *corpora* de aprendizes também podem auxiliar no desenvolvimento de materiais de ensino e de currículos” (2004, p. 271). Os materiais de ensino desenvolvidos a partir de uma variedade nativa, como os que têm por base a Abordagem Lexical (*Lexical Approach*), ignoram as necessidades reais dos alunos e assumem uma visão idealizada da aprendizagem dessa variedade. Segundo esse autor, um *corpus* de aprendiz pode fornecer para o profissional de ensino evidências de quais áreas os alunos têm maior ou menor dificuldade, assim como “quais as tendências de progressão natural do (ou resistência ao) aprendizado (idem, ibidem).

Granger, Gilquin e Meunier (2015) destacam que a LC de Aprendiz tem experimentado desenvolvimentos notáveis. A princípio limitada aos aprendizes de inglês, a LC de Aprendiz atualmente também está pesquisando outras L2s, tais como o alemão, o árabe, o francês, o coreano e o espanhol. O foco dominante ainda é a escrita, especialmente, o ensaio, mas há uma crescente diversificação dos tipos de dados, em especial um crescente número de projetos em oralidade, tais como o LINDSEI, anteriormente mencionado, ou o COREFL.

[45] https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html

1.3.1 Corpus de aprendiz: definição

Gilquin e Granger (2015) afirmam que embora haja muitos tipos de dados sobre desempenho de aprendizes, nem todos podem ser considerados como um *corpus* de aprendiz. Evidências provenientes de pesquisas de Aquisição de Segunda Língua, nas quais os estudantes têm de realizar tarefas determinadas, tais como a leitura em voz alta de um determinado texto, por exemplo, não o são (2015, p. 419). Como qualquer outro *corpus,* um *corpus* de aprendiz deve ser autêntico. Para Sinclair, isso significa ser "obtido a partir das comunicações genuínas de pessoas realizando suas atividades normais"[46] (SINCLAIR, 1996). Porém, esse critério precisou ser adaptado para *corpora* de aprendiz, uma vez que há ocasiões nas quais os estudantes não têm como usar a língua alvo em atividades rotineiras, já que não residem no país no qual ela é falada – por exemplo, um aluno espanhol ou brasileiro aprendendo inglês.

Para um *corpus* de aprendiz ser considerado autêntico, o objetivo da produção, seja ela escrita ou falada, tem de ser "a comunicação de mensagens e a capacidade dos aprendizes de expressarem-se com suas próprias palavras" [47] (GILQUIN; GRANGER, 2015, p. 419). A autenticidade, nesse sentido, pode abranger desde dados obtidos de contextos autênticos de comunicação até aqueles derivados de atividades em ambientes de sala de aula. Dado que a produção escrita de textos é uma atividade inerente ao contexto educacional, os *corpora* de aprendizes provenientes da produção escrita podem ser considerados autênticos (GRANGER, 2002, p. 5), assim como os resultantes de entrevistas informais "que são obtidas para o *corpus*, mas usam procedimentos exercendo um mínimo controle" [48] (NESSELHAULF, 2004 *apud* GILQUIN; GRANGER, 2015, p. 419), ou de descrições de gravuras, que são atividades mais guiadas, como por exemplo, o *Giessen-Long Beach Chaplin Corpus*[49] (GILQUIN; GRANGER, 2015, p. 419), ou o COREFL.

Além de autêntico, um *corpus* de aprendiz pode ser de diferentes tipos: pode ser geral ou específico, escrito ou falado, sincrônico ou longitudinal, com apenas uma língua materna ou com várias línguas maternas. Ao pesquisarem os *corpora* de aprendiz disponíveis, Gilquin e Granger (2015) observaram que alguns tipos são mais usuais do que outros: a maioria, por exemplo, era de estudantes mais proficientes; havia mais *corpora* escritos do que falados, além de normalmente não terem sido etiquetados (2015, p. 420).

Granger (2002) sugere adotar para *corpus* de aprendiz uma definição baseada na abordagem proposta por Sinclair (2002, p. 4):

---

[46] "[…] *gathered from the genuine communications business of people going about their normal business.*"

[47] "[…] *message conveyance and the possibility for learners to use their own wording*."

[48] "[…] *that are elicited for the corpus but that use procedures exerting very little control* […]"

[49]

> *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 procedências.[50]

A seguir versaremos sobre os critérios a serem empregados na compilação de um *corpus* de aprendiz.

1.3.2 A compilação de um *corpus* de aprendiz

Um dos critérios fundamentais na compilação de *corpora* é a autenticidade do material coletado. No que diz respeito aos *corpora* de aprendizes, a autenticidade desse termo abrange uma gama variada, que vai desde dados coletados em situações autênticas de comunicação até aqueles provenientes de ambientes legítimos de sala de aula (GRANGER, 2002, p. 5).

Ao compilar um *corpus* de aprendizes, Granger (1998) destaca que é crucial considerar tanto as características linguísticas quanto as individuais do aprendiz. Os critérios linguísticos são similares aos utilizados na compilação de *corpora* nativos e incluem considerações tais como se a comunicação é oral ou escrita, o tipo de gênero textual (por exemplo, argumentativo ou narrativo) e o contexto da interação (como conversa espontânea ou entrevista informal). A autora salienta que esse registro é essencial devido à variação na produção dos aprendizes, conforme a natureza da atividade proposta. Além disso, o tópico abordado também é um fator relevante, pois influencia a escolha lexical dos aprendizes. Outro elemento a ser considerado são as condições em que a atividade foi realizada, como a presença de um limite de tempo, se a tarefa fazia parte de uma avaliação formal e se os alunos tinham acesso a materiais de referência durante a execução da tarefa. Abaixo, o Quadro 2 resume esses critérios.

QUADRO 2: CRITÉRIOS PARA A COMPILAÇÃO DE UM *CORPUS* DE APRENDIZ

Língua	**Aprendiz**
Meio	Idade
Variedade Textual	Sexo
Tópico	Língua Materna

50 "*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.*"

Condições da Atividade	Região
	Outra Língua Estrangeira
	Nível de Proficiência
	Contexto de Aprendizagem
	Experiência Prática

Fonte: Adaptado de Granger, 1998.

Formulados em 1998, esses também foram os critérios usados na compilação do ICLE (GRANGER, 2008). A autora aponta a relevância de se ter conhecimento tanto do gênero quanto do tópico, mas não foram apresentados estudos empíricos que evidenciassem a influência de um ou de outro (GRANGER, 1998, 2008, 2015). Gilquin (2020) igualmente destaca que, apesar da riqueza de metadados que costuma acompanhar os *corpora* de aprendiz, essas informações ainda não foram usadas de forma a serem exploradas em todo o seu potencial (2020, p. 286). Segundo a autora, estudos que investiguem o impacto de diversas variáveis ainda são comparativamente raros, "embora [...] possam oferecer importantes ideias a respeito dos fatores que mais provavelmente afetam a linguagem de um aprendiz"[51] (idem, p.287).

1.4 PACOTES LEXICAIS E N-GRAMAS DE CLASSE SEMÂNTICA

Biber, Conrad e Cortes (2003) apontam que historicamente linguistas pressupunham que a "gramática é composicional": pequenas unidades, tais como fonemas, morfemas vão se associando e formando palavras; palavras, por sua vez, vão se combinando e formando frases, e assim por diante (2003, p. 71). Contudo, a partir da década de 1990, essa visão começou a ser questionada, argumentando-se que muito da linguagem usada rotineiramente é constituída por expressões pré-fabricadas, tais como, *in a nutshell, if you see what I mean*[52] (idem, ibidem), a partir da verificação de que falantes nativos "usam tanto unidades estruturalmente compostas quanto combinações de multipalavras funcionando como unidades pré-fabricadas" [53] (idem, ibidem).

No fim década de 1990 do século XX, numerosos estudos empíricos tinham por objetivo "definir e identificar [essas] combinações fixas de palavras"[54], assim como de

[51] "[…] *although such studies can offer important insights into the factors that are likely to affect learner language."*
[52] resumindo; se é que você me entende
[53] "[…] *use both structurally-composed units and multi-word combinations functioning as prefabricated units."*
[54] "[…] *define and identify fixed word combinations* […]"

analisar quais as suas funções discursivas (idem, p. ibidem), por exemplo, Sinclair (1991), Granger (1998) e Partington (1998).

Empregando uma abordagem empírica, indutiva, guiada pelo *corpus* (*corpus-driven*), Biber et al. (1999) desenvolveram o construto *lexical bundle* (pacote lexical). Essa expressão foi primeiramente usada em 1999, e depois em trabalhos posteriores, tais como, Biber, Conrad e Cortes (2003); Biber (2006), Berber Sardinha, Teixeira e São Bento Ferreira (2014); Cortes (2015); e Matte e Goulart (2020). Um pacote lexical são sequências recorrentes de três ou mais palavras que ocorrem em um *corpus* (BIBER et al. 1999; CORTES, 2015, 2023; MATTE; GOULART, 2020) – são sucessões contínuas de palavras, sem nenhum espaço vazio entre elas, que ocorrem em discursos naturais (CORTES, 2015, p. 204). Normalmente, não são unidades estruturais perfeitas (idem, p. 200; BIBER, 1999, p. 991).

Cortes (2015) destaca a diferenciação feita por Biber et al. (1999) entre pacotes lexicais de três palavras ou mais longos, com 4 ou mais: pacotes lexicais de três palavras "podem ser considerados uma forma estendida de associação colocacional"[55], e são bastante habituais. "Por outro lado, pacotes de 4, 5 ou 6 palavras têm natureza mais frasal e [são] correspondentemente menos comuns." [56] (CORTES, 2015, p. 204). Essa distinção é relevante porque quanto maior o pacote lexical, menos frequente ele será. A quantidade de palavras do pacote lexical também influencia quais as classes de palavra que o constituem assim como a sua função comunicativa (idem, ibidem).

Além de não se adequarem a nenhum padrão linguístico predefinido, outra característica importante de pacotes lexicais é a sua frequência. Segundo Cortes (2015), para uma sequência de palavras ser considerada um pacote lexical, sua ocorrência no corpus precisa ser significativa. Biber et al. (1999) definiram um ponto de corte arbitrariamente: 10 vezes por milhão de palavras, sendo que essa ocorrência tem de estar distribuída em pelo menos 5 textos. Tal condição é uma forma de evitar que possíveis idiossincrasias de um único falante ou escritor enviese o resultado. Cortes (2015) ressalta que foram testados outros pontos de corte (20, 40 ou mesmo 100 ocorrências por milhão de palavras) para demonstrar que sequências recorrentes de palavras identificadas como pacotes lexicais são bem mais constantes do que o estabelecido pelos pontos de corte (2015, p. 204). O limite de um milhão de palavras também é uma convenção, mas está relacionada à questão do tamanho do *corpus* ao se comparar "pacotes lexicais identificados em *corpora* pequenos ou em *corpora* de diferentes portes"[57] (idem, p. 205).

As unidades que formam os pacotes lexicais são unidades ortográficas. No caso do inglês, isso significou considerar uma contração (*contraction*) como uma palavra, por exemplo, ***don't,*** em vez das duas unidades (*do* + *not*) que foram

[55] '[…] *can be considered as a kind of extended collocational association* […]"
[56] "*On the other hand, four-word, five-word, and sir-word bundles are more phrasal in nature and correspondingly less common*."
[57] "[…] *lexical bundles identified in small corpora and corpora of different sizes.*"

combinadas.

Alguns exemplos de pacotes lexicais comumente usados em conversação na língua inglesa são 'você quer que eu' ('*do you want me to'*), 'não sei o quê' ('*I don't know what'*). Outros são frequentemente usados na prosa acadêmica 'não havia significante' ('*there was no significant*'), 'no caso de' ('*in the case of*').

Importante ressaltar que pacotes lexicais não são expressões idiomáticas (*idioms*) (BIBER et al., 1999; CORTES, 2015).

> Expressões idiomáticas são relativamente invariáveis, com um significado que não é derivado das partes, mas não são necessariamente expressões comuns. Em contrapartida, os pacotes lexicais são as sequências de palavras que ocorrem mais frequentemente num registro. Normalmente, não são expressões fixas, e não é possível substituir a sequência por uma única palavra; de fato, a maioria dos pacotes lexicais não está estruturalmente completa (como nos exemplos acima). [58] (BIBER, 1999, p. 989).

Além de invariáveis, as expressões idiomáticas não costumam ser muito frequentes (menos de cinco por milhão de palavras) e são eventualmente usadas em textos de ficção (idem, ibidem).

Os pacotes lexicais podem ser classificados a partir de sua estrutura, ainda que não sejam unidades completas em si mesmas. Cortes destaca que é um primeiro passo para ordená-los e apontar tendências no seu uso (2015, p. 208). Da mesma forma, podem ser classificados pela sua função. Biber, Conrad e Cortes (2003), propuseram uma taxonomia funcional. Cortes (2015) enfatiza a importância dessa identificação para que se possa analisar e entender como pacotes lexicais são usados, em quais registros e quais significados que expressam ao serem utilizados.

São três as principais categorias estruturais:

a) "Pacotes lexicais que incorporam fragmentos de frases verbais" [59] (2015, p. 207). Podem começar com um sujeito, um marcador discursivo, ou um verbo, como em *is going to be* (vai ser);

b) "Pacotes lexicais que incorporam fragmentos de orações subordinadas"[60]

---

[58] Idioms are relatively invariable expressions with a meaning not derivable from the parts, but they are not necessarily common expressions at all. In contrast, lexical bundles are the sequences of words that most commonly co-occur in a register. Usually they are not fixed expressions, and it is not possible to substitute a single word for the sequence; in fact, most lexical bundles are not structurally complete at all (as in the above examples).

[59] "*Lexical bundles that incorporate verb phrases fragments*: […]"
[60] "*Lexical bundles that incorporate dependent clauses fragments*: […]"

(idem. Ibidem), por exemplo, *if you want to* (se você quiser/ vocês quiserem), dentre outros;

c) "Pacotes lexicais que incorporam fragmentos de frases nominais (*noun phrase*) e preposicionais (*prepositional phrase*) (idem, ibidem). Iniciam com uma frase nominal ou preposicional, tais como, *the end of the* (o fim de/ do/dos da/das) ou *in the context of* (no contexto de/ do/dos da/das).

Há quatro categorias funcionais fundamentais de pacotes lexicais (BIBER et al, 1999, p. 79). Para o desenvolvimento dessas categorias, os pacotes lexicais foram esmiuçados para que as suas funções discursivas pudessem ser definidas[61].

a) pacotes lexicais referenciais (*referential lexical bundles*), como por exemplo, *the end of the* (o fim de/ do/dos da/das);

b) organizadores textuais (*text organizer*), por exemplo, *on the other hand* (por outro lado);

c) pacotes de posicionamento (*stance bundles*), tais como, *it is possible to* (é possível que);

d) pacotes interacionais (interactional bundles), como *what do you mean* (o que você quer/ vocês querem dizer.

Pacotes lexicais são expressões recorrentes em um *corpus,* independentemente de sua idiomaticidade e da sua situação em termos estruturais (Biber et al., 1999). Podem ser sequências de 3, 4, 5 ou 6 palavras. Quanto maior o pacote lexical, mais raro ele será (BIBER et al., 1999; CORTES, 2015, 2022). A frequência mínima de um pacote lexical em um *corpus* para que possa ser considerado como tal é de "pelo menos dez vezes por milhão de palavras, distribuídos por pelo menos cinco textos diferentes" [62] (GRAY; BIBER, 2015, p. 129). Essa frequência é arbitrária, uma convenção (BIBER, 2006; CORTES, 2015). Variam em sua estrutura e em sua função discursiva, mas mesmo assim após análise, puderam ser classificados em diferentes categorias, três em termos estruturais, e quatro em relação à sua função discursiva (BIBER; CONRAD; CORTES, 2003; CORTES, 2015). Podem ser considerados "um construto linguístico básico"[63] (BIBER, 2006, p. 172), cujo exame "em contextos textuais mostra que são importantes elementos

[61] Essas são as categorias principais; cada uma delas é dividida em subgrupos (BIBER; CONRAD; CORTES, 2003, p. 80).

[62] "[…] *at least ten times per million words in the target register, distributed across at least five different texts."*

[63] "[…] *a basic linguistic construct with important functions for the construction of discourse* […]"

constituintes do discurso” [64] (idem, p. 174).

Foram feitos diferentes estudos caracterizando o uso de pacotes lexicais em diferentes registros: conversação (BIBER et al., 1999), escrita acadêmica (CORTES, 2004), testes de proficiência em inglês (GRAY; BIBER, 2013), produção escrita em inglês de alunos não nativos (DUTRA; BERBER SARDINHA, 2013), diferentes registros universitários, tais como, livros didáticos, prosa acadêmica (BIBER et al. 1999, 2006), e também em outras línguas, como por exemplo, Berber Sardinha, Teixeira e São Bento Ferreira (2014), que investigaram pacotes lexicais no português brasileiro.

Cortes (2023) salienta que “todos os pacotes lexicais são em essência n-gramas (trigramas, quadrigramas, e assim por diante), mas apenas aqueles n-gramas que alcancem os critérios pré-estabelecidos de frequência e dispersão que podem ser considerados como tal”[65] (2023, p. 221). N-gramas são tradicionalmente usados em linguística computacional (CROSLLEY; LOUWERSE, 2007). Os dois autores ao classificar registros falados e escritos em inglês usaram a “frequência de bigramas compartilhados entre os *corpora*”[66] nesta investigação (2007, p. 454).

Neste trabalho as características do *corpus* de estudo a serem analisadas são n-gramas de classe semântica (BERBER SARDINHA, 2023b; RIBEIRO, 2023). A pesquisa com NGCSs foi introduzida por Berber Sardinha (2023b), que se inspirou em pacotes lexicais para a sua elaboração. Diferentemente dos últimos, contudo, pacotes lexicais são formados por unidades ortográficas, *going, don’t, I.* NGCSs, por sua vez, são “sequências de categorias semânticas”, sendo que cada categoria representa uma palavra no texto (BERBER SARDINHA, 2023b). Ou seja, NGCSs agrupam sequências de palavras que compartilham as mesmas categorias semânticas e cada categoria semântica expressa um sentido ou conceito. Em decorrência disso, as sequências de palavras incluídas em um mesmo NGCS possuem um sentido similar (idem, ibidem).

Uma vez que NGCSs agregam múltiplas sequências de palavras que compartilham as mesmas características semânticas, tal fato nos permite investigar uma proporção maior de agrupamentos lexicais no *corpus* (idem, ibidem). Como o presente trabalho se trata de uma pesquisa exploratória, optou-se por investigar a ocorrência de trigramas de classe semântica pelos mesmos motivos acima mencionados a respeito de pacotes lexicais com três palavras. Todavia, ao longo do trabalho, a terminologia empregada para se referir a esses agrupamentos semânticos será NGCS, ou NGCSs, no plural.

---

[64] “[…] *in textual contexts shows that they are important building blocks of discourse* […]”

[65] “*All lexical bundles are in essence n-grams (3-grams, 4-grams, and so on) but only those n-grams that meet the frequency and range thresholds can be considered lexical bundles.*”

[66] “[…] *the frequency of bigrams shared across corpora* […]”

1.5 CATEGORIAS SEMÂNTICAS E O ETIQUETADOR USAS

Como apontado por Delfino (2022, p. 12), estudos multidimensionais são baseados em "análise fatorial e seguem, basicamente, os mesmos passos metodológicos". A distinção essencial entre eles é a etiquetagem utilizada. Em outras palavras, a diferença está nas características do *corpus* que são selecionadas para serem as variáveis analisadas.

Nesta pesquisa foram utilizadas etiquetas semânticas. "Etiquetas semânticas indicam campos semânticos que agrupam sentidos de palavras que estão relacionados pelo fato de estarem conectados a um dado nível de generalidade com um mesmo conceito mental"[67] (ARCHER; WILSON; RAYSON, 2002, p. 1). Por exemplo, o campo semântico *Time*[68] agrupa palavras como *calendar, full-time, schedule, seasonality, historically, ago, current, simultaneous, future, tomorrow*[69], entre outras.

Para a etiquetagem do *corpus* desta pesquisa, utilizou-se o etiquetador semântico criado na Universidade de Lancaster: o Sistema de Análise Semântica UCREL (UCREL *Semantic Analysis* System, sigla USAS)[70], que será explicado mais detalhadamente na metodologia. As etiquetas estão organizadas hierarquicamente em 21 campos semânticos principais (ARCHER; WILSON; RAYSON, 2002, p. 2), como mostra a Figura 1.

Figura 1: Campos semânticos do etiquetador USAS[71]

[67] *The semantic tags show semantic fields which group together word senses that are related by virtue of their being connected at some level of generality with the same mental concept.*
[68] Tempo
[69] calendário, tempo integral, programação, sazonalidade, historicamente, atrás, atual, simultâneo, futura, amanhã.
[70] USAS online English tagger (lancaster.ac.uk)
[71] A - Termos gerais e abstratos; B- O corpo e o indivíduo; C- Artes e artesanato; E - Emoção; F - Comida e agricultura; G- Governo e domínio público; H - Arquitetura, habitação e lar; I - Dinheiro e comércio na indústria; K - Entretenimento, esportes e jogos; L - Vida e coisas vivas; M - Movimento, localização, viagem e transporte; N - Números e medidas; O - Substâncias, materiais, objetos e equipamento; P - Educação; Q – Linguagem e comunicação; S - Ações, estados e processos sociais; T - Tempo; W - O mundo e o meio ambiente; X - Ações, estados e processos psicológicos; Y - Ciência e tecnologia; Z - Nomes e palavras gramaticais.

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

Fonte: ARCHER; WILSON; RAYSON, 2002, p.2

Na Figura 2, exemplificamos as categorias do campo semântico *Time* no USAS com palavras do *corpus* de pesquisa.

FIGURA 2: CATEGORIAS E SUBCATEGORIAS DO CAMPO SEMÂNTICO *TIME* NO USAS

T1 Time
  T1.1 General
    T1.1.1 Past: then
    T1.1.2 Present: now
    T1.1.3 Future: will
  T1.2 Momentary: at that moment
  T1.3 Period: in the morning; every time
T2 Beginning/ending: at the end
T3 Old/new/young; age: baby; boy
T4 Early/late: nenhum exemplo encontrado

Fonte: elaborada pela autora

Outros trabalhos também usaram o etiquetador semântico. Berber Sardinha et al. (2022) ao analisar 2.600 imagens presentes em 17.267 tweets, do período que se estendeu do início da pandemia de COVID-19 em 2020 até novembro de 2021, "lança mão das etiquetas semânticas contidas no etiquetador USAS" que "contém diversas categorias [...] semelhantes ao Google Cloud Vision" (idem, p. 191). Delfino (2022) ao examinar as dimensões de variação multimodal da música popular em língua inglesa tem como uma das variáveis dependentes em seu estudo categorias semânticas. Berber Sardinha (2023b) investiga as dimensões de variação para a escrita e para o discurso acadêmico usando o mesmo etiquetador semântico. Igualmente Ribeiro (2023) utiliza esse etiquetador semântico ao pesquisar a linguagem telecinemática.

2. METODOLOGIA

Neste capítulo é descrita a metodologia utilizada, incluindo a apresentação do corpus de estudo, sua etiquetagem, extração e seleção das variáveis, bem como os procedimentos estatísticos empregados.

2. 1 COREFL – *CORPUS OF ENGLISH AS A FOREIGN LANGUAGE*

O COREFL, cujo acrônimo significa Corpus de Inglês como Língua Estrangeira (*Corpus of English as a Foreign Language*), foi o corpus empregado neste estudo. Esse corpus não foi coletado especificamente para esta pesquisa, tendo sido criado e disponibilizado para a comunidade de modo gratuito por pesquisadores da Universidade de Granada, sob os termos da *Creative Commons* (LOZANO; DÍAZ-NEGRILLO; CALLIES, 2020). O COREFL adere aos princípios da Ciência de Dados Aberta (*Open Data Science*). Para a realização deste trabalho, utilizou-se a Versão 1 do corpus, datada de outubro de 2021.

Composto por narrativas escritas e faladas produzidas por aprendizes espanhóis e alemães da língua inglesa com diferentes níveis de proficiência, sua coleta foi iniciada em 2012 com o objetivo de investigar como as pessoas aprendem esse idioma. Para propósitos comparativos, também foi coletado um subcorpus de controle de falantes nativos da língua inglesa principalmente da variedade britânica e estadunidense, além de outras poucas variedades não especificadas na descrição do corpus[72]. A versão usada tem um total de 530.392 palavras, sendo que 42.093 são de falantes que têm o inglês como língua materna, 124.798 que têm o alemão, e 363.501 que têm o espanhol (vide Tabela 3)[73].

TABELA 3: CARACTERÍSTICAS DO CORPUS COREFL

Língua materna	**Palavras**	**Documentos**
Inglês	42.093	175
Alemão	124.798	449
Espanhol	363.501	1.823

Fonte: a autora, adaptado de Lozano, Díaz-Negrillo e Callies, 2020.

[72] COREFL (learnercorpora.com)
[73] http://corefl.learnercorpora.com/statistics

Na sua fase inicial, conhecida como Corpus Piloto, foram incluídos apenas textos produzidos por estudantes espanhóis de inglês, com idades entre 12 e 18 anos, os quais foram obtidos durante aulas de sessenta minutos. Essa etapa fez parte de um programa de pós-graduação em Educação da Universidade de Granada, na Espanha (LOZANO; DÍAZ-NEGRILLO; CALLIES, p. 23). Durante essa fase, o corpus consistia em dois subcorpora distintos. Um deles foi compilado a partir de aulas de inglês frequentadas por alunos cujas outras disciplinas eram ministradas em espanhol. O segundo subcorpus foi constituído por textos de estudantes participantes do *Content and Language Integrated Learning* (CLIL – Aprendizagem Integrada de Língua e Conteúdo), onde algumas disciplinas, além de inglês, eram ministradas nessa língua. Essa distinção ocorreu devido ao tipo de escola onde o corpus foi coletado (LOZANO; DÍAZ-NEGRILLO; CALLIES, p. 23).

Inicialmente, os alunos foram solicitados a preencher um questionário contendo metadados (vide anexo) e a realizar um teste de classificação de seu nível de proficiência. Nessa fase, utilizou-se o teste *English Unlimited Placement Test* (Cambridge University Press, 2010), que estava disponível gratuitamente na internet na época. Esse teste abrange avaliações desde o nível A1 até o C1, conforme o Quadro Europeu Comum de Referência para Línguas (*Common European Framework of Reference for Languages* - CEFR[74]). Em outra ocasião, os alunos foram instruídos a escrever uma narrativa com base nas ilustrações do livro "Frog, Where Are You?" (Sapo, onde você está? - Mayer, 1969) (LOZANO; DÍAZ-NEGRILLO; CALLIES, p. 24). Com o intuito de facilitar a elaboração da narrativa, e considerando que o objetivo da pesquisa era investigar o uso de morfemas gramaticais, todos os alunos receberam um glossário, tiveram instruções fornecidas em espanhol e realizaram uma tarefa inicial padronizada (vide anexo 2) (LOZANO; DÍAZ-NEGRILLO; CALLIES, p. 24). A seguir, apresentamos na Tabela 4 detalhes da estrutura do Corpus Piloto.

[74] https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

TABELA 4: CARACTERÍSTICAS DO CORPUS PILOTO DO COREFL (L1 ESPANHOL)

Tipo de escola	Ano		A1		A2		B1		B2	total	total
		textos	palavras	textos	palavras	textos	palavras	textos	palavras	textos	palavras
Não-CLIL	1ESO*	26	1.687	2	229						
CLIL	1ESO	31	2.961	4	488						
Não-CLIL	2ESO	29	2.571	13	1.882	7	1.216	1	120		
CLIL	2ESO	10	976	23	3.136	12	1.659				
Não-CLIL	3ESO	23	2.232	17	2.401	14	2.373	1	143		
CLIL	3ESO	9	754	25	3.493	9	1.291	3	566		
Não-CLIL	4ESO	11	570	9	1.440	12	2.499	5	1238		
CLIL	4ESO	7	865	28	4.324	8	1.350	1	199		
Não-CLIL	1BAC**	33	2.522	28	3.440	4	652	5	952		
CLIL	1BAC			11	1.593	19	3193	11	2.245		
Não-CLIL	2BAC	18	1.760	15	2.265	3	402				
CLIL	2BAC										
Não-CLIL		140	11.342	84	11.657	40	7.142	12	2.453	276	32.594
CLIL		57	5.556	91	13.034	48	7.493	15	3.010	211	29.093
Total		197	16.898	175	24.691	88	14.635	27	5.463	**487**	**61.687**

*ESO = 'Educação Secundária Obrigatória', educação básica obrigatória (12-16 anos)[75]. **BAC = '*Bachillerato*', educação básica não obrigatória (16-18 anos).

Fonte: a autora, adaptado de Lozano, Díaz-Negrillo e Callies, 2020.

Neste momento, conforme visto na Tabela 4, tinham sido coletados 487 textos produzidos por alunos de 12 a 18 anos, CLIL e não-CLIL, com nível de proficiência de A1 a B2, totalizando 61.687 palavras (LOZANO; DÍAZ-NEGRILLO; CALLIES, 2020, p. 25).

Após essa fase inicial com um Corpus Piloto, estendeu-se a coleta para alunos universitários do primeiro, segundo e terceiro anos da Universidade de Granada que tinham inglês como disciplina obrigatória. Os alunos tinham de 18 a 21 anos e o nível de proficiência variava entre A2 e C1 (LOZANO; DÍAZ-NEGRILLO; CALLIES, 2020, p. 25). A coleta foi similar à que havia sido feita na fase do Corpus Piloto. As atividades foram feitas em sala, fora de uma sala de aula ou em ambas as possibilidades. Os aprendizes também puderam ter acesso a diferentes materiais de referência: corretor ortográfico, dicionários monolíngues ou bilíngues, livros de gramática e leituras de apoio, se assim desejassem. Nesta segunda fase, foi incluída também a produção oral desses alunos, baseada na mesma tarefa escrita que tinha sido dada aos mesmos alunos – em outras palavras, o aluno e a tarefa foram mantidos. Entre a produção escrita e a oral havia um intervalo de dias. Ainda foram coletadas mais 100 produções orais e escritas de aprendizes com o alemão como língua materna da Universidade de Bremen, Alemanha. A maioria tinha nível de proficiência C1 e C2 (LOZANO; DÍAZ-NEGRILLO; CALLIES, 2020, p. 26).

Além da inclusão da produção falada dos aprendizes, foram acrescentadas

[75] A educação básica espanhola é diferente da brasileira. É obrigatória dos 6 aos 16 anos e divide-se em educação primária (6 aos 12) e secundária (12 aos 16). A partir de então o estudante pode optar por parar de estudar, cursar o Bachirellato, que corresponderia ao nosso Ensino Médio e prepara o aluno para ingressar na universidade ou um fazer um curso técnico. https://www.espanhalegal.info/p/estudos/.

mais três tarefas, duas das quais já haviam sido usadas no CEDEL276 - versão 1: fale sobre uma pessoa famosa e resuma um filme que você viu recentemente. A terceira tarefa era o resumo de um trecho de 4 minutos de um filme mudo de Charles Chaplin. Esta segunda fase foi chamada de versão beta. Na Tabela 5, podemos observar como essas tarefas estão distribuídas no COREFL.

TABELA 5: TAREFAS INCLUÍDAS NO CORPUS COREFL

Tarefa	Número de palavras	Número de textos
2. Pessoa famosa	8.322	40
3. Filme	81.427	374
13. Sapo	157.903	928
14. Chaplin	282.704	1.105

Fonte: a autora, adaptado de Lozano, Díaz-Negrillo e Callies, 2020

Na segunda versão do COREFL planejou-se que haveria três corpora de controle nativos: um de inglês, um de espanhol e outro de alemão. Nesta pesquisa, só foi usado o subcorpus de inglês como L1.

O COREFL é mantido em formato de arquivo de texto. Conforme indicado no site do corpus, os textos escritos foram coletados online, com as instruções fornecidas na língua materna dos aprendizes. Alguns dos textos de alunos com níveis mais baixos de proficiência foram inicialmente escritos à mão e depois digitalizados. Os textos orais foram coletados na Universidade de Granada de quatro maneiras distintas, como especificado no corpus: 1) gravados em um laboratório ('*WRITING/AUDIO DETAILS: spoken_offline_lab*'); 2) gravados em um laptop durante uma aula presencial ('*WRITING/AUDIO DETAILS: spoken_offline_classroom*'); 3) durante a pandemia de Covid-19, gravados durante as aulas no Google Meet ('*WRITING/AUDIO DETAILS: spoken_offline_googlemeet*'); e 4) gravados pelos próprios aprendizes em seus laptops e posteriormente transferidos para o banco de dados do corpus ('*WRITING/AUDIO DETAILS: spoken_online*'). Na mesma página do site, que trata do design do corpus, destaca-se que foram seguidos protocolos para a coleta dos áudios, embora esses protocolos não tenham sido explicitados. Os áudios foram transcritos e arquivados em formato de texto, assim como todo o restante do COREFL. As convenções para a transcrição podem ser acessadas no site, na aba 'Guia do Usuário' > 'Convenções de Transcrição'[77].

No site COREFL há o esclarecimento de que esse corpus de aprendizes se inspirou no WriCLE (*Written Corpus of Learner English* - Corpus Escrito de Aprendizes de Inglês) e que está alinhado com outros projetos internacionais nos quais corpora maiores estão sendo coletados, como por exemplo o ICLE (I*nternational Corpus of*

[76] Para mais detalhes a respeito do CEDEL2, consulte CEDEL2: Corpus Escrito del Español L2 (version 2) (learnercorpora.com)
[77] http://corefl.learnercorpora.com/user_guide/conventions

Learner English - Corpus Internacional de Aprendizes de Inglês), da Universidade Católica de Lovaina na Bélgica[78].

Para a avaliação do nível de proficiência dos aprendizes, encontramos no mesmo sítio que foram usadas três medidas no COREFL, todas baseadas no CEFR: uma medida objetiva baseada em dois testes classificatórios da língua inglesa, o *Oxford Quick Placement Test* (2003), que avalia todos os níveis do CEFR (A1, A2, B1, B2, C1, C2), e o *Cambridge Unlimited Placement Test* (2010), que só avalia até C1, e que segundo consta na página, foi o mais usado; uma medida subjetiva, na qual os próprios alunos se atribuem um nível de proficiência, e por fim outra medida objetiva, os certificados de proficiência dos alunos. Para o agrupamento final por nível de proficiência, foi feito um cálculo da porcentagem de acertos do teste classificatório realizado pelo aluno, como explicitado no site.

2.2 ANOTAÇÃO SEMÂNTICA, EXTRAÇÃO E SELEÇÃO DOS NGCS

A primeira etapa da análise consistiu em dividir o COREFL em subgrupos, organizados de acordo com a língua materna do aluno: espanhol (es), alemão (de) ou inglês (en); o modo do texto: escrito (wr) ou falado (sp); e o nível de proficiência do aluno: A1, A2, B1, B2, C1, C2, falante nativo. Com base nessa divisão, o corpus foi seccionado em subcorpora, conforme apresentado na Figura 3 abaixo.

FIGURA 3: LISTA DOS SUBCORPORA

1	de_sp_b1
2	de_sp_b2
3	de_sp_c1
4	de_sp_c2
5	de_wr_b1
6	de_wr_b2
7	de_wr_c1
8	de_wr_c2
9	en_sp
10	en_wr
11	es_sp_a1
12	es_sp_a2
13	es_sp_b1
14	es_sp_b2
15	es_sp_c1
16	es_sp_c2
17	es_wr_a1
18	es_wr_a2
19	es_wr_b1
20	es_wr_b2
21	es_wr_c1
22	es_wr_c2

Fonte: Autora

[78] http://corefl.learnercorpora.com/user_guide/corpus_design: COREFL in context

A seguir, para que os NGCSs pudessem ser identificados, os textos do *corpus* foram etiquetados por categoria semântica. Utilizou-se o etiquetador USAS (*UCREL*[79] *Semantic Analysis System* – UCREL Sistema de Análise Semântica), concebido na Universidade de Lancaster[80], por meio de um script desenvolvido pelo professor pesquisador que acessava o USAS automaticamente.

O USAS é um conjunto de ferramentas para a realização da análise semântica de textos escritos e falados (RAYSON et al., 2004, p. 2). A etiquetagem dos textos é feita com base em um conjunto de etiquetas que se referem dis, e incluem não apenas sinônimos e antônimos, mas também hiperônimos e hipônimos (ARCHER; WILSON; RAYSON, 2002, p. 1). Originalmente, as etiquetas semânticas usadas basearam-se nas classificações tipológicas do *Longman Lexicon of Contemporary English* (LLOCE), de McArthur (1981) (ARCHER; WILSON; RAYSON, 2002, p. 2). Devido a problemas práticos de etiquetagem ao longo de sua utilização, essas categorias foram revistas e modificadas. Por exemplo, a categoria "*Arts and crafts, science and technology, industry and education*"[81] do LLOCE foi dividida em três: '*Arts and crafts*', '*Science and technology' e 'Industry and education*' (RAYSON et al., 2004, p. 3). Além disso, como o USAS etiqueta todas as palavras em um texto, também foi acrescentada a categoria '*Names and grammatical words*'[82] (RAYSON et al., 2004, p. 4).

A etiquetagem semântica tem duas fases. Na Fase I (*Tag assignment* – Atribuição de etiqueta), anexa-se "um conjunto de etiquetas semânticas potenciais a cada unidade lexical"[83] e, na Fase II (*Tag desambiguation – desambiguação de etiquetas),* seleciona-se "a etiqueta semântica contextualmente apropriada do conjunto fornecido pela Fase I"[84] (RAYSON et al., 2004, p. 4).

O USAS foi inicialmente usado apenas para a língua inglesa, mas atualmente também está disponível para o chinês, o holandês, o finlandês, o francês, o italiano, o português, o espanhol e o galês.[85] Encontra-se disponível na internet de graça.

Presentemente, as etiquetas do USAS incluem 21 grandes campos de discurso, os quais, por sua vez, se ampliam em 232 categorias. São usadas letras para designar os principais campos semânticos, enquanto números são usados para

---

[79] UCREL originariamente era o acrônimo de Unit for Computer Research on the English Language (Unidade para Pesquisa Computacional sobre a Língua Inglesa). Em 1995 o nome foi mudado para University Center for Computer Corpus Research on Language (Centro Universitário para Pesquisa de Corpus Computacional sobre Linguagem) devido à mudança da natureza das pesquisas, que passaram a abarcar diferentes línguas, dentre outros projetos, assim como para ressaltar a importância desse centro na universidade, mas optou-se por manter o mesmo acrônimo. (https://ucrel.lancs.ac.uk/history.html#:~:text=Hence%20the%20Unit%20for%20Computer,retain%20the%20acronym%20of%20UCREL).

[80] USAS online English tagger (lancaster.ac.uk)

[81] Arte e artesanato, ciência e tecnologia, indústria e educação

[82] Nomes e palavras gramaticais - 'Grammatical words' são palavras consideradas 'vazias', tais como, pronomes, artigos, preposições, conjunções.

[83] "[...] *a set of potential semantic tags to each lexical unit* […]"

[84] […] *the contextually appropriate semantic tag from the set provided by Phase I*."

[85] https://ucrel.lancs.ac.uk/usas/

assinalar as subdivisões nesses campos. A Figura 4 contém o total das categorias e subcategorias semânticas do USAS.

FIGURA 4: CATEGORIAS E SUBCATEGORIAS DO ETIQUETADOR USAS

A GENERAL & ABSTRACT TERMS
A1 General
A1.1.1 General actions, making etc.
A1.1.2 Damaging and destroying
A1.2 Suitability
A1.3 Caution
A1.4 Chance, luck
A1.5 Use
A1.5.1 Using
A1.5.2 Usefulness
A1.6 Physical/mental
A1.7 Constraint
A1.8 Inclusion/Exclusion
A1.9 Avoiding
A2 Affect
A2.1 Affect: Modify, change
A2.2 Affect: Cause/Connected
A3 Being
A4 Classification
A4.1 Generally kinds, groups, examples
A4.2 Particular/general; detail
A5 Evaluation
A5.1 Evaluation: Good/bad
A5.2 Evaluation: True/false
A5.3 Evaluation: Accuracy
A5.4 Evaluation: Authenticity
A6 Comparing
A6.1 Comparing: Similar/different
A6.2 Comparing: Usual/unusual
A6.3 Comparing: Variety
A7 Definite (+ modals)
A8 Seem
A9 Getting and giving; possession
A10 Open/closed; Hiding/Hidden; Finding; Showing
A11 Importance
A11.1 Importance: Important
A11.2 Importance: Noticeability
A12 Easy/difficult
A13 Degree
A13.1 Degree: Non-specific
A13.2 Degree: Maximizers
A13.3 Degree: Boosters
A13.4 Degree: Approximators
A13.5 Degree: Compromisers
A13.6 Degree: Diminishers
A13.7 Degree: Minimizers

A14 Exclusivizers/particularizers
A15 Safety/Danger
B THE BODY & THE INDIVIDUAL
B1 Anatomy and physiology
B2 Health and disease
B3 Medicines and medical treatment
B4 Cleaning and personal care
B5 Clothes and personal belongings
C ARTS & CRAFTS
C1 Arts and crafts
E EMOTIONAL ACTIONS, STATES & PROCESSES
E1 General
E2 Liking
E3 Calm/Violent/Angry
E4 Happy/sad
E4.1 Happy/sad: Happy
E4.2 Happy/sad: Contentment
E5 Fear/bravery/shock
E6 Worry, concern, confident
F FOOD & FARMING
F1 Food
F2 Drinks
F4 Farming & Horticulture
G GOVT. & THE PUBLIC DOMAIN
G1 Government, Politics & elections
G1.1 Government etc.
G1.2 Politics
G2 Crime, law and order
G2.1 Crime, law and order: Law & order
G2.2 General ethics G3 Warfare, defence and the army; Weapons
H ARCHITECTURE, BUILDINGS, HOUSES & THE HOME
H1 Architecture, kinds of houses & buildings
H2 Parts of buildings
H3 Areas around or near houses
H4 Residence
H5 Furniture and household fittings
I MONEY & COMMERCE
I1 Money generally
I1.1 Money: Affluence
I1.2 Money: Debts
I1.3 Money: Price
I2 Business
I2.1 Business: Generally
I2.2 Business: Selling
I3 Work and employment
I3.1 Work and employment: Generally
I3.2 Work and employment: Professionalism
I4 Industry
K ENTERTAINMENT, SPORTS & GAMES
K1 Entertainment generally

K2 Music and related activities
K3 Recorded sound etc.
K4 Drama, the theatre & show business
K5 Sports and games generally
K5.1 Sports
K5.2 Games
K6 Children's games and toys
L LIFE & LIVING THINGS
L1 Life and living things
L2 Living creatures generally
L3 Plants
M MOVEMENT, LOCATION, TRAVEL & TRANSPORT
M1 Moving, coming and going
M2 Putting, taking, pulling, pushing, transporting &c.
M3 Movement/transportation: land
M4 Movement/transportation: water
M5 Movement/transportation: air
M6 Location and direction
M7 Places
M8 Remaining/stationary
N NUMBERS & MEASUREMENT
N1 Numbers
N2 Mathematics
N3 Measurement
N3.1 Measurement: General
N3.2 Measurement: Size
N3.3 Measurement: Distance
N3.4 Measurement: Volume
N3.5 Measurement: Weight
N3.6 Measurement: Area
N3.7 Measurement: Length & height
N3.8 Measurement: Speed
N4 Linear order
N5 Quantities
N5.1 Entirety; maximum
N5.2 Exceeding; waste
N6 Frequency etc**.**
O SUBSTANCES, MATERIALS, OBJECTS & EQUIPMENT
O1 Substances and materials generally
O1.1 Substances and materials generally: Solid
O1.2 Substances and materials generally: Liquid
O1.3 Substances and materials generally: Gas
O2 Objects generally
O3 Electricity and electrical equipment
O4 Physical attributes
O4.1 General appearance and physical properties
O4.2 Judgement of appearance (pretty etc.)
O4.3 Colour and colour patterns
O4.4 Shape
O4.5 Texture

O4.6 Temperature
P EDUCATION
P1 Education in general
Q LINGUISTIC ACTIONS, STATES & PROCESSES
Q1 Communication
Q1.1 Communication in general
Q1.2 Paper documents and writing
Q1.3 Telecommunications
Q2 Speech acts
Q2.1 Speech etc: Communicative
Q2.2 Speech acts
Q3 Language, speech and grammar
Q4 The Media
Q4.1 The Media: Books
Q4.2 The Media: Newspapers etc.
Q4.3 The Media: TV, Radio & Cinema
S SOCIAL ACTIONS, STATES & PROCESSES
S1 Social actions, states & processes
S1.1 Social actions, states & processes
S1.1.1 General
S1.1.2 Reciprocity
S1.1.3 Participation
S1.1.4 Deserve etc.
S1.2 Personality traits
S1.2.1 Approachability and Friendliness
S1.2.2 Avarice
S1.2.3 Egoism
S1.2.4 Politeness
S1.2.5 Toughness; strong/weak
S1.2.6 Sensible
S2 People
S2.1 People: Female
S2.2 People: Male
S3 Relationship
S3.1 Relationship: General
S3.2 Relationship: Intimate/sexual
S4 Kin
S5 Groups and affiliation
S6 Obligation and necessity
S7 Power relationship
S7.1 Power, organizing
S7.2 Respect
S7.3 Competition
S7.4 Permission
S8 Helping/hindering
S9 Religion and the supernatural
T TIME
T1 Time
T1.1 Time: General
T1.1.1 Time: General: Past

T1.1.2 Time: General: Present; simultaneous
T1.1.3 Time: General: Future
T1.2 Time: Momentary
T1.3 Time: Period
T2 Time: Beginning and ending
T3 Time: Old, new and young; age
T4 Time: Early/late
W THE WORLD & OUR ENVIRONMENT
W1 The universe
W2 Light
W3 Geographical terms
W4 Weather
W5 Green issues
X PSYCHOLOGICAL ACTIONS, STATES & PROCESSES
X1 General
X2 Mental actions and processes
X2.1 Thought, belief
X2.2 Knowledge
X2.3 Learn
X2.4 Investigate, examine, test, search
X2.5 Understand
X2.6 Expect
X3 Sensory
X3.1 Sensory: Taste
X3.2 Sensory: Sound
X3.3 Sensory: Touch
X3.4 Sensory: Sight
X3.5 Sensory: Smell
X4 Mental object
X4.1 Mental object: Conceptual object
X4.2 Mental object: Means, method
X5 Attention
X5.1 Attention
X5.2 Interest/boredom/excited/energetic
X6 Deciding
X7 Wanting; planning; choosing
X8 Trying
X9 Ability
X9.1 Ability: Ability, intelligence
X9.2 Ability: Success and failure
Y SCIENCE & TECHNOLOGY
Y1 Science and technology in general
Y2 Information technology and computing
Z NAMES & GRAMMATICAL WORDS
Z0 Unmatched proper noun
Z1 Personal names
Z2 Geographical names
Z3 Other proper names
Z4 Discourse Bin
Z5 Grammatical bin

Z6 Negative Z7 If Z8 Pronouns etc. Z9 Trash can Z99 Unmatched

Fonte: Archer, Wilson e Rayson (2002)

Como pode ser visto na Figura 4, as etiquetas semânticas aqui listadas são compostas de:

1. uma letra maiúscula, que indica o campo semântico principal;
2. um número, para indicar uma primeira subdivisão do campo semântico;
3. um ponto decimal seguido de um número, que indica uma subdivisão mais refinada do campo semântico (opcional) (ARCHER; WILSON; RAYSON, 2002, p. 2).

Além da composição indicada acima, há mais detalhes nas categorias semânticas do USAS para que a etiquetagem seja a mais precisa possível. Por exemplo, "um ou mais sinais de 'positivo' ou 'negativo' para indicar uma posição positiva ou negativa em uma escala semântica"[86], "uma barra seguida de uma segunda etiqueta indicando claramente dupla filiação de categorias"[87]. Também são usados outros símbolos, tais como, *f* para designar feminino, *m* para masculino (ARCHER; WILSON; RAYSON, 2002, p. 2-3). Abaixo, a título de ilustração, temos alguns exemplos de duas unidades lexicais etiquetadas:

1. *dowry* = S4/I1/A9-
2. *accountant* = I2.1/S2mf

Na primeira unidade lexical, *dowry* (dote), temos a etiqueta S4, separada por uma barra da etiqueta I1, separada por outra barra da etiqueta A9, seguida de um sinal negativo. A etiqueta S representa o campo semântico de 'ações, estados e processos sociais' (*Social actions, states & processes*) (idem, p. 25). O número 4 indica uma das primeiras subdivisões desse campo. A etiqueta S4 agrupa "termos relacionados a vínculos entre membros de uma família/ familiares"[88] (idem, p. 27) separada de outra etiqueta, I1, por uma barra, o que indica a dupla filiação de categorias. I representa outro campo semântico, que é relacionado a dinheiro e comércio na indústria (*Money & commerce in industry*) (idem, p. 14). O número 1 indica a primeira subdivisão desse campo semântico. A etiqueta semântica I1 reúne "termos relacionados a dinheiro em geral"[89] (idem, p. 14), e também está separada da

[86] "[…] *one or more 'pluses' or 'minuses' to indicate a positive or negative position on a semantic scale."*
[87] "[…] *a slash followed by a second tag to indicate clear double membership of categories."*
[88] "*Terms relating to relationships between family members/ familiars"*
[89] "*Terms relating to money generally"*

próxima etiqueta, A9, por outra barra, indicando a dupla, nesse caso tripla filiação da categoria. A categoria semântica A representa termos gerais e abstratos (*General & abstract terms*) (idem, p. 1). O número 9 assinala uma das primeiras subdivisões desse campo. A classe semântica A9 designa os campos de 'receber e dar: posse', juntando "termos gerais/ abstratos relacionados a alocação, abrir mão de, adquirir, ganhar, etc."[90] (idem, p. 7). O sinal negativo indica a posição negativa dessa unidade lexical em uma escala semântica.

A segunda unidade lexical, *accountant* (contador), foi etiquetada com I2.1, separada por uma barra de S2mf. O campo semântico I foi explicado acima. O número 2 indica uma das primeiras subdivisões desse campo. É seguido de um ponto decimal e outro número, 1, que assinala uma subdivisão mais refinada desse campo semântico. A etiqueta I2.1 representa 'negócios em geral' (*Business: generally*), "termos relacionados a negócios em geral"[91] (idem, p. 14). A barra assinala a dupla filiação semântica dessa unidade lexical. A próxima etiqueta, S2mf, tem por campo semântico o S, também explicado acima. O número 2 designa uma das primeiras subdivisões desse campo. S2 agrupa termos relacionados a 'pessoas' (*People*) "termos indicando palavras específicas relacionadas a/ que denotam pessoas, por exemplo, estudantes" [92] (idem, p. 27). O símbolo *m* representa palavras masculinas, e *f* femininas, indicando que essa unidade lexical tanto pode se referir a uma pessoa do sexo masculino ou do feminino.

Após a etiquetagem segundo os procedimentos acima, os textos do corpus foram pós-processados de tal modo que foi retida apenas a primeira etiqueta indicada para cada palavra, por ser aquela que possuía a maior probabilidade. Em seguida, foi mantida apenas a classificação primária da classe semântica, composta por uma letra e um número.

[90] "*General/ abstract terms relating to allocating/ relinquishing/acquiring/receiving, etc"*
[91] "*Terms relating to business generally*"
[92] "*Terms indicating that particular words relate to/denote people – e.g. students"*

2.3 CÁLCULO DA CHAVICIDADE DOS NGCSS

Na presente pesquisa, após a divisão do COREFL em subgrupos e a etiquetagem semântica de suas palavras, foi feita a extração dos n-gramas de classe semântica (NGCSs). Da lista inicial dos NGCSs, aqueles que apresentavam em sua composição a categoria Z99, atribuída pelo USAS quando não havia correspondência da palavra com as categorias disponíveis – seja porque houve algum erro de ortografia ou porque se tratava de alguma palavra ainda não incorporada ao dicionário do etiquetador (ARCHER; WILSON; RAYSON, 2002, p. 36) –, foram desconsiderados para a análise. Com os NGCSs restantes, um script identificou os NGCSs únicos (*type*) a fim de realizar a computação dos valores necessários para o cálculo da chavicidade (*keyness*).

Chavicidade é uma "medida que compara as frequências relativas de uma palavra em um texto versus o *corpus* de referência"[93]. Uma palavra tem alta chavicidade se é frequente em um texto e, simultaneamente, infrequente no *corpus* de referência (CANTOS GÓMEZ, 2013, p. 153). Dentre as várias medidas estatísticas possíveis, optou-se nesta pesquisa pelo log-likelihood (logaritmo da função verossimilhança), medida amplamente usada em estudos baseados em corpus (GABRIELATOS, 2018; EGBERT, BIBER, 2019), no Grupo de Estudos de Linguística de Corpus (GELC) e na suíte de programas de análise linguística WordSmith Tools, criado por Mike Scott em 1996 (BERBER SARDINHA, 2009).

Gabrielatos (2018) ressalta que embora tenha sido inicialmente usada como medida para palavras-chave, estudos exploratórios de chavicidade foram usados com outras unidades linguísticas, tais como lemas, n-gramas, unidades de várias palavras (*multi-word units*), classes de palavras (*part of speech tags*), padrões lexicogramaticais, e campos semânticos (2018, p. 228).

Para a etapa de computação dos valores necessários para o cálculo da chavicidade, foram formados corpora de estudo e de referência dinamicamente, de acordo com os subcorpora apresentados anteriormente. Para cada corpus de estudo, que corresponde a cada um dos subcorpora, o corpus de referência era formado por todos os demais subcorpora. Por exemplo, quando o *corpus* de estudo era *de_sp_b1*, seu corpus de referência eram todos os outros 21 subgrupos restantes (vide Figura 4). Esse processo de comparação dinâmica de cada corpus de estudo ao *corpus* de referência foi repetido 22 vezes, uma vez para cada corpus de estudo. Ou seja, os corpora de estudo e de referência não eram fixos, mas sim ajustados de acordo com o subcorpus sob análise. Todo esse processo foi conduzido automaticamente por meio de um script especializado preparado pelo professor orientador.

Outro script, também elaborado pelo professor orientador, foi usado para computar os valores necessários para o cálculo da chavicidade com base nas seguintes variáveis:

[93] "[…] *measure compares the relative frequencies of a word in a text versus a reference corpus*."

- *targetfreq*: a frequência do NGCS; nesta pesquisa é a contagem do número de textos nos quais essa variável ocorreu no subcorpus-alvo;
- *targetwca*: total de textos do subcorpus-alvo;
- *perthoua*: a frequência normalizada (por mil textos) do NGCS no subcorpus-alvo;
- *perthoub*: a frequência normalizada (por mil textos) do NGCS no subcorpus de referência;
- *reffreq*: a frequência do NGCS no corpus de referência nesta pesquisa a contagem do número de textos no qual essa variável ocorreu no subcorpus de referência;
- *refwca*: total de textos do subcorpus de referência;

Na Figura 5, temos um exemplo dos valores necessários de um NGCS para o cálculo da chavicidade.

FIGURA 5: EXEMPLO DOS VALORES OBTIDOS PARA O CÁLCULO DA CHAVICIDADE DE UM NGCS

NGCS	targetfreq	targetwca	perthoua	perthoub	reffreq	refwca
A10_A6_I3	3	77	38,91	4,72	9	1905

Fonte: Autora

Ao escolhermos por computar o número de textos nos quais cada NGCS ocorreu no corpus de estudo e no *corpus* de referência, em vez da frequência de ocorrência total dos NGCS, levamos em conta a dispersão dos NGCSs nos textos do COREFL. Podemos observar na Figura 5 acima, por exemplo, que o NGCS A10_A6_I3 ocorreu em três textos dos 77 textos do *corpus* de estudo, e em nove textos do corpus de referência, que tem um total de 1.905 textos. Dito de outra forma, levamos em consideração a distribuição dos NGCSs no COREFL, evitando-se assim que variáveis com alta frequência total, resultado da ocorrência frequente em poucos textos, causassem viés no cálculo (EGBERT; BIBER, 2019).

A seguir foi calculada a chavicidade de cada NGCS com base nos valores computados na etapa anterior. Como explicado acima, a medida usada para o cálculo da chavicidade foi o log-likelihood.

Após essa rotina, foram selecionados os NGCSs-chave de acordo com o valor da chavicidade calculado[94]. Um NGCS-chave é o que ocorre estatisticamente com mais frequência do que esperado em um subgrupo do *corpus* do que no *corpus* de

[94] Como se calcular log-likelihood: https://ucrel.lancs.ac.uk/llwizard.html

referência correspondente. Os NGCSs com valores de log-likelihood maiores do que 3,84 foram considerados NGCSs-chave. Os NGCSs com valores menores ou iguais a 3,84 não foram considerados NGCSs-chave[95].

Em seguida, foram selecionados os 100 NGCSs-chave de cada subcorpus-alvo, mais especificamente, aqueles com maior frequência normalizada por mil palavras.

Depois dessa seleção, a etapa seguinte foi a filtragem dos NGCSs-chave com base em índice de correlação, visando à robustez da Análise Fatorial (doravante AF). Para tal, foi feita uma matriz de correlação para detectar os NGCSs com maior atração mútua (BIBER, 1988). Assim, foram escolhidos os 400 NGCSs com maiores correlações positivas e as 400 variáveis com maiores correlações negativas. Essas duas etapas foram realizadas com scripts elaborados pelo professor orientador.

2.4 ANÁLISE FATORIAL

A AF é um procedimento estatístico multivariado empregado para "investigar as correlações subjacentes entre um grupo de variáveis observadas"[96] (LOEWEN; GONULAL, 2015, p. 182). É usada com a finalidade de revelar "quais variáveis em um grupo formam subgrupos coerentes que são relativamente independentes uns dos outros. Esses subgrupos de variáveis correlacionadas, mas em grande medida independentes de outros subgrupos de variáveis, são os fatores" [97] (TABACHNICK; FIDELL, 2014, p. 660). Pode-se dizer, portanto, que o propósito da AF é "definir o menor número possível de variáveis capaz de explicar um montante vultoso de variação nos dados"[98] (LOEWEN; GONULAL, 2015, p. 182), "sem que haja muita perda de informação"[99], e revelar padrões de relações latentes entre as variáveis (CANTOS GÓMEZ, 2013, p. 113).

Segundo Biber (1988), o propósito da AF é "reduzir o número de variáveis observáveis a um número relativamente pequeno de construtos subjacentes" [100] (BIBER, 1988, p. 82). Segundo Kaufmann (2020) e Egbert e Staples (2019), é o procedimento estatístico mais comumente empregado na AMD por ser "a técnica estatística ideal para modelar os padrões de MD de coocorrência linguística" [101]

---

[95] https://ucrel.lancs.ac.uk/llwizard.html

[96] "[...] *to investigate the underlying correlations among a set of observed variables."*

[97] "[...] *which variables in the set form coherent subsets that are relatively independent of one another. Variables that are correlated with one another but largely independent of other subsets are combined into factors."*

[98] "[…] *to determine the fewest number of variables that will still explain a substantial amount of variance in the data."*

[99] "[…] *with the minimum loss of information* […]"

[100] "[...] *to reduce the number of observable variables to a relatively small number of underlying constructs.*"

[101] "[...] *factor analysis provide the ideal statistical technique for modeling the MD patterns of linguistic*

(EGBERT; STAPLES, 2019, p. 125).

Para a realização da AF, é preciso escolher o método de extração dos fatores (BIBER, 1988, p. 81). Nesta pesquisa, em virtude de sua natureza exploratória[102], usou-se a AF por eixos principais (*principal axis factoring*, ou PAF), que procura levar em conta somente a **variação compartilhada** nas variáveis (BIBER, 1988; TABACHNICK; FIDELL, 2014; LOEWEN; GONULAL, 2015). Em outras palavras, a PAF escolhe as variáveis com alta covariação para a elaboração dos fatores, diferentemente da análise de componentes principais, também usada em estudos de variação linguística (EGBERT; STAPLES, 2019). Essa última se diferencia do método PAF por incluir **toda a variação** nos dados, englobando também a parte da variação proveniente de erro e a variação única de determinada variável (BIBER, 1988; TABACHNICK; FIDELL, 2014; LOEWEN;, GONULAL, 2015; EGBERT; STAPLES, 2019). A PAF segue o padrão normalmente empregado em estudos de AMD, por exemplo, Biber (1988), Kauffmann (2020), Zuppardi (2020), Delfino (2022) e Berber Sardinha (2023).

Uma vez definidas as variáveis para a AF e escolhido o método, foi realizada a extração inicial não rotacionada para que o maior número de fatores pudesse ser extraído (ZUPPARDI, 2020, p. 64). Em seguida, avaliou-se a comunalidade das variáveis e determinou-se o número ideal de fatores para a extração final. Comunalidade (h^2) "indica a relação de cada variável com todo o conjunto de dados"[103] (LOEWEN; GONULAL, 2015, p. 190). Uma comunalidade alta assinala que os fatores extraídos representam bem as variáveis. Valores baixos, por outro lado, apontam que as variáveis não se encaixam bem na solução fatorial, e deveriam ser descartadas (CANTOS GÓMEZ, 2013). Nesta pesquisa, variáveis com comunalidade inferior a 0,15 foram descartadas (BIBER, 2006, p. 183; BERBER SARDINHA, 2023, p. 72).

Existem diversos procedimentos para a determinação do número de fatores (BIBER, 1988; TABACHNICK; FIDELL, 2014; LOEWEN; GONULAL, 2015). Normalmente, analisam-se os autovalores (*eigenvalues*), com base em um gráfico de sedimentação (*scree plot*) (BIBER, 1988; EGBERT; STAPLES, 2019; ZUPPARDI, 2020; KAUFFMANN, 2020; DELFINO, 2022; BERBER SARDINHA, 2023). Um autovalor é "uma medida de quanta variação um fator explica nos dados – quanto mais alto, melhor" (BREZINA, 2018, p. 166). Segundo Brezina, um autovalor é "a soma dos quadrados das cargas fatoriais de todas as variáveis" (2018, p. 166). De acordo com Egbert e Staples (2019), os "fatores com autovalor maior do que 1 poderiam ser considerados para a fatoração" (2019, p. 129). Mas, como na primeira extração

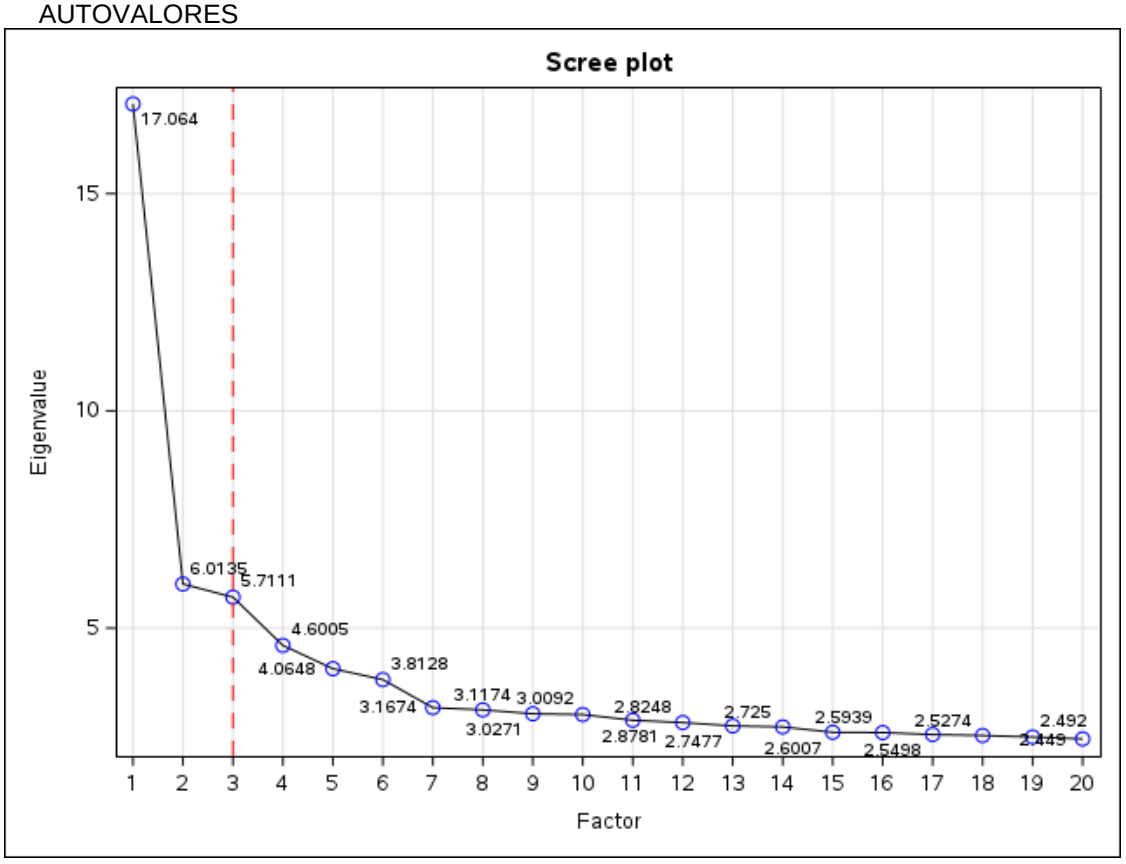
co-occurrence [...]"

[102] Uma pesquisa de natureza exploratória pretende "descrever e condensar os dados ao agrupar variáveis que se correlacionam" (TABACHNIK; FIDELL, 2014, p. 662) em um grupo de textos que não tenha sido anteriormente examinado (EGBERT; STAPLES, 2019).

[103] "[...] *communalities (h^2) can provide an indication of the relationship of each variable to the entire data set.*"

sempre há um número considerável de fatores com esse perfil, há que se levar em conta o gráfico de sedimentação, que é uma representação visual dos autovalores, organizados por ordem descendente (TABACHNICK; FIDELL, 2014). Para a seleção do número de fatores é preciso examinar o gráfico e observar onde há um ponto de inflexão, que é "onde a queda acentuada de valores termina e a curva do gráfico começa a ficar mais plana"[104] (BREZINA, 2018, p. 166). O número de autovalores à esquerda desse ponto de inflexão é selecionado para a AF final. Loewen e Gonulal (2015), Egbert e Staples (2019) e Delfino (2022), dentre outros, ressaltam o caráter subjetivo dessa etapa. Na Figura 6, podemos ver o gráfico de sedimentação da AF inicial.

FIGURA 6: GRÁFICO DE SEDIMENTAÇÃO DA SOLUÇÃO NÃO ROTACIONADA COM AUTOVALORES



Fonte: Elaborado pela autora

Como se pode observar, há dois pontos de inflexão, um no fator 2, mais acentuado e outro no fator 4, menos acentuado. Assim, foram experimentadas soluções de dois, três e quatro fatores, sendo a de três aquela considerada mais interpretável.

Uma vez determinado em três o número de fatores, foi feita a extração final rotacionada. Segundo Biber (1988), essa é uma etapa importante uma vez que cada

[104] "[...] *where the sharp drop in eigenvalues ends and the curve of the plot starts levelling off.*"

fator explica a maior quantidade possível de variação. Ou seja, cada fator captura a maior quantidade de variação disponível, deixando o restante para os fatores subsequentes. Em outras palavras, cada fator responde por uma fatia maior da variação do que os consecutivos. Com a extração rotacionada, "cada fator é caracterizado pelas poucas características que são mais representativas de uma determinada quantidade de variação compartilhada"[105] (BIBER, 1988, p. 84). Dito de outra forma, a rotação sublinha a força com que cada variável carrega em cada fator ao distribuir os carregamentos pelos fatores, simplificando e facilitando a sua posterior interpretação (LOEWEN, GONULAL, 2015).

Há basicamente dois tipos de métodos de rotação: a ortogonal e a oblíqua. A ortogonal é usada quando se espera que os fatores sejam independentes ou não se relacionem muito; a oblíqua é utilizada quando se espera que haja correlação. (BIBER, 1988; LOEWEN; GONULAL, 2015; EGBERT; STAPLES, 2019). O método de rotação oblíqua usado foi o Promax, acatando a sugestão de Biber (1988), uma vez que variáveis linguísticas são normalmente interrelacionadas (EGBERT; STAPLES, 2019). Essa extração reteve 323 variáveis, na forma de NGCSs.

A próxima etapa foi o exame das cargas fatoriais dos NGCSs. Cargas fatoriais indicam "a força de associação entre cada variável e cada fator" [106] (LOEWEN; GONULAL, 2015, p. 199). Como de praxe na AMD, assim como nesta pesquisa, as variáveis com valor de carregamento menores do que 0,30 foram descartadas (BIBER, 1988, p. 87).

As variáveis podem ter cargas fatoriais positivas ou negativas, o que não significa que sejam mais ou menos importantes (BIBER, 1988; EGBERT; STAPLES, 2019). Um carregamento positivo ou negativo apenas reflete a relação positiva ou negativa de uma variável com a composição geral dos fatores (EGBERT; STAPLES, 2019). Segundo Biber (1988) e Egbert e Staples (2019), um fator com carregamentos positivos e negativos "exibe conjuntos de características que estão distribuídas nos textos de forma complementar"[107] (BIBER, 1988, p. 88). Ou seja, quando variáveis positivas ocorrem em um texto, as variáveis negativas tendem a não ocorrer, e quando variáveis negativas ocorrem, as positivas tendem a não estar presentes. As cargas fatoriais positivas, chamamos de polo fatorial positivo; as negativas, chamamos de polo fatorial negativo.

Segundo Biber (1988), é desejável ter no mínimo cinco variáveis em um polo fatorial, para permitir uma interpretação robusta dos padrões de relações latentes entre as variáveis. Assim, polos com menos do que cinco variáveis foram descartados na interpretação dos fatores em dimensões, em etapa subsequente da pesquisa.

A etapa seguinte foi a padronização dos valores dos NGCSs. Essa etapa é necessária porque "dá sentido e significado aos escores possibilitando a sua

---

105 "[...] *each factor is characterized by the few features that are most representative of a particular amount of shared variance."*

106 "[...] *the strength of the association between each variable and each factor.*"

107 "[…] *show groups of features that are distributed in texts in a complementary pattern.*"

interpretação e permite a comparação direta entre dois ou mais escores"[108] (CANTOS GÓMEZ, 2013, p. 10) É comumente usada em AMD para o cálculo posterior dos escores dos fatores (BIBER, 1988; ZUPPARDI, 2020).

Os escores das variáveis, nesta pesquisa os NGCSs, foram padronizados em uma única escala: os escores Z (CANTOS GÓMEZ, 2018; ZUPPARDI, 2020), os quais são usados para dar a cada variável o mesmo peso, independentemente de "ser frequente ou rara" (BIBER, 1988; BREZINA, 2018, p. 168). Expressam "quão distantes um dado escore bruto está da média em unidades de desvio padrão"[109] (CANTOS GÓMEZ, 2013, p. 12).[110]

Para ilustrar a relevância desse procedimento, em duas tabelas (6A e 6B), Cantos Gómez (2013) dá o exemplo hipotético da nota de 5 alunos em dois testes diferentes, e demonstra como fica mais fácil a análise da performance dos alunos após a padronização das variáveis (2013, p. 12). Na Tabela 6A, abaixo, temos os resultados brutos.

TABELA 6A: RESULTADOS EM DOIS TESTES DIFERENTES DE LÍNGUA

	ALUNO 1	ALUNO 2	ALUNO 3	ALUNO 4	ALUNO 5
Teste **A**	120	90	130	125	110
Teste **B**	4	3	5	4	3

Fonte: Cantos Gómez, 2013, p. 10

Não é possível comparar, por exemplo, o resultado do aluno 1 no teste A com o resultado do aluno 3 no teste B, uma vez que são dois testes diferentes. A seguir, a Tabela 6B tem os resultados dos alunos, agora padronizados.

TABELA 6B: COMPARAÇÃO DOS ESCORES Z PARA OS TESTES A E B

	ALUNO 1	ALUNO 2	ALUNO 3	ALUNO 4	ALUNO 5
Teste **A**	0,31	- 1,58	0,94	0,63	- 0,31
Teste **B**	0,24	- 0,96	1,44	0,24	- 0,96

[108] "[…] *it gives sense and meaning to the scores and it allows their interpretation, and it allows direct comparison between two scores or more."*

[109] "[...] *in standard deviation units how far a score is from the mean.*"

[110] A média é sempre zero após a padronização dos escores. O valor positivo indica que o escore é acima da média e o negativo que está abaixo (CANTOS GÓMEZ, 2013).

Fonte: Cantos Gómez, 2013, p. 12

Com essa padronização pode-se observar, por exemplo, que a performance do aluno 1 foi semelhante nos dois testes e o aluno 5, por sua vez, ficou abaixo da média nos dois testes, sendo que foi pior no teste B. O aluno 3 também foi bem melhor que o aluno 1, tanto no teste A quanto no teste B. Esses dados não estavam claros antes (CANTOS GÓMEZ, 2013).

A seguir, foram calculados os escores de fator de cada texto somando-se os escores Z de cada variável componente de cada fator, tipificando, dessa forma, o texto em relação a cada fator. Ou seja, os textos foram pontuados, recebendo um escore para cada fator, como pode ser observado na Tabela 7, abaixo.

TABELA 7: EXEMPLO DE ESCORES DE FATORES EM CADA TEXTO

TEXTO	FATOR 1	FATOR 2	FATOR 3
t000186	9,47185945	59,97835953	2,903653959
t000457	29,56981669	6,652714297	108,4836458
t000599	51,22642143	4,981316946	5,64927826

Fonte: Autora

Analisando a tabela acima, de acordo com os escores computados para cada texto em cada fator, podemos dizer que o texto t000186 é mais representativo do fator 2, o t000457 do fator 3, e o t000599 do fator 1.

Após efetuar a pontuação, foram analisadas qualitativamente dezenas de textos com os maiores escores em cada fator e anotados os NGCSs que ocorreram em cada texto, a fim de auxiliar a interpretação dos fatores por meio de análise qualitativa. Os textos foram formatados para facilitar a análise interpretativa qualitativa, por meio de um script desenvolvido pelo professor orientador. Em seguida, foram listados os NGCSs mais frequentes nesses textos. Também foram criadas duas nuvens de palavras, também por intermédio de um script especializado criado pelo professor orientador: uma com as etiquetas semânticas com maior peso em cada fator, e a outra com as palavras mais usadas nessas etiquetas semânticas.

Resumidamente, a metodologia consistiu nas seguintes etapas (adaptado de Berber Sardinha, 2023b):

1. Etiquetagem semântica;
2. Extração dos NGCSs e seleção dos NGCSs;
3. Normalização e classificação dos NGCSs;
4. AF inicial não rotacionada;
5. AF rotacionada;
6. Cálculo dos escores de fator, em cada texto;
7. Análise qualitativa: interpretação dos fatores e nomeação das dimensões.

3. RESULTADOS E DISCUSSÃO

Nesta seção apresentamos e analisamos os resultados alcançados em resposta às perguntas que orientaram este trabalho. A primeira delas foi identificar as dimensões de variação de uso de NGCSs na fala e na escrita de alunos de inglês como língua estrangeira; a segunda buscou elucidar o quanto da variação no uso desses NGCSs pode ser explicada pelo modo (falado ou escrito), pela língua materna (espanhol, alemão ou inglês), pelo nível de proficiência dos alunos (A1, A2, B1, B2, C1, C2), pela idade, pela quantidade de anos que estudou a língua inglesa ou pela tarefa designada aos alunos. Para tal, foi feita a análise qualitativa dos fatores e a nomeação das dimensões assim como a análise de variância (ANOVA) em cada dimensão.

3.1 FATOR 1

Por sua natureza, na AF, o primeiro fator é o que concentra a maior porcentagem de variação explicada do *corpus* (cf. BIBER, 1988; KAUFFMAN, 2020; DELFINO, 2022), uma vez que extrai “o maior grupo de coocorrências nos dados” (BIBER, 1988, p. 82).

A partir dos resultados da AF rotacionada, chegamos à Tabela 8 que foi organizada em ordem decrescente dos pesos das 38 variáveis que carregaram neste fator.

TABELA 8: PADRÃO FATORIAL DO FATOR 1

	Variável - NGCS	**Peso**
1	Grammatical bin (Z5): Age (T3): Grammatical bin (Z5)	0,64003
2	Getting (A9): Grammatical bin (Z5): Age (T3)	0,61786
3	Grammatical bin (Z5): Age (T3): Direction (M6)	0,57462
4	Age (T3): Direction (M6): Grammatical bin (Z5)	0,55380
5	Grammatical bin (Z5): Grammatical bin (Z5): Age (T3)	0,54773
6	Direction (M6): Grammatical bin (Z5): Transport (M3)	0,52794
7	Closed (A10): Grammatical bin (Z5): Age (T3)	0,52580
8	Age (T3): Grammatical bin (Z5): Grammatical bin (Z5)	0,48800
9	Coming (M1): Grammatical bin (Z5): Age (T3)	0,47566

10	Grammatical bin (Z5): Grammatical bin (Z5): People (S2)	0,47532
11	Age (T3): Grammatical bin (Z5): Pronouns (Z8)	0,46673
12	Getting (A9): Concern (E6): Grammatical bin (Z5)	0,45941
13	Grammatical bin (Z5): Age (T3): Getting (A9)	0,42994
14	Grammatical bin (Z5): Getting (A9): Grammatical bin (Z5)	0,42473
15	Grammatical bin (Z5): Crime (G2): Sensory (X3)	0,41207
16	Grammatical bin (Z5): Getting (A9): Concern (E6)	0,40861
17	Direction (M6): Grammatical bin (Z5): Age (T3)	0,40566
18	Crime (G2): Sensory (X3): Pronouns (Z8)	0,40130
19	Grammatical bin (Z5): Age (T3): Frequency (N6)	0,40080
20	Being (A3): Coming (M1): Direction (M6)	0,39920
21	Grammatical bin (Z5): Transport (M3): Grammatical bin (Z5)	0,39053
22	Deciding (X6): Grammatical bin (Z5): Getting (A9)	0,37224
23	Grammatical bin (Z5): People (S2): Grammatical bin (Z5)	0,36401
24	Pronouns (Z8): Trying (X8): Grammatical bin (Z5)	0,36303
25	(Pulling (M2): Grammatical bin (Z5): Age (T3)	0,36271)
26	Grammatical bin (Z5): Age (T3): Being (A3)	0,36246
27	Trying (X8): Grammatical bin (Z5): Getting (A9)	0,35435
28	Being (A3): Negative (Z6): Pronouns (Z8)	0,35268
29	Pronouns (Z8): Being (A3): Coming (M1)	0,34221
30	(Direction (M6): Pronouns (Z8): Closed (A10)	0,34166)
31	Grammatical bin (Z5): People (S2): Pronouns (Z8)	0,32480
32	Transport (M3): Grammatical bin (Z5): Pronouns (Z8)	0,32343
33	Mental processes (X2): Pronouns (Z8): Grammatical bin (Z5)	0,31121
34	Concern (E6): Grammatical bin (Z5): Pronouns (Z8)	0,30857
35	Getting (A9): Pronouns (Z8): Grammatical bin (Z5)	0,30770
36	Personal (Z1): Getting (A9): Grammatical bin (Z5)	0,30470

37	Grammatical bin (Z5): People (S2): Sensory (X3)	0,30158
38	Concern (E6): Grammatical bin (Z5): Grammatical bin (Z5)	0,30016

Fonte: Elaborado pela autora

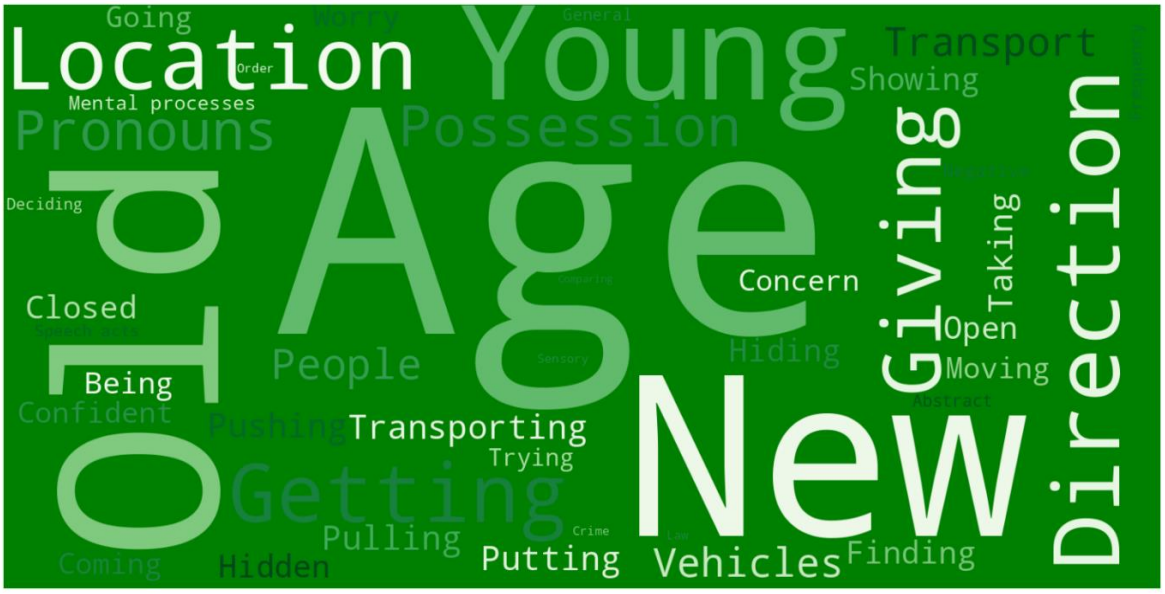
O polo negativo no Fator 1 foi desconsiderado porque havia apenas 3 variáveis, quantidade insuficiente para permitir a sua interpretação (cf. BIBER, 1988, p. 88).

Pode-se observar também na Tabela 8 que 37% das classes semânticas dos NGCSs que carregaram neste fator são Z5, que significa literalmente 'caixa ou depósito gramatical' (*grammatical bin*). Essa classe semântica agrupa preposições, advérbios, por exemplo, *before, after*[111], conjunções etc. (ARCHER; WILSON; RAYSON, 2002, p. 36). Já 12,6% são T3, que significa 'Tempo: velho, novo e jovem; idade' (*Time: old, new and young; age*), que agrupa termos relacionados à idade/ maturidade (idem, p. 30). Foram as etiquetas que mais carregaram.

Os NGCSs 25 e 30 estão em parênteses porque, além de terem carregado no Fator 1, também carregaram no Fator 2, só que com maior peso. O carregamento dessas variáveis no Fator 1, por ter sido menor, ainda que acima do valor de corte, como explicado na metodologia, chama-se carregamento secundário. O carregamento com maior peso (das variáveis 25 e 30 no Fator 2) denomina-se carregamento primário.

A ocorrência das etiquetas semânticas que carregaram no Fator 1 pode ser visualmente observada na nuvem de palavras abaixo (Figura 7).

FIGURA **7**: NUVEM DE PALAVRAS COM AS ETIQUETAS SEMÂNTICAS QUE CARREGARAM NO FATOR 1



Fonte: Elaborado pela autora

O objetivo dessa nuvem de palavras foi destacar visualmente as etiquetas semânticas que carregaram neste fator. Dito de outra forma, as categorias principais

[111] antes, depois

foram acentuadas, realçando o que há de mais importante. Foram incluídas as etiquetas semânticas que possuem significado semântico substancial, em inglês content words (palavras com conteúdo), como substantivos, verbos, adjetivos e advérbios. A categoria Z5, caixa gramatical, conforme já explicado, é uma etiqueta destinada a palavras gramaticais, não foi incluída ao se gerar essa nuvem de palavras de etiquetas semânticas.

Podemos observar nessa nuvem a proeminência de termos relacionados à idade, etiqueta semântica T3. As outras classes que sobressaem são relacionadas a verbos que descrevem ações: dar, esconder, encontrar, obter, deslocar, transportar, processos mentais (*giving, hiding, finding, getting, going/ moving, transporting, mental processes).*

Para auxiliar na interpretação desta dimensão foi gerada outra nuvem, desta vez com a ocorrência das palavras presentes nas categorias semânticas que carregaram no Fator 1**.** Na primeira, há uma menor concentração de itens, pois há menos categorias do que palavras. Na segunda, além das palavras com significado semântico substancial, as chamadas *content words,* também aparecem as que são classificadas como caixa gramatical (*grammatical bin*) como podemos ver na Figura 8 abaixo.

FIGURA 8: NUVEM DE PALAVRAS COM AS PALAVRAS PRESENTES NAS ETIQUETAS SEMÂNTICAS DO FATOR 1

Fonte: Elaborado pela autora

Podemos observar que há menor concentração de itens na Figura 8 do que na Figura 9, uma vez que há menos categorias do que palavras.

Nessa última figura, as palavras que mais carregaram nesse fator incluem *baby, take, woman, street, he, him, care, finds, found, where, tries, into, inside,*

pram[112], além de vários verbos que descrevem ações: *gives, puts, walking, give, know, leave, keep, get, places, do, left, sees, decided*[113], sugerindo a narração de ações de personagens - um bebê, uma mulher, um policial - numa trama.

Também foram analisadas as sequências de palavras que mais carregaram nos NGCSs. O primeiro deles, de maior peso no Fator 1, foi Z5 (caixa gramatical) _T3 (Tempo: Velho, novo e jovem; idade) _Z5 (caixa gramatical). Na Tabela 9 podemos observar a frequência assim como as sequências de palavras que compartilham as mesmas categorias semânticas desse NGCS.

TABELA 9: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 1

Contagem	NGCS 1	Sequências de palavras
742	Z5_T3_Z5	the_baby_and
406	Z5_T3_Z5	the_baby_to
248	Z5_T3_Z5	the_baby_with
150	Z5_T3_Z5	the_baby_into
62	Z5_T3_Z5	the_baby_so
40	Z5_T3_Z5	a_baby_and
31	Z5_T3_Z5	the_baby_‘
26	Z5_T3_Z5	the_baby_because
24	Z5_T3_Z5	the_baby_at
18	Z5_T3_Z5	the_baby_the

Fonte: Elaborado pela autora

O segundo NGCS que mais carregou foi A9_Z5_T3, sendo que a categoria semântica A9 significa ‘recebendo e dando: posse’[114]. As palavras agrupadas no campo semântico de A9 são termos gerais e abstratos que se relacionam a atribuir, ceder, adquirir, receber[115] etc. (ARCHER; WILSON; RAYSON, 2002, p. 7). Na Tabela 10, podemos observar as sequências de palavras.

TABELA 10: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 1

Contagem	NGCS 2	Sequências de palavras
377	A9_Z5_T3	takes_the_baby
157	A9_Z5_T3	gives_the_baby

[112] bebê, apanhar, mulher, rua, ele, lhe/ o, cuidado, encontra, encontrou, onde, tenta, em, dentro, carrinho de bebê
[113] dá, põe, andando, dão, sabe, deixa, guarda, consegue, coloca, faz, deixou, vê, decidiu
[114] *Getting and giving: possession*
[115] *Allocating/ relinquishing/acquiring/ receiving, etc.*

118	A9_Z5_T3	give_the_baby
94	A9_Z5_T3	keep_the_baby
92	A9_Z5_T3	take_the_baby
82	A9_Z5_T3	took_the_baby
72	A9_Z5_T3	holding_the_baby
46	A9_Z5_T3	gave_the_baby
28	A9_Z5_T3	taking_the_baby
27	A9_Z5_T3	has_the_baby

Fonte: Elaborado pela autora

A terceira variável foi Z5_T3_M6, sendo que M6 significa 'localização e direção' [116] . As palavras agrupadas nesse campo semântico são termos que descrevem a posição de/ ponto de referência para X[117] (idem, p. 18). Abaixo, na Tabela 11, há exemplos de sequências de palavras desse NGCS.

TABELA 11: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 1

Contagem	NGCS 3	Sequências de palavras
771	Z5_T3_M6	the_baby_in
141	Z5_T3_M6	the_baby_on
100	Z5_T3_M6	a_baby_on
98	Z5_T3_M6	a_baby_in
94	Z5_T3_M6	the_baby_there
79	Z5_T3_M6	the_baby_where
78	Z5_T3_M6	the_baby_inside
64	Z5_T3_M6	the_baby_for
30	Z5_T3_M6	the_baby_away
19	Z5_T3_M6	the_baby_by

Fonte: Elaborado pela autora

A quarta variável foi T3_M6_Z5. Abaixo a Tabela 12 apresenta exemplos de sequências de palavras desse NGCS.

[116] Location and direction
[117] Terms depicting position of/ point of reference for X.

TABELA 12: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 1

Contagem	NGCS 4	Sequências de palavras
474	T3_M6_Z5	baby_in_the
212	T3_M6_Z5	baby_on_the
92	T3_M6_Z5	baby_in_a
47	T3_M6_Z5	baby_for_a
27	T3_M6_Z5	baby_inside_the
13	T3_M6_Z5	baby_inside_of
12	T3_M6_Z5	baby_on_a
11	T3_M6_Z5	baby_there_and
9	T3_M6_Z5	baby_inside_and
9	T3_M6_Z5	baby_in_an

Fonte: Elaborada pela autora

Abaixo, na Tabela 13, a quinta variável Z5_Z5_T3 e exemplos de sequências de palavras desse NGCS.

TABELA 13: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 1

Contagem	NGCS 5	Sequências de palavras
606	Z5_Z5_T3	with_the_baby
348	Z5_Z5_T3	of_the_baby
157	Z5_Z5_T3	with_a_baby
145	Z5_Z5_T3	at_the_baby
62	Z5_Z5_T3	to_an_old
46	Z5_Z5_T3	to_the_baby
26	Z5_Z5_T3	into_the_baby
26	Z5_Z5_T3	and_the_baby
23	Z5_Z5_T3	and_the_old
22	Z5_Z5_T3	the_the_baby

Fonte: Elaborado pela autora

Em todos os exemplos acima de sequências de palavras, aparece a palavra

baby, às vezes seguida de conjunções (*because, and, so* [118]) ou preposições (*to, into, at* [119]), outras vezes precedida por verbos reunidos na categoria semântica A9 (recebendo e dando: posse), preposições (*with, of* [120]*, at, to, into*) ou artigos: definidos (*the*[121]), na sua grande maioria, ou indefinido (*a*[122]). Tal fato, indica a relevância de *baby* no Fator 1.

Para a análise qualitativa desta dimensão, foram analisados os textos que mais carregaram no Fator 1. A tarefa feita nesses textos foi a 14, que descreve uma parte do filme O garoto (1921), dirigido e estrelado por Charles Chaplin. Observou-se que 46 foram de produção escrita. A distribuição por proficiência, modo (escrito ou falado), a língua materna (L1), idade dos aprendizes assim com o total de anos estudando o idioma inglês podem ser verificados nas Tabelas 14 e 15, abaixo.

TABELA 14: DISTRIBUIÇÃO POR PROFICIÊNCIA, MODO E LÍNGUA MATERNA DOS 50 TEXTOS QUE MAIS CARREGARAM NO FATOR 1

FATOR 1				
Nível de Proficiência	Produção Escrita	Produção Oral	L1: Espanhol	L1: Alemão
A1	3		3	
A2	9		9	
B1	15	1	16	
B2	12	1	12	1
C1	5	3	7	1
C2				
L1: Inglês	1			

Fonte: Elaborado pela autora

TABELA 15: DISTRIBUIÇÃO POR IDADE E ANOS DE ESTUDO DA LÍNGUA INGLESA DOS 50 TEXTOS QUE MAIS CARREGARAM NO FATOR 1

FATOR 1			
Grupos por idade	Total de aprendizes	Anos de estudo da língua inglesa	Total de aprendizes
1 (<=15)		1 (<=3)	2
2 (>15 até <=20)	15	2 (>3 até <=6)	4
3 (>20 até <=30)	33	3 (>6 até <=9)	4
4 (>30 até <=40)	2	4 (>9 até <=12)	8
5 (>40)		5 (>12)	31
		L1 - inglês	1

Fonte: Elaborado pela autora

[118] porque, e, então
[119] para, em, na/no
[120] com, da/das/do/dos
[121] o - escolheu-se para a tradução apenas o artigo definido masculino porque nos exemplos dados essa palavra refere-se a 'bebê', substantivo masculino em português.
[122] um - escolheu-se para a tradução apenas o artigo indefinido masculino porque nos exemplos dados essa palavra refere-se a 'bebê', substantivo masculino em português.

Na Tabela 14, com a distribuição por proficiência, modo e língua materna, podemos observar que além da produção ter sido majoritariamente escrita, com apenas 4 dos textos sendo de produção oral, 60% dos aprendizes não têm um nível muito alto de proficiência, de A1 até B1. Apenas dois aprendizes têm por língua materna o alemão, e um deles é um falante nativo da língua inglesa; os outros 47% têm o como língua materna o espanhol. A maioria, 66%, têm de 20 até 30 anos, 30% de 15 até 20, e somente dois aprendizes têm de 30 até 40. Um detalhe que chamou a atenção é que, embora não sejam muito proficientes, 62% dizem ter estudado inglês por mais de 12 anos.

Abaixo podemos observar três textos. Não foi feita nenhuma adaptação da grafia das composições nos exemplos.

O primeiro exemplo (arquivo t001271) foi o texto com o escore mais alto no Fator 1: 93,3622475. Pelos metadados do arquivo, es_wr_b1_19_7_14_emgv, sabemos que a sua L1 é o espanhol (**es**_wr_b1_19_7_14_emgv), que a tarefa foi escrita (es_**wr**_b1_19_7_14_emgv), que o seu nível de proficiência é B1 (es_wr_**b1**_19_7_14_emgv), que o aluno tem 19 anos (es_wr_b1_**19**_7_14_emgv), que estuda inglês há 7 (es_wr_b1_19_**7**_14_emgv), que fez a tarefa 14 (es_wr_b1_19_7_**14**_emgv), que suas iniciais são emgv (es_wr_b1_19_7_14_**emgv**)[123].

> Charles Chaplin is Walking in the street when he stops to smoke a cigarette and he find *a baby on*[124] the floor.
> He *takes the baby*[125] *and*[126] tries to give the baby[127] to a woman with another baby but the woman doesn't want so he tries to give *the baby*[128] *to*[129] *a* men who is walking in the street but he puts the baby with the woman other time and the woman gets angry with Charles Chaplin and gives he back. Finally, Charles Chaplin find a note in the baby's[130] blanket that says please take care of him and Charles decides to keep the baby.

O segundo exemplo (arquivo t001104) foi o texto com o segundo escore mais alto: 90,0878076. Pelos metadados do arquivo, es_wr_a2_25_1_14_ag, sabemos que a sua L1 também é o espanhol (**es**_wr_a2_25_1_14_ag), que a tarefa foi escrita (es_**wr**_a2_25_1_14_ag), que o seu nível de proficiência é A2

[123] COREFL (learnercorpora.com)
[124] Z5 (caixa gramatical)_T3 (idade)_Z5 (caixa gramatical)
[125] A9_Z5_T3 segunda variável, o campo semântico de A9 é recebendo e dando: posse.
[126] Z5_T3_Z5
[127] A9_Z5_T3
[128] A9_Z5_T3
[129] Z5_T3_Z5
[130] Z5_T3_Z5

(es_wr_**a2**_25_1_14_ag), que o aluno tem 25 anos (es_wr_a2_**25**_1_14_ag), que estuda inglês há 1 ano (es_wr_a2_25_**1**_14_ag), que fez a tarefa 14 (es_wr_a2_25_1_**14**_ag), que suas iniciais são ag (es_wr_a2_25_1_14_**ag**).

Chaplin found a baby[131] and he decided to look the baby's family. First, he found a woman with a baby[132] carriage but she wasn't his mother.

Then, Chaplin left the baby[133] at the same place where he found, but a policeman saw him. Also, Chaplin gave the baby[134] to[135] a man and he went away.

The man put the baby at[136] the same baby carriage and the woman saw chaplin again and she hit him. Finally, Chaplin decided adopt the baby.

O terceiro exemplo (arquivo t001352) foi o texto com o quarto escore mais alto: 72,922683. Pelos metadados do arquivo, es_wr_b1_23_17_14_aj, sabemos que a sua L1 é o espanhol (**es**_wr_b1_23_17_14_aj), que a tarefa foi escrita (es_**wr**_b1_23_17_14_aj), que o seu nível de proficiência é B1 (es_wr_**b1**_23_17_14_aj), que o aluno tem 23 anos (es_wr_b1_**23**_17_14_aj), que estuda inglês há 17 anos (es_wr_b1_23_**17**_14_aj), que fez a tarefa 14 (es_wr_b1_23_17_**14**_aj), que suas iniciais são aj (es_wr_b1_23_17_14_**aj**).

The video is about the character Chaplin who finds a little *baby in the*[137] street.

Firstly, he tries to leave *the baby with*[138] a woman who was passing through the same street, but she realized he was not her baby, she gave back *the baby to*[139] Chaplin, and he left the *baby in the*[140] same place he found him, but suddenly, he turned around and found a policeman, and he had to *keep the baby*[141] *with*[142] him. He was sitting on the sidewalk and he found a note in the baby clothes.

[131] A10_Z5_T3; sétima variável; A10 - o campo semântico dessa etiqueta é: Aberto/ fechado; escondendo/ escondido; encontrando; mostrando – termos gerais e abstratos relacionados a (nível de) abertura/ ocultação/ visibilidade etc. *(open/closed; hiding/ hidden; finding; showing General abstract terms relating to (level of) openness/ concealment/ exposure* etc.)
[132] Z5_Z5_T3
[133] M1_Z5_T3, nona variável, M1 – o campo semântico dessa etiqueta é movimento: movendo, vindo, indo - termos retratando movimento (*moving, coming, going – terms depicting movement*)
[134] A9_Z5_T3
[135] Z5_T3_Z5
[136] Z5_T3_Z5
[137] T3_Z5_Z5
[138] Z5_T3_Z5
[139] Z5_T3_Z5
[140] T3_Z5_Z5
[141] A9_Z5_T3
[142] Z5_T3_Z5

Nos exemplos acima, podemos verificar que os textos são sucintos ao descrever a interação entre os personagens. Há o uso adequado dos poucos conectivos utilizados, e alguns erros de diferentes naturezas, tais como concordância verbal (*he find*), padrão verbal (decided adopt), o uso de determinantes (*Other time*), dentre outros, apontando familiaridade, mas não muita correção ao se expressar no idioma inglês.

3.1.1 ANOVA da Dimensão 1

A análise de variância (ANOVA) é utilizada para avaliar a quantidade de variação explicada por cada fator (BERBER SARDINHA; VEIRANO PINTO, 2019, p.6) e determinar se essa variação é estatisticamente significativa (OWA, 2021). O resultado da ANOVA fornece duas medidas estatísticas importantes, frequentemente usadas em análises multidimensionais: a razão F e o coeficiente de determinação (R^2) (BERBER SARDINHA; VEIRANO PINTO, 2019).

A razão F é usada para comparar a variância entre os grupos com a variância dentro dos grupos (BERBER SARDINHA; VEIRANO PINTO, 2019, p. 6). Um valor de F alto sugere que há uma diferença significativa na variância entre os grupos. No entanto, é o valor de p associado ao teste F que indica se essa diferença é estatisticamente significativa. Se o valor de p for menor que o nível de significância (geralmente 0,05), podemos concluir que as diferenças entre os grupos são estatisticamente significativas.

O coeficiente de determinação (R^2) demonstra a proporção da variação total que é explicada pela variável independente (BERBER SARDINHA; VEIRANO PINTO, 2019, p.6). Em outras palavras, R^2 indica o quão bem a variável independente consegue prever ou explicar a variação na variável dependente[143].

Abaixo, na Tabela 16, temos os resultados dessas medidas estatísticas para a Dimensão 1.

TABELA 16: RESULTADO DA ANOVA DA DIMENSÃO 1

VARIÁVEL INDEPENDENTE	**F**	**p**	**R^2**
L1	0,26	0,7744	0,000259
Modo	0,04	0,8348	0,000022
Nível de proficiência	31,12	<0,0001	0,086408

[143] Variáveis dependentes são as variáveis observadas, medidas; variáveis independentes são as selecionadas “para determinar a relação com ou o efeito na variável dependente” (CANTO GÓMEZ, 2013, p. 39).

Tarefa	1307,84	<0,0001	0,725836
Idade	93,32	<0,0001	0,158895
Anos de estudo de inglês	53,29	<0,0001	0,118883

Fonte: Elaborado pela autora

Como se pode observar pelas medidas estatísticas acima, nesta dimensão tanto a língua materna, com F=0,26, p>0,7744 e R^2= 0,00%, como o modo, com F=0,04, p>0,8348 e R^2 =0,00%, não foram estatisticamente significativos. A variável mais preditória dos resultados foi a tarefa, que previu 72,6% da variação, com R^2 de 0,725836 e p<0,0001 e F=1307,84. As demais variáveis independentes, nível de proficiência, anos de estudo de inglês e idade, ainda que tenham valores para o teste estatístico p abaixo do valor crítico, são menos preditoras, sendo responsáveis respectivamente por 8,6% (R^2=0,086408), 11,9% (R^2=0,118883) e 15,9% (R^2=0,158895) de variação nesta dimensão.

Nas Tabelas 17, 18, 19 e 20, abaixo, podemos ver a medida de dispersão das variáveis independentes significativas (p<0,0001).

TABELA 17: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL NÍVEL DE PROFICIÊNCIA NA DIMENSÃO 1

Nível de proficiência	Média	Desvio Padrão
A1	-11,3671873	16,4174046
A2	-7,0765050	19,6321358
B1	3,3598079	22,4955684
B2	5,2360245	19,9753788
C1	3,6379404	17,1463455
C2	1,2173367	14,2684857
Falante nativo de inglês	-0,2699145	14,6756265

Fonte: Elaborado pela autora

Na Tabela 17 acima temos os subgrupos da variável independente Nível de proficiência e os seus subgrupos (A1, A2, B1, B2, C1, C2, Falante nativo do inglês). A média representa o valor médio dos escores dos textos no fator, enquanto o desvio padrão mede a variação em torno da média, indicando "quanta variação podemos observar nos dados" (BREZINA, 2018, p. 49). Observando os valores do desvio padrão podemos concluir que existe uma dispersão considerável em torno das médias nos subgrupos dessa variável. Com 7 subgrupos e um R^2 de 0,086408, podemos inferir que essa variável não possui uma capacidade significativa de diferenciação entre os subgrupos no Fator 1.

TABELA 18: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL ANOS DE ESTUDO DA LÍNGUA INGLESA NA DIMENSÃO 1

Anos de estudo da língua inglesa	Média	Desvio Padrão

1 (<=3)	-2,2053596	25,5608659
2 (>3 até <=6)	-10,1872216	17,6866464
3 (>6 até <=9)	-10,5088901	15,8607197
4 (>9 até <=12)	-1,5430900	18,1052490
5 (>12)	7,3831444	19,3856734
L1 – inglês	-0,2699145	14,6756265

Fonte: Elaborada pela autora

Na Tabela 18 acima temos a variável Anos de estudo da língua inglesa e os seus subgrupos (1, 2, 3, 4, 5). O desvio padrão nos subgrupos dessa variável está bem alto, o que significa que há muita dispersão em torno das médias dos subgrupos dessa variável. Com 6 subgrupos e o R^2 de 0,118883, podemos depreender que essa variável não tem alto poder de predição entre os subgrupos do Fator 1.

TABELA 19: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL IDADE NA DIMENSÃO 1

Idade	Média	Desvio padrão
1 (<=15)	-17,4857931	4,7963384
2 (>15 até <=20)	-0,0773733	18,9278274
3 (>20 até <=30)	5,9862620	19,7423545
4 (>30 até <=40)	4,8520935	17,1958767
5 (>40)	2,1162485	15,5673166

Fonte: Elaborada pela autora

Na variável Idade, Tabela 19, podemos observar que os desvios padrão dos subgrupos estão altos, com exceção do subgrupo 1, indicando que há muita dispersão em torno das médias dos subgrupos dessa variável. Com 5 subgrupos e o R^2 de 0,158895, podemos deduzir que essa variável não tem alto poder de diferenciação entre os subgrupos do Fator 1.

TABELA 20: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL TAREFA NA DIMENSÃO 1

Tarefa	Média	Desvio padrão
13 Sapo	-16,5258929	3,5991013
14 Chaplin	17,6894126	14,3216148
2 Pessoa famosa	-12,9713633	2,1327001
3 Filme	-11,1529819	4,7860107

Fonte: Elaborada pela autora

Na Tabela 20, variável tarefa, constatamos que há baixa dispersão em torno da média de 3 subgrupos: tarefas 2, 3 e 13, o que não acontece com a tarefa 14, que tem uma alta dispersão. Com 4 subgrupos e um R^2 de 0,725836 podemos inferir que nessa variável há um pequeno poder de diferenciação entre os subgrupos, apesar de ter um alto poder de predição, 72,6%, da variação entre as variáveis dependentes.

O conjunto dessas observações indica que no Fator 1 carregaram textos de alunos não muito proficientes, a maioria de nível A 1 a B1 (54%) na amostra selecionada, que a produção desses aprendizes nesse fator foi majoritariamente

escrita, na qual narraram ações de personagens com escolhas lexicais não muito sofisticadas. Nos textos selecionados, que foram os que mais carregaram neste fator, os alunos eram predominantemente do grupo de idade 3, composto de jovens adultos maiores de 20 até 30 anos, inclusive, e com o espanhol como língua materna. 78% desses aprendizes disseram ter estudado a língua inglesa por mais de nove anos, apesar de não serem muito proficientes. A combinação dessas análises juntamente com a análise das etiquetas semânticas deste fator direcionou a denominação do polo positivo como **Dimensão 1: Cuidado, movimento, idade e interações sociais**.

3.2 FATOR 2

Como de costume na AF, o segundo fator extrai o máximo de variação compartilhada das variáveis que sobraram após a extração do primeiro fator (BIBER, 1988). A segunda dimensão possui apenas o polo positivo, apresentando 16 variáveis. A partir dos resultados da AF, chegamos à Tabela 21, que foi organizada em ordem decrescente dos pesos que carregaram no Fator 2.

TABELA 21: PADRÃO FATORIAL DO FATOR 2

	Variável - NGCS	**Peso**
1	Pulling (M2): Grammatical bin (Z5): Age (T3)	0,55389
2	Grammatical bin (Z5): Pulling (M2): Grammatical bin (Z5)	0,47371
3	Grammatical bin (Z5): Crime (G2): Army (G3)	0,46285
4	Coming (M1): Direction (M6): Grammatical bin (Z5)	0,46063
5	Grammatical bin (Z5): Coming (M1): Direction (M6)	0,41071
6	Crime (G2): Army (G3): Coming (M1)	0,40754
7	Age (T3): Direction (M6): Pronouns (Z8)	0,39064
8	Grammatical bin (Z5): Concern (E6): Direction (M6)	0,37115
9	(Grammatical bin (Z5): Age (T3): Direction (M6)	0,36171)
10	Direction (M6): Pronouns (Z8): Closed (A10)	0,36035
11	Grammatical bin (Z5): Pulling (M2): Pronouns (Z8)	0,34377
12	Army (G3): Coming (M1): Direction (M6)	0,34083
13	Coming (M1): Linear (N4): Pronouns (Z8)	0,33302

14	Grammatical bin (Z5): People (S2): Coming (M1)	0,32753
15	Linear (N4): Coming (M1): Direction (M6)	0,30916
16	(Grammatical bin (Z5): Transport (M3): Grammatical bin (Z5)	0,30390)

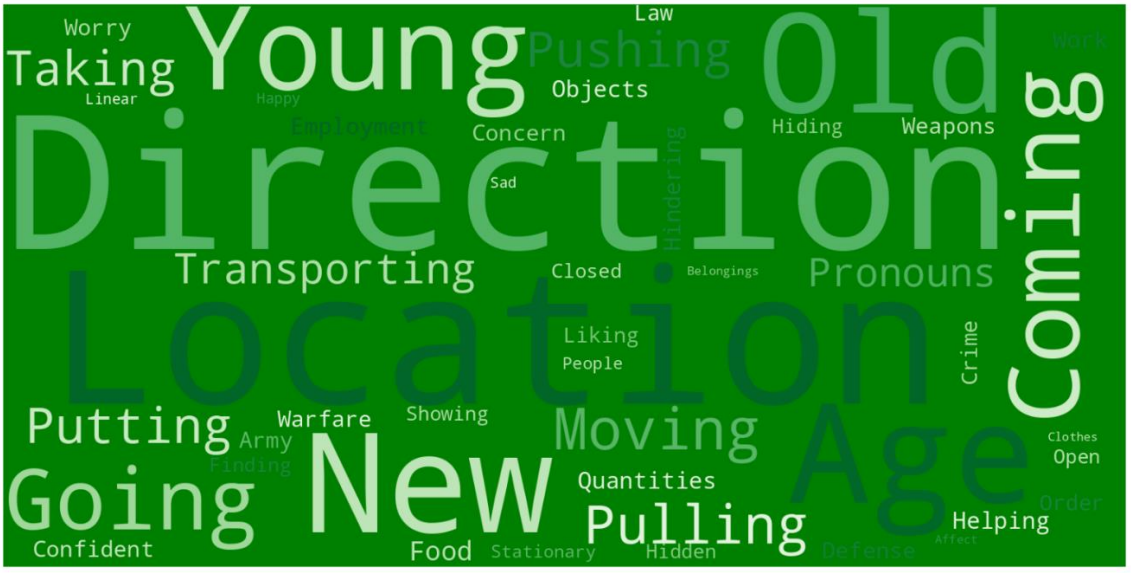
Fonte: Elaborada pela autora

Podemos observar na tabela acima que Z5 é a classe semântica mais saliente, com 25%, seguida de M6, que agrupa palavras com o sentido de localização e direção[144] (ARCHER; WILSON; RAYSON, 2002, p. 18), com 14,5%, e M1, que reúne palavras com o sentido de mover, vir, e ir[145] (idem, p. 17), com 12, 5%.

Os NGCSs 9 e 16 estão em parênteses porque, além de terem carregado no Fator 2, também carregaram no Fator 1, só que com maior peso. O carregamento dessas variáveis no Fator 2 chama-se carregamento secundário. O carregamento com maior peso das variáveis 9 e 16 no Fator 1 denomina-se carregamento primário.

A ocorrência das etiquetas semânticas que carregaram no Fator 2 pode ser visualmente observada na nuvem de palavras abaixo (Figura 9).

FIGURA 9: NUVEM DE PALAVRAS COM AS ETIQUETAS SEMÂNTICAS QUE CARREGARAM NO FATOR 2



Fonte: Elaborado pela autora

O objetivo dessa nuvem de palavras foi deixar em evidência as classes semânticas que carregaram neste fator, realçando o que há de mais importante. Como explicado anteriormente, categorias semânticas visam classificar palavras que possuem significado semântico substancial, em inglês *content words* (palavras com

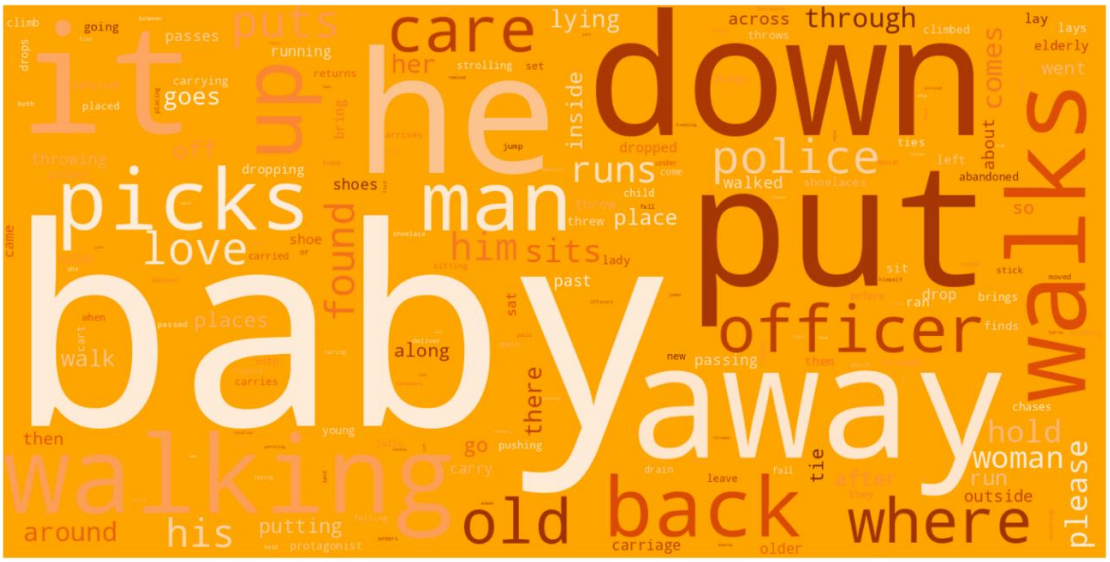
[144] *Location and direction*
[145] *Moving, coming and going*

conteúdo), tais como substantivos, verbos, adjetivos e advérbios. Z5, caixa gramatical, que foi uma etiqueta pensada para palavras gramaticais, normalmente consideradas desprovidas de sentido, não foi incluída ao se gerar essa nuvem de palavras.

Podemos observar as etiquetas salientes: direção, localização, jovem, velho, novo, e verbos de movimento. É a nuvem com maior densidade de etiquetas, sugerindo maior variedade lexical.

Para auxiliar na interpretação da dimensão também foi gerada outra nuvem, desta vez com a ocorrência de palavras das etiquetas semânticas que carregaram no Fator 2. Como no Fator 1, aqui aparecem as palavras classificadas como caixa gramatical. Há menor concentração de itens na Figura 9 acima do que na Figura 10, abaixo, uma vez que há menos categorias do que palavras.

FIGURA 10: NUVEM DE PALAVRAS COM AS PALAVRAS PRESENTES NAS ETIQUETAS SEMÂNTICAS DO FATOR 2



Fonte: Elaborado pela autora

A palavra que sobressai é *baby,* seguida por *down* e *away*[146], e verbos, como *put, walks, walking, picks, love*[147]*.* Aqui aparecem com mais robustez os pronomes *he* e *it*[148]*, e police officer*[149]*,* mais um personagem na narrativa. Da mesma forma que a nuvem de etiquetas, essa também é mais densa do que as nuvens de palavras dos outros fatores, sinalizando maior variedade lexical.

Foram analisadas as sequências de palavras que mais carregaram nos NGCSs no Fator 2. O primeiro exemplo é do NGCS M2_Z5_T3. A etiqueta M2 significa mover, ir e vir (*moving, coming and going*). Agrupa termos que descrevem movimento em

[146] para baixo, para longe
[147] colocar, andar, andando, pegar, amar
[148] ele, ele/ ela neutro
[149] agente da polícia

direção a/ longe de (*towards and away from X)* (ARCHER; WILSON; RAYSON, 2002, p.17). Na Tabela 22 podemos observar a frequência assim como as sequências de palavras que compartilham essas mesmas categorias semânticas,

TABELA 22: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 2

Contagem	NGCS 1	Sequências de palavras
378	M2_Z5_T3	put_the_baby
305	M2_Z5_T3	puts_the_baby
110	M2_Z5_T3	hold_the_baby
68	M2_Z5_T3	places_the_baby
64	M2_Z5_T3	putting_the_baby
36	M2_Z5_T3	dropped_the_baby
34	M2_Z5_T3	throwing_the_baby
27	M2_Z5_T3	place_the_baby
25	M2_Z5_T3	holds_the_baby
24	M2_Z5_T3	drop_the_baby

Fonte: Elaborado pela autora

Abaixo mais um exemplo de NGCS, o segundo que mais carregou, e as sequências de palavras dessa variável: Z5_M2_Z5. Na Tabela 23 podemos observar as frequências e as sequências de palavras.

TABELA 23: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 2

Contagem	NGCS 2	Sequências de palavras
196	Z5_M2_Z5	to_put_the
94	Z5_M2_Z5	to_hold_the
92	Z5_M2_Z5	and_puts_the
34	Z5_M2_Z5	and_put_the
25	Z5_M2_Z5	and_places_the
21	Z5_M2_Z5	about_putting_the
18	Z5_M2_Z5	the_carriage_and
17	Z5_M2_Z5	to_throw_the
17	Z5_M2_Z5	to_place_the
15		about_throwing_the

	Z5_M2_Z5	

Fonte: Elaborado pela autora

Outra variável, a quarta que mais carregou, é M1_M6_Z5. M6 representa localização e direção (*location and direction*[150]*).* Agrupa termos descrevendo a posição de/ o ponto de referência para X[151] (ARCHER; WILSON; RAYSON, 2002, p.18). Abaixo na Tabela 24 os exemplos de sequências de palavras desse NGCS assim como a sua frequência.

TABELA 24: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3

Contagem	NGCS 4	Sequências de palavras
105	M1_M6_Z5	lying_on_the
62	M1_M6_Z5	walking_in_the
58	M1_M6_Z5	walking_on_the
53	M1_M6_Z5	walks_away_with
40	M1_M6_Z5	walking_through_the
39	M1_M6_Z5	runs_away_and
38	M1_M6_Z5	walks_by_and
31	M1_M6_Z5	walks_by_the
25	M1_M6_Z5	walks_away_and
25	M1_M6_Z5	walking_around_the

Fonte: Elaborado pela autora

Outro exemplo de variável é Z5_M1_M6. Abaixo a Tabela 25 apresenta exemplos de sequências de palavras desse NGCS assim como quantas vezes ocorreu.

TABELA 25: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3

Contagem	NGCS 5	Sequências de palavras
130	Z5_M1_M6	and_walks_away
104	Z5_M1_M6	and_runs_away
37	Z5_M1_M6	and_run_away

[150] localização e direção
[151] *Terms depicting means of position of/ point of reference for X*

34	Z5_M1_M6	and_goes_away
25	Z5_M1_M6	to_run_away
23	Z5_M1_M6	and_walks_off
15	Z5_M1_M6	to_walk_away
14	Z5_M1_M6	'_walks_away
14	Z5_M1_M6	and_ran_away
12	Z5_M1_M6	and_runs_off

Fonte: Elaborado pela autora

Abaixo, na Tabela 26, mais uma variável T3_M6_Z8, e os exemplos de sequências de palavras desse NGCS bem como a sua frequência.

TABELA 26: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3

Contagem	NGCS 7	Sequências de palavras
144	T3_M6_Z8	baby_in_his
103	T3_M6_Z8	baby_in_her
88	T3_M6_Z8	baby_in_it
65	T3_M6_Z8	baby_where_he
29	T3_M6_Z8	baby_on_his
13	T3_M6_Z8	baby_inside_it
9	T3_M6_Z8	baby_on_it
7	T3_M6_Z8	baby_where_it
7	T3_M6_Z8	baby_for_him
6	T3_M6_Z8	baby_on_her

Fonte: Elaborado pela autora

No fator 2, *baby* é uma das palavras salientes. As outras palavras que se destacam são verbos que descrevem ações, movimentos distintos, como podemos atestar pelas várias sequências de palavras acima, por exemplo, *puts the baby, lying on the, walks away*[152]*.*

Para a análise qualitativa desta dimensão foram selecionados os textos que mais carregaram neste fator, sendo que 43 deles foram escritos. Todas as produções dos alunos eram da tarefa 14, a narração das cenas de uma parte do filme O garoto,

[152] coloca o bebê, deitando no, vai embora

de Charles Chaplin. A distribuição por proficiência, modo (escrito ou falado), a língua materna, a idade e o total de anos aprendendo inglês podem ser verificados nas Tabelas 27 e 28 abaixo.

TABELA 27: DISTRIBUIÇÃO POR PROFICIÊNCIA, MODO E LÍNGUA MATERNA DOS 50 TEXTOS QUE MAIS CARREGARAM NO FATOR 2

FATOR 2				
Nível de Proficiência	Produção Escrita	Produção Oral	L1: Espanhol	L1: Alemão
A1				
A2				
B1	2		2	
B2	6		4	2
C1	11	3	2	12
C2	11	1		12
L1: Inglês	13	3		

Fonte: Elaborado pela autora

Nesta amostra de textos que mais carregaram neste fator,16 deles (32%) são de falantes com o inglês como L1, sendo que 3 de produção oral e 13 de produção escrita. Outro fato que chamou a atenção foi o nível de proficiência dos aprendizes, 52% eram C1 e C2. Não há produções de alunos A1 e A2. 12% são de alunos B2, e apenas 4% de alunos B1. 48% dos aprendizes tinham o alemão como L1 e 16% o espanhol.

TABELA 28: DISTRIBUIÇÃO POR IDADE E ANOS DE ESTUDO DA LÍNGUA INGLESA DOS 50 TEXTOS QUE MAIS CARREGARAM NO FATOR 2

FATOR			
Grupos por idade	Total de aprendizes	Anos de estudo da língua inglesa	Total de aprendizes
1 (<=15)		1 (<=3)	
2 (>15 até <=20)	19	2 (>3 até <=6)	
3 (>20 até <=30)	27	3 (>6 até <=9)	3
4 (>30 até <=40)	2	4 (>9 até <=12)	12
5 (>40)	2	5 (>12)	19
		L1 - inglês	16

Fonte: Elaborado pela autora

Na Tabela 28, acima, podemos observar que os aprendizes são na sua maioria jovens adultos, 54% com 20 até 30 anos, inclusive. 38% têm de 15 até 20 anos. Como

esperado pelo seu nível de proficiência, a maioria dos alunos, 62%, diz ter estudado inglês por muitos anos, sendo que 24% estudaram de 9 até 12 anos, e 38% mais do que 12.

Abaixo podemos observar três textos. Não foi feita nenhuma adaptação da grafia das composições nos exemplos.

O primeiro exemplo foi o texto com o escore mais alto neste fator: 60,43869014. Pelos metadados do arquivo, de_wr_b2_18_9_14_tb, sabemos que a sua L1 é o alemão (**de**_wr_b2_18_9_14_tb), que a tarefa foi escrita (de_**wr**_b2_18_9_14_tb), que seu nível de proficiência é B2 (de_wr_**b2**_18_9_14_tb), que o aluno tem 18 anos (de_wr_b2_**18**_9_14_tb), estuda inglês há 9 (de_wr_b2_18_**9**_14_tb), que fez a tarefa 14 (de_wr_b2_18_9_**14**_tb), e que suas iniciais são tb (de_wr_b2_18_9_14_**tb**),

Charles walks down a street trying to dodge trash that people throw out of their windows. After that he finds a lonely baby in a corner while he just started smoking a cigarette. He decides to pick up the baby and discovers a mother with a baby carriage and a child in it. He thought the baby belongs to her so he puts him in the stroller. The mother gets upset telling him to take his baby back. So he took the baby back and placed it right where he found him but this time a police officer walks by. He takes the baby again back **and walks around**[153] **with**[154] him until he finds another person. He gives the **baby in his**[155] arms **and runs away**[156]**.** The guy with the baby finds the mother with the stroller as well and so he **puts the baby**[157] **in it**[158] **and walks away**[159]. Charles now randomly **walks by the**[160] baby carriage as the mother sees the random **baby in it**[161] again. She hits Charles a few times and force him to take the baby out by speaking to the police officer. So he picks the baby back up and sits at a roadside. There he finds a paper which was hidden in the clothing of the baby. It leads to him being happy and deciding to keep the baby.

O segundo exemplo foi o texto com o segundo escore mais alto: 60,373200415, Com os seus metadados, de_wr_c1_23_20_14_cb, sabemos que a sua L1 é o alemão (**de**_wr_c1_23_20_14_cb), que a tarefa foi escrita (de_**wr**_c1_23_20_14_cb), que seu nível de proficiência é C1 (de_wr_**c1**_23_20_14_cb), que o aluno tem 23 anos (de_wr_c1_**23**_20_14_cb),

[153] Z5_M1_M6
[154] M1_M6_Z5
[155] T3_M6_Z8
[156] Z5_M1_M6
[157] M2_Z5_T3
[158] T3_M6_Z8
[159] Z5_M1_M6
[160] M1_M6_Z5
[161] T3_M6_Z8

estuda inglês há 20 (de_wr_c1_23_**20**_14_cb), que fez a tarefa 14 (de_wr_c1_23_20_**14**_cb), e que suas iniciais são cb (de_wr_c1_23_20_14_**cb**),

At first, you can see Charles Chaplin walking down an alley while somebody is throwing (probably bricks or similar) out of a window. Charles has to move out of the way to not get hit. He continues walking while he's havin a smoke. He then gets hit by somebody else throwing out bricks. He dusts himself off and lights another cigarette. Afterwards he finds a baby laying on the floor which he wants to give to a woman who is **strolling by with**[162] a trolley but it appears that it's not her kid. Charles takes the kid and wants to put it back where he found it. But a police officer **walks by and**[163] gives him a warning look. Charles then picks the baby back up and continues walking. He passes an older gentleman, **puts the baby**[164] **in his**[165] arms **and runs away**[166] **to**[167] hide. The old man sees the trolley **and puts the**[168] **baby**[169] into it without the woman noticing **and walks off**[170]. Next thing you see is Charles coming out of his hiding place. He **walks past the**[171] woman who sees the baby back in her trolley and follows Charles and calls the police officer. Charles then takes the baby and continues his journey. He sits down and seems to think. He then finds a note in the baby 's pocket which reads: Please love and care for this orphan child. He smiles at the baby **and walks off**[172] **the**[173] scene with the kid in his arms.

O terceiro exemplo foi o texto com quarto escore mais alto: 57,085231977, Pelos seus metadados, de_wr_c1_21_11,5_14_rd, sabemos a sua L1 é o alemão (**de**_wr_c1_21_11,5_14_rd), que a tarefa foi escrita (de_**wr**_c1_21_11,5_14_rd), que o seu nível de proficiência é C1 (de_wr_**c1**_21_11,5_14_rd), que o aluno tem 21 anos (de_wr_c1_**21**_11,5_14_rd), estuda inglês há 11 anos e meio (de_wr_c1_21_**11,5**_14_rd), que fez a tarefa 14 (de_wr_c1_21_11,5_**14**_rd), que suas iniciais são rd (de_wr_c1_21_11,5_14_**rd**),

Charles Chaplin **walks around an**[174] alley and discovers a baby **lying on the**[175]

[162] M1_M6_Z5
[163] M1_M6_Z5
[164] M2_Z5_T3
[165] T3_M6_Z8
[166] Z5_M1_M6
[167] Z5_M1_M6
[168] Z5_M2_Z5
[169] M2_Z5_T3
[170] Z5_M1_M6
[171] M1_M6_Z5
[172] Z5_M1_M6
[173] M1_M6_Z5
[174] M1_M6_Z5
[175] M1_M6_Z5

ground next to a garbage load. He picks up the baby and looks for the mother. After he sees a mother with a baby carriage, he gives her the child and says that she forgot something. The mother gets mad, because it is not her child and Charles takes it back. Than, he wants to **lay the baby**[176] **back**[177] to the place where he found it, but there is a police man behind him, who watches him, so he picks the baby up again. After that, Charles goes to a street with the **baby in his**[178] arms and sits on the curb. He discovers a drain and thinks **about putting the**[179] **baby**[180] in that drain. Suddenly, got pricked by something under the baby's clothes and notices a piece of paper saying "Please love and care for this orphan child ". He smiles at the baby, stands up **and walks away**[181] **with**[182] the **baby in his**[183] arms,

Nesses exemplos, podemos constatar o uso adequado de conectivos, vocabulário preciso e variado, adequada concordância verbal, e quase nenhum erro ortográfico em textos fluentemente escritos.

3.2.1 ANOVA da Dimensão 2

Na Tabela 29, abaixo, temos os resultados da razão F, o valor de p e o coeficiente de determinação (R^2) para a Dimensão 2.

TABELA 29: RESULTADO DA ANOVA DA DIMENSÃO 2

VARIÁVEL INDEPENDENTE	**F**	**p**	**R^2**
L1	192,88	<0,0001	0,163196
Modo	3,71	<0,0543	0,001871
Nível de proficiência	70,37	<0,0001	0,176205
Tarefa	517,74	<0,0001	0,511734
Idade	54,97	<0,0001	0,100133
Anos de estudo de inglês	39,15	<0.0001	0,090166

[176] M2_Z5_T3
[177] Z5_T3_S8
[178] T3_M6_Z8
[179] Z5_M2_Z5
[180] M2_Z5_T3
[181] Z5_M1_M6
[182] M1_M6_Z5
[183] T3_M6_Z8

Fonte: Elaborado pela autora

Como podemos observar, o modo (escrito ou falado), apresenta um F baixo (F=3,71), um p que não é estatisticamente significante (p=0,0543) e R^2=0,0018, o que nos indica que essa variável não é importante para esta dimensão. O valor mais saliente foi para tarefa, com F=517,74 e p <0,0001, com R^2 de 0,511734 apontando que 51,8% da variação nesta dimensão é explicada por essa variável. As outras variáveis independentes Anos de estudo de inglês, Idade, L1 e Nível de proficiência, têm valor estatístico significante com p>0,0001, mas são responsáveis respectivamente por apenas 9% (R^2=0,090166), 10% (R^2=0,100133), 16,3% (R^2=0,163196) e 17,6% (R^2=0,176205) de variação na Dimensão 2.

Nas Tabelas 30, 31, 32, 33 e 34, abaixo, podemos ver a medida de dispersão das variáveis independentes significativas (p< 0,0001).

TABELA 30: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL ANOS DE ESTUDO DE INGLÊS NA DIMENSÃO 2

Anos de estudo de inglês	Média	Desvio Padrão
1 (<=3)	-5,21595834	8,8982685
2 (>3 até <=6)	-7,40864152	7,0650905
3 (>6 até <=9)	-6,54727615	9,8936154
4 (>9 até <=12)	-0,38544638	14,2434064
5 (>12)	2,68144097	14,3435649
L1 – inglês	7,89651585	19,1609407

Fonte: Elaborado pela autora

Na Tabela 30, observamos que o desvio padrão nos subgrupos da variável Anos de estudo de inglês está alto, o que sinaliza que há alta dispersão em torno da média. Com 6 subgrupos e R^2 de 0,090166, podemos deduzir que essa variável não possui capacidade significativa de diferenciação entre os subgrupos no Fator 2.

TABELA 31: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL IDADE NA DIMENSÃO 2

Idade	Média	Desvio padrão
1 (<=15)	-10,0769145	2,2190393
2 (>15 até <=20)	-0,4501586	14,2804473
3 (>20 até <=30)	3,6294553	15,3192553
4 (>30 até <=40)	4,4013717	15,3875302
5 (>40)	2,6519950	14,2912784

Fonte: Elaborado pela autora

Na Tabela 31, observamos que o desvio padrão nos subgrupos da variável Idade também está alto, com exceção do grupo 1, de até 15 anos, indicando uma alta dispersão entre torno da média. Portanto, com 5 subgrupos e R^2 de 0.100133, podemos concluir que essa variável não possui capacidade significativa de diferenciação entre os subgrupos no Fator 2.

TABELA 32: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL L1 NA DIMENSÃO 2

L1	Média	Desvio Padrão
Alemão	8,87825336	18,0322643
Espanhol	7,89651585	19,1609407
Falante nativo de inglês	-3,94650057	10,0200274

Fonte: Elaborado pela autora

Na Tabela 32, observamos que os subgrupos da variável L1 apresentam um desvio padrão alto, o que nos permite inferir que há expressiva dispersão em torno da média. Com 3 subgrupos e R^2 de 0,163196, podemos concluir que essa variável não possui capacidade significativa de diferenciação entre os subgrupos no Fator 2.

TABELA 33: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL NÍVEL DE PROFICIÊNCIA NA DIMENSÃO 2

Nível de Proficiência	Média	Desvio Padrão
A1	-7,78848566	7,1194904
A2	-7,32687655	6,6685972
B1	-3,95137866	9,4111002
B2	0,39669646	13,1205740
C1	4,70959584	15,7149276
C2	9,93174787	18,1774493
Falante nativo de inglês	7,89651585	19,1609407

Fonte: Elaborado pela autora

Na Tabela 33, observamos que o desvio padrão nos subgrupos da variável Nível de proficiência são altos, o que nos permite concluir que há significativa dispersão em torno das médias nos subgrupos. Com 7 subgrupos e R^2 de 0.176205, podemos concluir que essa variável não possui capacidade significativa de diferenciação entre os subgrupos no Fator 2.

TABELA 34: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL TAREFA NA DIMENSÃO 2

Tarefa	Média	Desvio Padrão
13 Sapo	-9,6230284	2,3998138
14 Chaplin	11,1233310	14,5890485
2 Pessoa famosa	-10,5184864	1,2152124
3 Filme	-9,3934849	2,8613738

Fonte: Elaborado pela autora

Na Tabela 34, variável Tarefa, podemos verificar que há baixa dispersão em torno da média de 3 subgrupos: tarefas 2, 3 e 13, o que não acontece com a tarefa 14, que tem uma alta dispersão. Com 4 subgrupos e um R^2 de 0.511734 podemos

inferir que nessa variável há um pequeno poder de diferenciação entre os subgrupos deste fator.

A dimensão 2 foi majoritariamente escrita. No Fator 2 carregaram textos de alunos mais proficientes na língua inglesa e alunos cuja L1 é o inglês. Os aprendizes eram na sua maioria jovens adultos, 54%, com 20 até 30 anos, inclusive. Esses estudantes narraram as ações de personagens com escolhas lexicais variadas, o que podemos observar nos exemplos. Essa variedade lexical também já havia sido sinalizada nas nuvens de palavras deste Fator. A combinação dessas análises direcionou a denominação do polo positivo como **Dimensão 2: Localização, deslocamento, idade, autoridade e emoção**.

3.3 FATOR 3

Como de praxe na AF, o terceiro fator extrai o máximo de variação compartilhada das variáveis que sobraram após a extração do primeiro e do segundo fator (BIBER, 1988). A Dimensão 3 só carregou no polo positivo, com 14 variáveis. A partir dos resultados desse procedimento estatístico, chegamos à Tabela 35 que foi organizada em ordem decrescente dos pesos que carregaram no Fator 3.

TABELA 35: PADRÃO FATORIAL DO FATOR 3

	Variável - NGCS	**Peso**
1	Discourse (Z4): Discourse (Z4): Grammatical bin (Z5)	0,59966
2	Discourse (Z4): Discourse (Z4): Pronouns (Z8)	0,58241
3	Discourse (Z4): Discourse (Z4): Z4	0,52878
4	Discourse (Z4): Grammatical bin (Z5): People (S2)	0,47534
5	Grammatical bin (Z5): People (S2): Discourse (Z4)	0,46981
6	Grammatical bin (Z5): Discourse (Z4): Discourse (Z4)	0,45883
7	Grammatical bin (Z5): Age (T3): Discourse (Z4)	0,43881
8	Discourse (Z4): Direction (M6): Grammatical bin (Z5)	0,41482
9	Personal (Z1): Personal (Z1): Discourse (Z4)	0,40590
10	Direction (M6): Discourse (Z4): Grammatical bin (Z5)	0,37477
11	Coming (M1): Direction (M6): Discourse (Z4)	0,35959

12	Discourse (Z4): Linear (N4): Discourse (Z4)	0,35815
13	Discourse (Z4): Linear (N4): Pronouns (Z8)	0,33405
14	Discourse (Z4): Pronouns (Z8): Closed (A10)	0,31546

Fonte: Elaborado pela autora

Nesta dimensão só houve carregamentos primários. 52% das etiquetas semânticas são Z4, que significa literalmente 'caixa de discurso' (*discourse bin*). Nesse campo semântico foram agrupados marcadores de discurso e termos enfáticos de comunicação (ARCHER; WILSON; RAYSON, 2002, p, 36). 19% são Z5, caixa gramatical.

A ocorrência das etiquetas semânticas que carregaram no Fator 3 pode ser visualmente observada na nuvem de palavras abaixo (Figura 11).

FIGURA 11: NUVEM DE PALAVRAS COM AS ETIQUETAS SEMÂNTICAS QUE CARREGARAM NO FATOR 3

Fonte: Elaborado pela autora

Podemos observar nessa nuvem a proeminência da classe semântica Z4. Z8 (pronomes) também se destaca. Outras classes que sobressaem são relacionadas a verbos que descrevem ações: ir/ vir, esconder, mostrar, mover, abrir (*coming/going, hiding, showing, moving, open),* a tempo: velho, jovem, idade (*old, young, age*) e a pessoas: pessoas, pessoal (*people, personal*),

Para auxiliar na interpretação desta dimensão foi gerada outra nuvem, desta vez com a ocorrência das palavras das etiquetas semânticas que carregaram no Fator 3 (Figura 12).

FIGURA 12: NUVEM DE PALAVRAS COM AS PALAVRAS PRESENTES NAS ETIQUETAS SEMÂNTICAS DO FATOR 3

Fonte: Elaborado pela autora

Nessa figura podemos constatar o que é mais saliente nesta dimensão - uh, hhh, yeah - características da produção oral. Além delas, outras palavras são bastante semelhantes às que carregaram nos outros fatores, sugerindo a narração de ações de personagens - um bebê/ menino/ criança, mulher – numa trama: bebê, menino, criança, mulher, protagonista, entretanto, depois, novamente, encontra, encontrou (*baby, boy, child, woman, protagonist, however, after, again, finds, found*).

Como nas outras dimensões, foram analisadas as sequências de palavras dos NGCS que carregaram no Fator 3. Um deles é Z4_(discurso)_Z4 (discurso)_Z8 (pronomes). Na Tabela 36 podemos observar a frequência assim como as sequências de palavras que compartilham as mesmas categorias semânticas desse NGCS.

TABELA 36: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3

Contagem	NGCS 2	Sequência de palavras
29	Z4_Z4_Z8	hhh_uh_he
18	Z4_Z4_Z8	uh_hhh_he
11	Z4_Z4_Z8	uh_yeah_he
7	Z4_Z4_Z8	hhh_uh_she
6	Z4_Z4_Z8	hhh_uh_they
6	Z4_Z4_Z8	hhh_uh_I
5	Z4_Z4_Z8	hhh_uh_it
4	Z4_Z4_Z8	uh_hhh_she

3	Z4_Z4_Z8	uh_yeah_I
3	Z4_Z4_Z8	uh_uh_she

Fonte: Elaborado pela autora

Abaixo mais um exemplo de NGCS e as sequências de palavras dessa variável: Z5_S2_Z4, sendo que a classe semântica S2 significa pessoa'[184]. As palavras agrupadas no campo semântico de S2 indicam palavras que são relacionadas a ou denotam pessoas[185] (ARCHER; WILSON; RAYSON, 2002, p, 27). Na Tabela 37 podemos observar as sequências de palavras.

TABELA 37: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3

Contagem	NGCS 5	Sequências de palavras
27	Z5_S2_Z4	the_woman_uh
19	Z5_S2_Z4	the_child_uh
9	Z5_S2_Z4	the_woman_hhh
7	Z5_S2_Z4	a_woman_uh
4	Z5_S2_Z4	a_child_uh
3	Z5_S2_Z4	the_protagonist_uh
3	Z5_S2_Z4	the_lady_uh
2	Z5_S2_Z4	the_lady_hhh
2	Z5_S2_Z4	the_girl_uh
2	Z5_S2_Z4	the_child_there

Fonte: Elaborado pela autora

Outra variável é Z4_Z5_S2. Abaixo na Tabela 38 os exemplos de sequências de palavras desse NGCS assim como a sua frequência.

TABELA 38: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3

Contagem	NGCS 4	Sequências de palavras
38	Z4_Z5_S2	uh_the_child
11	Z4_Z5_S2	uh_the_woman
8	Z4_Z5_S2	uh_the_protagonist
6	Z4_Z5_S2	uh_a_woman

[184] *People*

[185] *Terms indicating that particular words relate to/ denote people*

5	Z4_Z5_S2	hhh_the_woman
3	Z4_Z5_S2	yeah_the_woman
3	Z4_Z5_S2	however_the_woman
3	Z4_Z5_S2	hhh_a_woman
2	Z4_Z5_S2	uh_the_women
2	Z4_Z5_S2	uh_the_male

Fonte: Elaborado pela autora

Outro exemplo de variável é M6_Z4_Z5. Abaixo a Tabela 39 apresenta exemplos de sequências de palavras desse NGCS e a sua frequência.

TABELA 39: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3

Contagem	NGCS 10	Sequências de palavras
4	M6_Z4_Z5	away_uh_the
4	M6_Z4_Z5	away_hhh_and
3	M6_Z4_Z5	away_uh_when
2	M6_Z4_Z5	there_uh_and
2	M6_Z4_Z5	there_hhh_so
2	M6_Z4_Z5	there_hhh_and
2	M6_Z4_Z5	in_uh_to
2	M6_Z4_Z5	in_uh_and
2	M6_Z4_Z5	for_uh_the
2	M6_Z4_Z5	away_uh_and

Fonte: Elaborado pela autora

Abaixo, na Tabela 40, mais uma variável M1_M6_Z8 e exemplos de sequências de palavras desse NGCS e a sua frequência. M1, mover, ir e vir (*moving, coming and going*), agrupa termos descrevendo movimento (em direção a e afastado de X[186]) (ARCHER; WILSON; RAYSON, 2002, p. 17).

[186] *towards and away from X*

TABELA 40: EXEMPLOS DE SEQUÊNCIAS DE PALAVRAS DO FATOR 3

Contagem	NGCS 11	Sequências de palavras
7	M1_M6_Z8	runs_away_uh
6	M1_M6_Z8	walks_away_uh
4	M1_M6_Z8	runs_away_hhh
3	M1_M6_Z8	walking_by_uh
2	M1_M6_Z8	walks_past_uh
2	M1_M6_Z8	walks_off_uh
2	M1_M6_Z8	walks_by_uh
2	M1_M6_Z8	running_away_uh
2	M1_M6_Z8	run_away_hhh
1	M1_M6_Z8	went_outside_uh

Fonte: Elaborado pela autora

Em todos os exemplos de sequências de palavras acima aparecem proeminentemente interjeições, *uh, 'hhh'.* Podemos notar nos exemplos que foram normalmente usadas indicando hesitação, que é uma característica da produção oral. Também se observam pronomes, a palavra mulher, senhora, menina, criança (*woman, lady, girl, child*), palavras assinalando localização/ movimento (*there*, *away, off, by, outside*[187]). A saliência de interjeições no Fator 3 indica que é marcadamente oral, diferentemente das outras dimensões.

Para a análise qualitativa desta dimensão foram selecionados textos que mais carregaram no Fator 3. Todos são de produção oral que neste *corpus* era a narrativa individual de tarefas pré-determinadas, sem nenhuma forma de interação entre os aprendizes. Diferentemente dos outros fatores, neste havia outras tarefas (Tabela 43), além de O garoto, de Charles Chaplin. ¨Seis aprendizes falaram sobre um filme de sua escolha, e 10 narraram a estória de um sapo, que tinha ilustrações para guiá-los assim como um glossário (vide anexo). A distribuição por proficiência e L1 podem ser verificadas na Tabela 41 abaixo.

[187] lá/ ali, embora, (ir) embora, (passar) por, exterior

TABELA 41: DISTRIBUIÇÃO POR PROFICIÊNCIA, MODO DE PRODUÇÃO E L1 DOS 50 TEXTOS QUE MAIS CARREGARAM NO FATOR 3

FATOR 3				
Nível de Proficiência	Produção Escrita	Produção Oral	L1: Espanhol	L1: Alemão
A1		2	2	
A2		7	7	
B1		12	12	
B2		10	8	2
C1		12	6	6
C2		7	1	6
L1: Inglês				

Fonte: Elaborado pela autora

Este fator está distribuído entre todos os 6 níveis de proficiência, sendo que 42% dos aprendizes não eram muito proficientes, indo de A1 até B1 e 58% de B2 até C2, como podemos observar na Tabela 41 acima. A maior concentração, 68%, está nos níveis B1, B2, C1, dos quais 24% eram B2, 20% eram B1, e 24%, C1.

Na Tabela 42 abaixo, com a distribuição por idade e anos de estudo da língua inglesa, podemos constatar que a maioria era composta por adolescentes de 15 até 20 anos (50%) e jovens adultos de 20 até 30 anos, inclusive (42%). 54% dos estudantes dizem ter estudado inglês por mais de 12 anos, e 26%, de 9 até 12 anos.

TABELA 42: DISTRIBUIÇÃO POR IDADE E ANOS DE ESTUDO DA LÍNGUA INGLESA DOS TEXTOS QUE MAIS CARREGARAM NO FATOR 3

FATOR			
Grupos por idade	Total de aprendizes	Anos de estudo da língua inglesa	Total de aprendizes
1 (<=15)		1 (<=3)	2
2 (>15 até <=20)	25	2 (>3 até <=6)	4
3 (>20 até <=30)	21	3 (>6 até <=9)	4
4 (>30 até <=40)	1	4 (>9 até <=12)	13
5 (>40)	3	5 (>12)	27
		L1 - inglês	

Fonte: Elaborado pela autora

Na Dimensão 3, apesar da maioria dos textos, 70%, ter sido sobre a tarefa 14, a respeito de uma parte do filme O garoto, de Charles Chaplin, tivemos também 18% sobre a estória em sequência sobre o sapo, e 12% a respeito de um filme. A Tabela 43, adicionalmente, apresenta dados sobre a proficiência nesse fator.

TABELA 43: DISTRIBUIÇÃO POR TAREFA E NÍVEL DE PROFICIÊNCIA DOS TEXTOS QUE MAIS CARREGARAM NO FATOR 3

	Nível de Proficiência					
Tarefa	A1	A2	B1	B2	C1	C2
2. Pessoa famosa						
3. Filme	1			2	3	
13. Sapo			5	2	2	
14 Charles Chaplin	1	7	8	5	6	8

Fonte: Elaborado pela autora

Constatamos que nos textos selecionados foram usadas interjeições em todos os níveis de proficiência, ainda que mais frequentemente nos níveis B1, B2 e C1, e que não houve nem produção oral nem escrita de falantes nativos de inglês.

Abaixo podemos observar três textos. Como nos outros exemplos, não foi feita nenhuma intervenção na produção dos alunos.

O primeiro exemplo foi o segundo texto com o escore mais alto no Fator 3: 95,778573545, Pelos metadados do arquivo, es_sp_b1_20_15_13_mhl, sabemos que a sua L1 é o espanhol (**es**_sp_b1_20_15_13_mhl), que a tarefa foi oral (es_**sp**_b1_20_15_13_mhl), que o seu nível de proficiência é B1 (es_sp_**b1**_20_15_13_mhl), que o aluno tem 20 anos (es_sp_b1_**20**_15_13_mhl), que estuda inglês há **15** (es_sp_b1_20_15_13_mhl), que a sua tarefa foi a 13 (es_sp_b1_20_15_**13**_mhl), e que suas iniciais são mhl (es_sp_b1_20_15_13_**mhl**),

uh is a child that uh have a frog / and the frog escape fo= from his jail hhh and **the child uh** [188] was sleeping uh while the child was sleeping uh the next morning **uh the child**[189] **uh**[190] / hhh uh / uh go to look to look **uh hhh his**[191] fro= uh his frog and then uh he / **uh hhh he**[192] know that the frog uh hhh was escape / was esc= / uh xxx hhh she go to look after / no to look / to look **for uh the**[193] frog uh she go uh he go to the / to the / uh / hhh he **go in uh**[194] **to**[195] look for the hhh the frog and a wild animal throw the / **uh the child**[196] **uh**[197]/ and then the the child go to a lake/ hhh and in the lake was the frog and / the hhh end with his frog and / I think that is a happy

[188] Z5_S2_Z4
[189] Z4_Z5_S2
[190] Z5_S2_Z4
[191] Z4_Z4_Z8
[192] Z4_Z4_Z8
[193] M6_Z4_Z5
[194] M1_M6_Z4
[195] M6_Z4_Z5
[196] Z4_Z5_S2
[197] Z4_Z5_S2

end_

O segundo exemplo do Fator 3 foi o texto com o quinto escore mais alto: 71,346764322, Observando os metadados, es_sp_b1_18_10_14_atm, verificamos que a L1 é espanhol (**es**_sp_b1_18_10_14_atm), que a tarefa foi falada (es_**sp**_b1_18_10_14_atm), que o seu nível de proficiência é B1 (es_sp_**b1**_18_10_14_atm), que o aluno tem 18 anos (es_sp_b1_**18**_10_14_atm), que estuda inglês há 10 (es_sp_b1_18_**10**_14_atm), que a tarefa é a de número 14 (es_sp_b1_18_10_**14**_atm), e que suas iniciais são atm (es_sp_b1_18_10_14_**atm**),

ok uh hhh at the beginning of the video uh we can see Chaplin smoking in the street hhh uh when he suddenly sees a baby on the ground hhh and pick its up / uh and this moment **uh a woman**[198] appear with a baby car hhh and Chaplin uh chase here to return / uh the baby to here / hhh **uh the woman**[199] **uh**[200] gets angry and tells him to take it **away hhh because**[201] that's not her baby hhh uh Chaplin uh decides to returns the baby to the ground/ uh but a policeman appear / uh so he picks up uh the baby and leaves hhh uh then hhh uh Chaplin sees an older man /uh and gives him the baby and **run away hhh**[202] **and**[203] the older man hhh uh puts the baby in the woman 's car baby /again hhh and and leaves hhh **uh the woman**[204] sees Chaplin in the street running hhh and chase him and hits him with an umbrella hhh and explain to the policeman what 's happening/ hhh Chaplin uh takes the baby a sits on the sidewalk hhh and hhh and the end of the video uh Chaplin realise that the baby uh has a letter / on his clothes hhh that say something like hhh uh please love and care for this orphan child hhh and Chaplin sad uh decides to take care of the baby / and **go away hhh**[205] uh that's all I think_

O terceiro exemplo foi o oitavo texto com o escore mais alto: 58,731821729. Com a leitura dos metadados, de_sp_c1_20_12_14_ns, sabemos que a sua L1 é o alemão (**de**_sp_c1_20_12_14_ns), que a tarefa foi falada (de_**sp**_c1_20_12_14_ns), que o seu nível de proficiência é C1 (de_sp_**c1**_20_12_14_ns), que o aluno tem 20 anos (de_sp_c1_**20**_12_14_ns), estuda inglês há 12 (de_sp_c1_20_**12**_14_ns, a tarefa é de

[198] Z5_S2_Z4
[199] Z4_Z5_S2
[200] Z5_S2_Z4
[201]M6_Z4_Z5)
[202] M1_M6_Z4
[203] M6_Z4_Z5
[204] Z4_Z5_S2)
[205] M1_M6_Z4

número 14 (de_sp_c1_20_12_**14**_ns), e suas iniciais são ns (de_sp_c1_20_12_14_**ns**),

hhh in the beginning we see Charlie Chaplin uh walking around in an what appears to be an industrial area uh when he wants to light up a cigarette uh he finds a baby lying on the floor uh / he then uh sees a woman in the background running around with a baby wagon he then assumes that the baby belongs to the uh to the to the woman with the baby wagon and uh just drops it **off uh the**[206] baby actually does not belong to the woman and **uh yeah she**[207] forces uh the baby more or less out of the bab= baby wagon onto Charlie Chaplin once more **uh yeah he**[208] then wants to drop the kid off to on the ground but uh yeah a nearby police officer stops him from doing that and uh yeah Chaplin walks around with the with the kid yeah not knowing what to do with it uh when he sees another man he then pretends that he has something on his shoe and that he wants to clean it and therefore the man has to take care of the child for for a second uh yeah but when he hands uh the the child to the man he just **runs away hhh**[209] uh yeah in the next sequence we see the man uh dropping the the child **off uh into**[210] the baby wagon of the of the women that we have seen in the in the first sequence as well uh to just to get rid of it uh yeah meanwhile uh Chaplin walks by once more on the street uh and **yeah the woman**[211] **uh**[212] assumes that Chaplin uh it was Chaplin who dropped off the the baby once more into the baby wagon and then uh yeah more or less beats him up uh forces the baby onto him once more uh yeah then uh Charlie Chaplin uh **walks around uh**[213] yeah looking a bit cluelessly uh sits down on the ground and uh he then finds finds a note uh that that came with the with the baby and the note says uh please take care and love this orphan child hhh yeah uh the short film ends uh with Charlie Chaplin uh standing up still holding the the baby but looking a bit more happy so it looks like he he now wants to take care of the baby_

3.3.1 ANOVA da Dimensão 3

Na Tabela 44 abaixo, temos os resultados da razão F, o valor de p e o coeficiente de determinação (R^2) para a Dimensão 3.

[206] M6_Z4_Z5
[207] Z4_Z4_Z8
[208] Z4_Z4_Z8
[209] M1_M6_Z4
[210] M6_Z4_Z5
[211] Z4_Z5_S2
[212] Z5_S2_Z4
[213] M1_M6_Z4

TABELA 44: RESULTADO DA ANOVA DA DIMENSÃO 3

Variável Independente	**F**	**p**	**R^2**
L1	12,83	<0,0001	0,012807
Modo	1180,99	<0,0001	0,373733
Nível de proficiência	6,52	<0,0001	0,019435
Tarefa	7,59	<0,0001	0,015141
Idade	9,40	<0,0001	0,018682
Anos de estudo de inglês	4,45	<0,0005	0,011134

Fonte: Elaborado pela autora

Como podemos verificar pelas medidas estatísticas acima, a variável mais saliente é o Modo, com F=1180.99, p <0,0001 e R^2=0.373733 indicando que 37,4% da variação nesta dimensão é explicada por essa variável independente. As demais variáveis, apesar de estatisticamente significantes, possuem um R^2 baixo, e consequentemente, têm um baixo poder de predição na variação da Dimensão 3, como pode ser atestado pelos valores de suas medidas estatísticas. L1, com F=12.83, p <0,0001 e R^2=0.012807, prediz apenas 1,3% de variação; o Nível de proficiência, com F= 6.52, p<0,0001 e R^2=0,019435, prevê apenas 1,9%; a Tarefa, com F=7,59, p<0,0001 e R^2= 0,015141, apenas 1,5%; a Idade, com F=9,40, p <0,0001 e R^2=0,018682, apenas 1,8% e Anos de estudo de inglês, com F=4,45, p=0,0005 e R^2=0,011134, apenas 1,1%.

Nas Tabelas 45, 46, 47, 48 e 49, temos a medida de dispersão das variáveis desta Dimensão.

TABELA 45: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL L1 NA DIMENSÃO 3

L1	Média	Desvio Padrão
Alemão	1,67436565	9,48483358
Espanhol	-0,32915690	8,83973690
Falante nativo de inglês	-1,73211852	3,69224700

Fonte: Elaborado pela autora

Na Tabela 45, observamos que o desvio padrão nos subgrupos da variável L1 são altos, o que nos permite concluir que há significativa dispersão em torno das médias nos subgrupos. Com R^2 de 0.012807, podemos concluir que essa variável não possui capacidade significativa de diferenciação entre os subgrupos no Fator 3.

TABELA 46: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL MODO NA DIMENSÃO 3

Modo	Média	Desvio Padrão
Falado	11,0109062	15,5230728
Escrito	-2,5795020	1,5638349

Fonte: Elaborado pela autora

Na Tabela 46, variável Modo, constatamos que o desvio padrão no subgrupo Falado é alto e que no subgrupo Escrito é baixo. Com um R^2 de 0.373733, podemos inferir que a variável independente Modo tem um poder considerável de predição de variação nas variáveis dependentes do *corpus.*

TABELA 47: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL NÍVEL DE PROFICIÊNCIA NA DIMENSÃO 3

Nível de Proficiência	Média	Desvio Padrão
A1	-1,94241915	4,5540534
A2	-0,92376982	9,3825502
B1	0,39302150	10,7970406
B2	0,17129608	8,0529703
C1	1,38002326	9,6994644
C2	1,62519094	9,4129106
Falante nativo de inglês	-1,73211852	3,6922470

Fonte: Elaborado pela autora

Na Tabela 47, observamos que o desvio padrão nos subgrupos da variável Nível de proficiência são altos, com exceção dos subgrupos Falante nativo de inglês e A1, o que nos permite concluir que há significativa dispersão em torno das médias dos subgrupos. Com 7 subgrupos e um R^2 de 0,019435, podemos concluir que essa variável não possui capacidade significativa de diferenciação entre os subgrupos no Fator 3.

TABELA 48: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL TAREFA NA DIMENSÃO 3

Tarefa	Média	Desvio Padrão
13 Sapo	-1,20764560	6,3545658
14 Chaplin	1,05246395	10,4596197
2 Pessoa famosa	-2,07040771	3,4355623
3 Filme	0,18732720	8,0110609

Fonte: Elaborado pela autora

Na Tabela 48, variável Tarefa, podemos verificar que há que há alta dispersão em torno da média de 3 subgrupos: tarefas 3, 13 e 14, com exceção da tarefa 2. Com 4 subgrupos e um R^2 de 0,015141 podemos inferir que nessa variável não há poder de diferenciação entre os subgrupos do Fator 3.

TABELA 49: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL IDADE NA DIMENSÃO 3

Idade	Média	Desvio padrão
1 (<=15)	-2,66881575	1,6987635
2 (>15 até <=20)	0,93446979	10,7888710
3 (>20 até <=30)	0,10812053	7,8205006
4 (>30 até <=40)	-1,11529688	5,6341192
5 (>40)	0,09672967	9,3598651

Fonte: Elaborado pela autora

Na Tabela 49, observamos que o desvio padrão nos subgrupos da variável Idade também está alto, com exceção do grupo 1, de até 15 anos, o que significa que há uma dispersão considerável em torno das médias nos subgrupos dessa variável. Com 5 subgrupos e R^2 de 0,018682, podemos concluir que essa variável não possui capacidade significativa de diferenciação entre os subgrupos no Fator 3.

TABELA 50: DESVIO PADRÃO DOS SUBGRUPOS DA VARIÁVEL ANOS DE ESTUDO DE INGLÊS NA DIMENSÃO 3

Anos de estudo de inglês	Média	Desvio Padrão
1 (<=3)	0,74525710	16,0141383
2 (>3 até <=6)	-0,19633715	8,1057143
3 (>6 até <=9)	-1,40884975	5,8133340
4 (>9 até <=12)	0,32492298	9,1339462
5 (>12)	0,74033230	9,4654315
L1 – inglês	-1,73211852	3,6922470

Fonte: Elaborado pela autora

Na Tabela 50, observamos que o desvio padrão nos subgrupos da variável Anos de estudo de inglês está alto, com exceção de Falantes nativos de inglês (L1 – inglês), indicando alta dispersão em torno da média. Com 6 subgrupos e R^2 de 0,011134, podemos deduzir que essa variável não possui capacidade significativa de diferenciação entre os subgrupos no Fator 3.

O conjunto dessas observações indica que alunos de todos os níveis de proficiência usaram interjeições na sua produção oral, independentemente da L1. Esse uso parece ter sido fruto de pausas preenchidas, talvez também uma pausa para reorganizar o fluxo de ideias, pensar qual a próxima palavra deveria ser usada para poder narrar o que estava sendo visto. Não houve textos de alunos com o inglês como L1 nos textos que mais carregaram neste fator. A combinação dessas análises direcionou a denominação do polo positivo como **Dimensão 3: Narrativa oral, marcadores de discurso, pausas preenchidas.**

3.4 DISCUSSÃO

Duas questões nortearam esta pesquisa. Uma delas foi a identificação das dimensões de variação de uso de NGCSs na fala e na escrita de alunos de inglês como língua estrangeira. A outra examinou o quanto da variação no uso desses NGCSs foi explicada pelo modo (falado ou escrito), pela língua materna (espanhol, alemão ou inglês), pelo nível de proficiência dos alunos (A1, A2, B1, B2, C1, C2), pela idade, pela quantidade de anos que estudou a língua inglesa ou pela tarefa designada.

Ao analisarmos quantitativa e qualitativamente os resultados dos procedimentos estatísticos empregados identificamos 3 dimensões: 1. **Cuidado, movimento, idade e interações sociais, 2. Localização, deslocamento, idade, autoridade e emoção, 3. Narrativa oral, marcadores de discurso, pausas preenchidas.**

Na Dimensão 1: Cuidado, movimento, idade e interações sociais o poder de predição de variação da tarefa foi de 72,6%. Tanto o modo quanto a língua materna tiveram p acima do valor estatístico crítico, $p>0,7744$ e $p> 0,8348$, respectivamente. As outras variáveis independentes, nível de proficiência, idade e anos de estudo de inglês, apesar de terem p abaixo do valor crítico, $p<0,0001$, têm R^2 indicando pouco poder de predição.

Na Dimensão 2: Localização, deslocamento, idade, autoridade e emoção o poder de predição da tarefa também foi alto, 51,8%; proficiência tem o segundo maior valor com 17,6%, mas é bem menos preditor. O modo tem p acima do valor estatístico crítico, $p<0,543$, R^2 de 0,001871, claramente sem poder de predição nesta dimensão. L1, idade e anos de estudo têm p com valor estatístico significativo, e R^2 de 16,3%, 10% e 9% respectivamente igualmente sem poder de predição.

Na Dimensão 3: Narrativa oral, marcadores de discurso, pausas preenchidas todas as seis variáveis apresentam p abaixo do valor crítico, mas apenas o modo com R^2 de 37,4% tem um poder de predição nesta dimensão.

Nas duas primeiras dimensões constatamos a importância da tarefa designada na produção dos alunos; as duas também são do modo escrito. Igualmente relevante, embora menos saliente estatisticamente, foi o nível de proficiência, que ficou mais evidente quando da análise qualitativa dessas duas dimensões. Na terceira dimensão, marcada pela oralidade, independentemente do nível de proficiência, o modo foi mais determinante para a produção dos aprendizes.

Em relação aos falantes nativos da língua inglesa podemos observar que, dentre os 50 textos com escore mais alto em cada uma das três dimensões, somente na dimensão 2, textos produzidos por nativos tiveram algum destaque. Na dimensão 3 não há textos de falantes nativos; na dimensão 1 apenas 2%

Também observamos que as variáveis independentes Tarefa e Modo possuem um bom poder de predição, mas não entre os seus respectivos subgrupos, o que não chega a ser um problema. Ao contrário, revelou com evidências que a tarefa designada, o modo, escrito ou falado, e o nível de proficiência dos aprendizes, ainda que menos

significativo estatisticamente, apresentam um papel relevante na sua produção, seja ela oral ou escrita.

4. CONSIDERAÇÕES FINAIS

O trabalho aqui apresentado propôs-se a fazer uma investigação exploratória, guiada por *corpus* em um *corpus* de aprendiz. O objetivo foi descrever e analisar a variação do uso de NGCSs neste *corpus*, assim como avaliar o quanto dessa variação pode ser explicada pela tarefa, pelo modo, escrito ou falado, pelo nível de proficiência dos alunos, pela quantidade de anos estudando inglês, pela língua materna, ou pela idade dos alunos.

Primeiramente, o *corpus* foi subdividido em subgrupos organizados de acordo com a língua materna do aprendiz, o modo do texto, e o nível de proficiência. Na sequência foi etiquetado com categorias semânticas. Para tal, utilizou-se o UCREL *Semantic Analysis System*, da Universidade de Lancaster. Após a etiquetagem, foi feita a extração dos NGCSs e identificação dos NGCSs únicos para o posterior cálculo da chavicidade. Para esse cálculo, foi levado em consideração não apenas a frequência, mas também a dispersão dos NGCSs nos textos dos alunos. Em seguida, os NGCCs foram normalizados e classificados. A seguir foi realizada a extração fatorial não rotacionada para determinação do número de fatores. Uma vez escolhida a solução com 3 fatores, efetuou-se a AF rotacionada e depois foram calculados os escores de fator de cada texto. Por último, com a análise qualitativa, os fatores foram interpretados e as dimensões nomeadas.

Para a nomeação das dimensões, foi dada prioridade às etiquetas semânticas. A opção por uma investigação exploratória deveu-se ao fato de que uma tal abordagem permite a identificação de padrões que de outro modo poderiam passar despercebidos. Por ser também uma primeira análise com NGCSs em um *corpus* de aprendiz, escolheu-se usar um n-grama com três categorias semânticas, uma vez que presumimos que quantos maiores os NGCSs, mais escassos seriam, como acontece com pacotes lexicais (CORTES, 2023).

Neste trabalho identificamos três dimensões de variação no *corpus* de estudo: Dimensão 1: Cuidado, movimento, idade e interações sociais, Dimensão 2: Localização, deslocamento, idade, autoridade e emoção, e Dimensão 3: Narrativa oral, marcadores de discurso, pausas preenchidas, sendo as duas primeiras de textos escritos, e a terceira de falados. Observamos que na produção dos alunos tanto a tarefa quanto o modo são fatores relevantes na sua produção. O nível de proficiência também tem papel significativo, mas é menos importante do que a princípio supúnhamos. A língua materna, a idade e os anos estudando inglês não apresentaram resultado expressivo, apresentando pouco poder de predição. O uso de pausas preenchidas na dimensão 3, recursos marcadamente orais, foi usado em todos os níveis de proficiência, evidenciando o marcante papel dessa estratégia para os aprendizes.

Ainda que não seja possível propor generalizações baseadas apenas nesta pesquisa, as evidências em relação ao peso da tarefa assim como ao do modo foram

robustas, demonstrando a sua relevância tanto na produção escrita quanto na produção oral dos alunos, repleta de pausas preenchidas. Tais constatações nos indicam a importância de ter esses fatos em mente quando o educador for planejar suas aulas e atividades.

REFERÊNCIAS

ARCHER, D.; WILSON, A.; RAYSON, P. *Introduction to the USAS category system*. Benedict project report, 2002. Disponível em: UCREL Semantic Analysis System (USAS) (lancs.ac.uk). Acesso em: 9 de set. 2023.

BARON, A.; RAYSON, P.; ARCHER, D. Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies*. Germany, v.20. p. 41-67, 2009.

BERBER SARDINHA, T. Linguística de *Corpus*: histórico e problemática. *DELTA*. São Paulo, v. 16, n. 2, p. 323-367, 2000.

BERBER SARDINHA, T. *Linguística de corpus*. São Paulo: Manole, 2004.

BERBER SARDINHA, T. *Pesquisa em linguística de corpus com Wordsmith Tools*. Campinas: Mercado das Letras, 2009.

BERBER SARDINHA, T. Variação entre registros da internet. In: SALIÉS, T.; SHEPHERD, T.G. (Eds.). *Linguística da internet*. São Paulo: Editora Contexto, 2013.

BERBER SARDINHA, T. Looking at Collocations in Brazilian Portuguese through the Brazilian *corpus*. In: BERBER SARDINHA, T.; FERREIRA, T. *Working with Portuguese* corpora*.* London: Bloomsburry, 2014a. p. 9-32.

BERBER SARDINHA, T. Freedom of Combination and Heterogeneity: a *Corpus* Linguist's Look at Two Saussurean Insights. In: *Matraga*. Rio de Janeiro, v.21, n.34, jan./jun. 2014b.

BERBER SARDINHA, T.; TEIXEIRA, R.; SÃO BENTO FERREIRA, T. Lexical bundles in Brazilian Portuguese. In: BERBER SARDINHA, T.; SÃO BENTO FERREIRA, T. *Working with Portuguese corpora.* London: Bloomsburry, 2014c. p. 33-67.

BERBER SARDINHA, T.; KAUFFMANN, C.; ACUNZO, C. Dimensions of register variation in Brazilian Portuguese. In: BERBER SARDINHA, T.; VEIRANO PINTO, M. (Orgs.) *Multidimensional Analysis 25 Years On:* A Tribute to Douglas Biber. Amsterdam/Philadelphia: John Benjamins, 2014c.

BERBER SARDINHA, T. Register variation and metaphor: a multi-dimensional perspective. In: BERBER SARDINHA, T.; HERRMANN, B. (Eds.). *Metaphor in specialist discourse.* Amsterdam/Philadelphia: John Benjamins, 2015.

BERBER SARDINHA, T. A corpus-based history of Applied Linguistics. In: *World Congress of Applied Linguistics* (AILA) 2017. Comunicação. Rio de Janeiro: 2017.

BERBER SARDINHA, T.; VEIRANO PINTO, M. (Eds.) *Multi-dimensional Analysis:* research methods and current issues. London: Bloomsberry, 2019.

BERBER SARDINHA, T. Discourse of academia from a multidimensional perspective. In: FRIGINAL, E.; HARDY, J. A. (Eds.). The Routledge Handbook of Corpus Approaches to Discourse Analysis. New York: Routledge, 2021. p. 298–318.

BERBER SARDINHA, T. et al. #eunaovoutomarvacina: uma abordagem da Linguística de *Corpus* e da análise multimodal imagética. *Intercâmbio*, v.51. São Paulo: LAEL/PUCSP, 2022. p. 298-318. Disponível em: https://revistas.pucsp.br/index.php/intercambio/article/view/58515. Acesso em: 6 jul. 2024.

BERBER SARDINHA, T. Corpus linguistics and historiography: finding the major discourses in the first 50 years of TESOL Quarterly.In: *Journal of Research Design and Statistics in Linguistics and Communication Science*. *[S. l.]*, v. 7, n. 1, p. 69–90, 2023a. DOI: 10.1558/jrds.18538. Disponível em: https://journal.equinoxpub.com/JRDS/article/view/18538 Acesso em: 13 feb. 2024.

BERBER SARDINHA, T. Análise Multidimensional para escrita acadêmica. In: *Internacionalização da educação superior*: a Linguística de Corpus facilitando o uso de Inglês como Meio de Instrução (EMI) nas universidades brasileiras. Seminário. Porto Alegre, UFRGS, 15 mar. 2023b.

BIBER, D. Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics and philosophy*, v. 23, n. 2, p. 337–360, 1985.

BIBER, D. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, v. 62, n. 2, p. 384–414, 1986.

BIBER, D. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press, 1988.

BIBER, D.; CONRAD, S; REPPEN, R. *Corpus Linguistics*: Investigating Language Structure and Use. Cambridge: CUP, 1998.

BIBER, D. et al. *Longman Grammar of Spoken and Written English.* England: Pearson Education Limited, 1999.

BIBER, D.; CONRAD, S.; CORTES, V. Lexical bundles in speech and writing: an initial taxonomy. In: WILSON, A.; RAYSON, P.; MCENERY, T. *Corpus linguistics by the lune:* a festschrift for Geoffrey Leech. Berlin: Peter Lang, 2003. p. 71-92.

BIBER, D.; CONRAD, S.; CORTES, V. If you look at…: Lexical bundles in university teaching and textbooks*. Applied linguistics*, v. 25, n. 3, p. 371-405, 2004.

BIBER, D. Multi-dimensional patterns of variation among university registers. In: *University language*: a corpus-based study of spoken and written registers. Amsterdam/Philadelphia, PA: John Benjamins, 2006. p. 177-212.

BIBER, D. Lexical bundles in university teaching and textbooks. In: *University language*: a corpus-based study of spoken and written registers. Amsterdam/Philadelphia, PA: John Benjamins, 2006. p. 133-175.

BIBER, D.; CONRAD, S. *Register, genre and style*. Cambridge: CUP, 2009.

BIBER, D.; GRAY, B. Discourse Characteristics of Writing and Speaking Task Types on the TOEFL iBT® Test: A Lexico-Grammatical Analysis. TOEFL iBT ® Research Report TOEFL iBT–19. ETS, 2013. p. 33-37.

BREZINA, V. *Statistics in corpus linguistics: a practical guide.* Cambridge: CUP, 2018.

CANTOS GÓMEZ, P. *Statistical methods in language and linguistic research.* Sheffield: Equinox, 2013.

CANTOS GÓMEZ, P.; Multivariate Statistics Commonly Used in Multi-Dimensional Analysis. In: BERBER SARDINHA, T. e VEIRANO PINTO, M. (Eds.) *Multi-dimensional Analysis:* research methods and current issues. London: Bloomsberry, 2019. p.97-124

Common European Framework of Reference for Languages: learning, teaching, assessment. Disponível em: < http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf >. Acesso em: 4 jul. 2016.

CONRAD, S. e BIBER, D. *Variation in English*: multidimensional studies. London: Routledge, 2001.

CORTES, V. Situating lexical bundles in the formulaic language spectrum: origins and functional analysis developments. In: CORTES, V.; CSOMAY, E. (Eds). *Corpus-based research in Applied Linguistics:* Studies in honor of Doug Biber. Amsterdam/Philadelphia, PA: John Benjamins, 2015. p. 197-216.

CORTES, V. Lexical bundles in EAP. In: JABLONKAI, R. E CSOMAY, E. (Eds.) *The Routledge Handbook of corpora and English language teaching and learning.* London: Routledge, 2023. p. 218-233.

COWIE, A.P. Phraseology. In: ASHER, R.E. (ed.). *The Encyclopedia of Language and Linguistics*. Oxford: Oxford University Press, 1994. p. 3168–3171.

CROSSLEY, S.; LOUWERSE, M. Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics*, v. 12, n. 4, 2007, p. 453-478.

DE MÖNNINK, I. M., BROM, N., e OOSTDIJK, N. H. J. Using the MF/MD method for automatic text classification. In GRANGER, S. e PETCH TYSON, S. (Eds.). *Extending the scope of corpus based research: new applications new challenges*. Amsterdam: Rodopi, 2003.

DELFINO, M.C.N. *More than words:* análise multidimensional da música popular em língua inglesa. Tese (Doutorado em Linguística Aplicada). LAEL, PUC, São Paulo, 2022.

DELFINO, M. C. N.; ARAÚJO, R. F. de; BERBER SARDINHA, T. Revista Brasileira de Linguística Aplicada: multidimensões temáticas. In: M. J. B. FINATTO et al. (Orgs.). *Linguística de corpus*: Perspectivas. Porto Alegre: Instituto de Letras/UFRGS, 2018. p. 93– 126.

DIRDAL, H. et al. Design and construction of the Tracking Written Learner Language (TRAWL) Corpus: A longitudinal and multilingual young learner corpus. In: *Nordic Journal of Language Teaching and Learning*, Vol. 10, No. 2, 2022. p. 115-133. ISSN: 2703-8629 https://doi.org/10.46364/njltl.v10i2.1005

DUNCAN, C.; HOWITT, D. *The Sage Dictionary of Statistics:* a practical resource for students in the social sciences. London: Sage Publications, 2004.

EBELING, S. O.; HASSELGARD, H. Learner Corpora and Phraseology. In: GRANGER, Sylviane; GILQUIN, Gaëtanelle; MEUNIER, Fanny. (Eds.). *The Cambridge Handbook of Learner Corpus Research.* Cambridge: CUP, 2015.

EGBERT, J.; BIBER, D. Incorporating text dispersion into keyword analyses. *Corpora*, v. 14, n. 1, p. 77-104, 2019.

EGBERT, J.; STAPLES, S. Doing Multi-Dimensional Analysis in SPSS, SAS, and R. In: BERBER SARDINHA, T. e VEIRANO PINTO, M. *Multi-dimensional analysis*: Research methods and current issues. London: Bloomsbury, 2019. p. 125-144.

ELLIS, N.C. et al. Learner Corpora and Formulaic Language in Second Language Acquisition Research. In: GRANGER, Sylviane; GILQUIN, Gaëtanelle; MEUNIER, Fanny. (Eds.). *The Cambridge Handbook of Learner Corpus Research.* Cambridge: CUP, 2015.

FITZSIMMONS-DOOLAN, S. Using lexical variables to identify language ideologies in a policy corpus. *Corpora*, 9(1), 2014. p. 57–82. https://doi.org/10.3366/cor.2014.0051

GABRIELATOS, Costas. Keyness analysis. In: TAYLOR, C. e MARCHI, A. (eds). *Corpus approaches to discourse: A critical review.* Oxford: Routledge, *2018.* p. 225-258.

GIL, C. B. Revisitando Sinclair: o princípio idiomático e o princípio da escolha aberta em um corpus de aprendiz. *ReVEL*, v. 21, n. 40, 2023. [www.revel.inf.br]. https://www.revel.inf.br/pt/edicoes/?id=63

GILQUIN, G.; GRANGER, S. Learner language. In: BIBER, D.; REPPEN, R. (Eds.). *The Cambridge Handbook of English Corpus Linguistics.* Cambridge: CUP, 2015. p. 418-435.

GILQUIN, G. Learner *Corpora.* In: PAQUOT, M.; GRIES, S. Th. (eds.). *A practical handbook of Corpus Linguistics.* Switzerland: Springer, 2020.

GRANGER, S. *Learner English on Computer*. London; New York: Addison Wesley Longman, 1998. p. 3-18.

GRANGER, S. A Bird's Eye View of Learner *Corpus* Research. In: GRANGER, S., HUNG, J.; PETCH-TYSON, S. (eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins, 2002. p. 3-33.

GRANGER, S. Learner Corpora in Foreign Language Education. In: VAN DEUSEN-SCHOLL, N. e HORNBERGER N.H. (ed.) *Encyclopedia of Language and Education.* Volume 4. *Second and Foreign Language Education*. Springer, 2008a. p. 337-351.

GRANGER, S.; PAQUOT, M. Disentangling the phraseological web. In GRANGER, S. & MEUNIER, F. (eds.) *Phraseology*: An Interdisciplinary Perspective. Amsterdam & Philadelphia: Benjamins, 2008b. p. 27-49.

GRANGER, S.; BESTGEN, Y. The Use of Collocations by Intermediate vs. Advanced Non-Native Writers: A Bigram-Based Study. In: De Gruyter, *IRAL,* n.52, v.3, 2014. p. 229-25.

GRANGER, S.; GILQUIN, G.; MEUNIER, F. Introduction: learner corpus research – past, present and future. In: GRANGER, S.; GILQUIN, G.; MEUNIER, F. (eds.) *The Cambridge Handbook of Learner Corpus Research.* Cambridge: CUP, 2015. p. 1- 5.

GRAY, B.; BIBER, D. Phraseology. In: BIBER, D.; REPPEN, R. (Eds.) *The Cambridge Handbook of English Corpus Linguistics.* Cambridge: Cambridge University Press, 2015. p. 125-145.

HALLIDAY, M. Corpus studies and probabilistic grammar. ln: AIJMER,K.; ALTENBERG, B. (orgs.). *English corpus linguistics: studies in honour of Jan Svartvik.* London: Longman, 1991. p. 30-43.

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

HUNSTON, S.; FRANCIS, G. *Pattern Grammar:* a Corpus Driven Approach to the Lexical Grammar of English. Amsterdam: Benjamins, 2000.

KAUFMANN, C. H. *Linguística de corpus e estilo*: análises multidimensional e canônica na ficção de Machado de Assis. Tese (Doutorado em Linguística Aplicada). LAEL, PUC, São Paulo, 2020.

KAUFFMANN, C.; BERBER SARDINHA, T. Brazilian Portuguese literary style. In E. Friginal & J. Hardy (Eds.), *The Routledge Handbook of Corpus Approaches to Discourse Analysis* Abingdon: Routledge, 2021. (pp. 354–375). https://doi.org/10.4324/9780429259982-2

KIM, Y.; BIBER, D. A corpus-based analysis of register variation in Korean. In: BIBER, D.; FINEGAN, E. (Eds.). *Sociolinguistic Perspectives on Register*. Oxford: Oxford University Press, 1994.

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

LOEWEN, S., GONULAL, T. Exploratory Factor Analysis and Principal Component Analysis. In: PLONSKY, L. *Advancing Quantitative Methods in Second Language Research.* New York: Routledge, 2015.

LOZANO, C., DÍAZ-NEGRILLO, A., & CALLIES, M. Designing and compiling a learner corpus of written and spoken narratives: COREFL. In C. Bongartz & J. Torregrossa (Eds.), *What's in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy* (pp. 21-46). (2020). Peter Lang. https://doi.org/10.3726/978-3-653-05182-7

MATTE, M. L.; GOULART, L. Pacotes Lexicais e níveis de proficiência em Português como Segunda Língua: Uma investigação da função de pacotes lexicais. *Letras de Hoje*, *[S. l.]*, v. 55, n. 4, p. e38377, 2020. DOI: 10.15448/1984-7726.2020.4.38377.

McCARTHY, M.; O'KEEFE, A. Historical Perspective: What are Corpora and How Have they Evolved? In: O'KEEFE, A.; McCARTHY, M. (Eds.) *The Routledge Handbook of Corpus Linguistics.* Oxford: Routledge, 2010. p. 3-13.

McENERY, T.; HARDIE, A. *Corpus Linguistics*: Method, Theory and Practice. Cambridge: CUP, 2012.

OWA, D.L.M. *Estudo comparativo entre análise multidimensional lexical e modelagem de tópicos.* Tese (Doutorado em Linguística Aplicada). LAEL, PUC, São Paulo, 2021.

PAQUOT, M.; GRANGER, S. Formulaic language in learner corpora. In: *Annual Review of Applied Linguistics*, 32, March 2012. DOI: https://doi.org/10.1017/S0267190512000098

PAQUOT, M. et al. The Varieties for Specific Purposes dAtabase (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing. In: *Research in Corpus Linguistics,* 2022, 10/2, p. 1–15.

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

RAYSON, P. et al. The UCREL semantic analysis system. Conference Paper: Workshop: *Beyond Named Entity Recognition Semantic Labeling for NLP Tasks in LREC'04*. Lisbon, Portugal, Janeiro 2004.

RIBEIRO, N.L.A. *Portal multimodal/multilíngue para o avanço da ciência aberta nas humanidades*: linguagem telecinemática como recurso de ensino de inglês como

língua estrangeira através de n-grama de classe semântica (NGCS). Dissertação (Mestrado em Linguística Aplicada). LAEL, PUC, São Paulo, 2023.

Shin, Y.; Cortes, V.; & Yoo, I. Using lexical bundles as a tool to analyze definite article use in L2 academic writing: An exploratory study. In: *Journal of Second Language Writing*. 39. 10.1016/j.jslw.2017.09.004, 2018. p. 29-41.

SINCLAIR, J. *Corpus, Concordance, Collocation.* Oxford: OUP, 1991.

SINCLAIR, J. Beginning the study of lexis. In: BAZELL, C. et al. (eds.). *In memory of J. R. Firth*. London: Longman, 1966. p. 148–162.

SINCLAIR, J. *Trust the Text*: Language, Corpus and Discourse. London: Routledge, 2004.

STUBBS, M. Collocations and Semantic Profiles: on the Cause of the Trouble with Quantitative Studies. *Functions of language.* Amsterdam: Benjamin, n. 2, v. 2, p. 23-56, dez. 1995.

TABACHNICK, B.G.; FIDELL, L.S. *Using multivariate statistics*. Boston: Pearson Education, 2014.

TOGNINI-BONELLI, E. *Corpus linguistics at work.* Amsterdam & Philadelphia: Benjamins, 2001.

TOGNINI-BONELLI, E. Theoretical Overview of the Evolution of Corpus Linguistics. In: O'KEEFE, A.; McCARTHY, M. (Eds.) *The Routledge Handbook of Corpus Linguistics.* Oxford, 2010.

VEIRANO PINTO, M. *A linguagem dos filmes norte-americanos ao longo dos anos: uma abordagem multidimensional.* Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem). LAEL, PUC, São Paulo, 2013.

WESTIN, I.; GEISLER, C. A multi-dimensional study of diachronic variation in British newspaper editorials. *ICAME*, v. 26, 2002, p. 133-152.

ZUPPARDI, M.C. *Collocation dimensions in academic English.* Tese (Doutorado em Linguística Aplicada). LAEL, PUC, São Paulo, 2020.

ANEXOS

1. Perfil do aprendiz usado para a coleta de metadados para o *corpus* piloto[214]

[214] Learner language profile used for the collection of metadata for the pilot corpus

RH.D.

ugr Universidad de GrimaOa

(LHt saved: 31oct-16]

Fecha:_____

INFORMACION PERSONAL

■ TUS INICIALES___ ■TU NICK___ _ ■EDAD__ ■SEXO: ☐Hombre ☐Mujer

■ CURSO: ☐1QESO ☐2ºESO ☐3ºESO ☐4ºESO ☐1ºBachi. ☐2ºBachi.

☐PCPI ☐Grado administrativo Detro_____ _

■ INSTITUTO DONDE ESTAS ESTUDIANDO: _____

INFORMACION LINGUISTICA

■ Lengua materna: ☐espaiiol ☐otra (indicar): _____

■ Lengua materna de tu padre: Oespaiiol Ootra (indicar): _____

■ Lengua materna de tu madre: Despaiiol Dotra (indicar): _____

■ Lengua(s) que hablas en casa: Despaiiol Dotras (indicar): _____

■ Edad a la que empezaste a aprender ingles _____

■ ,Cual crees tu que es tu nivel de ingles?

SPEAKING:	LISTENING:	READING:	WRITING:
☐Principiante bajo (A1)	☐Principiante bajo (A1)	☐Principiante bajo (A1)	☐Principiante bajo (A1)
☐Principitnte alto (A2)	OPrincipiante alto (A2)	☐Principitrnte alto (A2)	☐Principiante alto (A2)
☐Interm.edio bajo (Bl)	☐Intermedio bajo (Bl)	☐Intermedio bajo (Bl)	☐Intennedio bajo (Bl)
☐Intermedio alto (B2)	☐Intermedio alto (B2)	☐Intermedio alto (B2)	☐Intennedio alto (B2)
☐Avanzado bajo (Cl)	☐Avanzado bajo (Cl)	DAvanzado bajo (Cl)	☐Avanzado bajo (Cl)
☐Avanzado alto (C2)	☐Avanzado alto (C2)	DAvanzado alto (C2)	☐Avanzado alto (C2)

■ ,Estas aprendiendo otro idioma ademas del ingles? Dsi ☐No

Si tu respuesta es st, <Cual? _____

■ Nota del curso pasado (a rellenar por el profesor de ingles): C:::::J

EXPOSICION LINGUiSTICA

■ ,Has hecho alguna estancia en un pais de habla inglesa? Dsi ☐No

Si tu respuesta es sf, <d6nde? _____

lCUando? _____

lCUantas semanas o meses estuviste alli? _____

■ ,Has estudiado o estudias ingles fuera del instituto? Dsi ☐No

Si has contestado st, <en que ano y cuanto tiempo (semanas/meses)?

_____

■ .:Haees algo fuera del colegio relacionado con el ingles? (ej: ver peliculas en ingles, leer internet en ingles, etc.) Osi ☐No

Especifica: _____

_____

■ .:Estas en algun programa de bilinguismo en el Instituto? Dsi ☐No

Si tu respuesta es st, <en que curso empezaste el bilingUismo?____ _

lQue asignaturas bilingues tienes?

cCuantas horas semanales de ingles tienes en esas asignaturas?____ _

■ consentimiento: 0 marca aqui para dar el consentimiento de qve tvs datos sean usados <on fines de investigación sobre el aprendizaje del ingles. Esto NO es un examen. Todos tus datos seran anOnimos y tratados confidencialmente. **Gracias por tu colaboración.**

2. Estória em imagens usada para a elicitação da narrativa na compilação do corpus piloto (baseada em Mayer 1969).[215]

[215] *Picture story used for narrative elicitation in the compilation of the pilot corpus (based on*

Mayer 1969)