CADERNOS
de_LINGUÍSTICA

cadernos.abralin.org

ENSAIO TEÓRICO

ANÁLISE MULTIDIMENSIONAL: OS NÚMEROS NA LINGUÍSTICA

Maria Claudia Nunes DELFINO
Pontifícia Universidade Católica de São Paulo (PUC–SP) | Faculdade de Tecnologia de Praia Grande (FATEC–PG) | Cambly Inc.

RESUMO

A Linguística de Corpus, um dos ramos da Linguística Aplicada, tem como um dos seus construtos metodológicos a análise multidimensional, uma metodologia que leva em consideração a parte quantitativa da linguística, onde grandes quantidades de textos que formam o corpus de análise passam por procedimentos estatísticos. As características linguísticas dos textos são agrupadas em fatores de acordo com sua coocorrência nos textos que, ao serem interpretados linguisticamente, são chamadas de dimensões. Essa abordagem metodológica teve início na década de 80 com o linguista Douglas Biber nos Estados Unidos implementando o que chamamos abordagem americana da Linguística de Corpus. No Brasil, esta abordagem é desenvolvida tanto em textos em língua inglesa, portuguesa, alemã e espanhola e o LAEL na PUC-SP é o polo de desenvolvimento desta metodologia. O presente trabalho é uma revisão de literatura dessa abordagem que a princípio foi desenvolvida para análise gramatical / funcional e, hoje em dia, já aborda trabalhos na área lexical, semântica e colocacional.

ABSTRACT

Corpus Linguistics, a branch of Applied Linguistics, has as one of its methodologic constructs, the multidimensional analysis, a methodology which takes into consideration the quantitative part of Linguistics, where large collection of texts make up the corpus of analysis which go through statistical procedures. The linguistic characteristics of the texts

OPEN ACCESS

EDITORES
– Miguel Oliveira, Jr. (UFAL)
– René Almeida (UFS)

AVALIADORES
– Adriana Lucena (UNP–UNINTA)
– Tiago Torrent (UFJF)

DATAS/ES
– Recebido: 08/08/2021
– Aceito: 02/09/2021
– Publicado: 11/09/2021

COMO CITAR
DELFINO, Maria Claudia Nunes (2021). Análise multidimensional: os números na Linguística. *Cadernos de Linguística*, v. 2, n. 4, e474.

are grouped into factors according to their coocurrence in the texts. When these texts are linguistically interpreted, they are named dimensions. This methodologic approach started in the 80s with the American linguist Douglas Biber in the United States giving start to the American approach of Corpus Linguistics. In Brazil, this approach is developed in texts in English, but also in Portuguese, German and Spanish at LAEL in PUC-SP, which is the main place of development of this methodology. This work is a literary review of this approach that originated as a grammar / functional analysis and, nowadays, there is research in the lexical, semantic and collocational areas.

PALAVRAS-CHAVE

Análise Multidimensional; Linguística de Corpus;
Abordagem Metodológica.

KEYWORDS

Multidimensional Analysis; Corpus Linguistics;
Methodologic Approach.

ANÁLISE MULTIDIMENSIONAL – ORIGEM

Na década de 80, mais especificamente no ano de 1985, o linguista norte-americano Douglas Biber trouxe-nos uma nova maneira de enxergar a língua, por meio da análise de vários textos ao mesmo tempo, valendo-se de uma abordagem estatística, envolvendo múltiplas variáveis, a Análise Multidimensional (AMD). O grande argumento de Biber foi o fato dele entender que a língua não pode ser caracterizada de maneira adequada a partir de uma única dimensão textual, mas deve ser descrita por meio de uma variedade de dimensões que consigam abarcar a descrição de aspectos diferentes dentro de um mesmo texto. O autor considerava inadequado pensar em uma caracterização linguística como dicotômica e que o correto seria considerar a língua como um *continuum*, onde um determinado tipo de texto seria descrito como contendo *mais* ou *menos* de uma determinada característica (por exemplo, mais ou menos formal/informal) (Biber, 1988: p.9).

Biber também propôs uma abordagem *bottom-up*[1] para se estudar a língua, pois as características estudadas seriam selecionadas por padrões de co-ocorrência que se revelariam através de uma análise fatorial exploratória, possibilitando a descrição de grandes quantidades de dados linguísticos em diferentes registros[2]. A prevalência da ocorrência de grupos de características identificadas na análise fatorial foi usada para identificar uma função comunicativa que essas características compartilham quando aparecem juntas e, quando interpretadas, recebem o nome de 'dimensão'. Biber descreve essas dimensões como sendo parâmetros subjacentes de variação linguística (BIBER, 1985, p.338).

O autor, em 1988, trabalhou com um *corpus* de textos que representasse a variedade de registros do inglês na época e, para atingir tal objetivo, os *corpora* escolhidos foram o LOB *corpus* (*Lancaster-Oslo-Bergen*[3] *Corpus*), de textos escritos em inglês britânico e o London-Lud, de transcrições de eventos falados, também do inglês britânico. Além desses dois *corpora,* foram adicionados mais dois registros, cartas pessoais e cartas profissionais, o que totalizou 481 textos e 960 mil palavras. O próximo passo foi selecionar características linguísticas que, segundo a literatura da época, seriam relevantes para a descrição da língua. Para tanto, 67 características de cunho lexical e estrutural foram reunidas e chamadas de variáveis. Com o *corpus* e as variáveis em mãos, o passo seguinte foi

1 Abordagem *bottom up* sugere que o estudo da língua deve surgir a partir do *corpus*, ou seja, o estudo será direcionado pelos resultados da pesquisa feita no *corpus* e não a partir de características definidas a partir da intuição do pesquisador.

2 Registro aqui é definido como uma 'variedade linguística definida por aspectos situacionais, incluindo o propósito do falante, a relação entre falante e ouvinte, e o contexto de produção.' (Biber, 2009: p. 823).

3 Lancaster, Oslo e Bergen são as três universidades que colaboraram para produzir o *corpus*. A universidade de Lancaster localiza-se na Inglaterra e as outras duas na Noruega.

etiquetar[4] esses textos com as variáveis selecionadas através de um etiquetador que o próprio autor desenvolveu especificamente para este estudo, o Biber Tagger. Então, as etiquetas foram contadas com outro programa desenvolvido especificamente para o estudo, o Biber Tag Count, que gerou uma planilha em Excel que foi então introduzida no pacote estatístico[5] para rodar a análise fatorial.

A primeira análise fatorial (chamada de não rotacionada) possibilitou a verificação do número de fatores a serem analisados. No caso do estudo de Biber (1988), 7 fatores foram extraídos, depois o próprio autor considerou os dois últimos fatores não representativos o suficiente e manteve 5 fatores como a solução ideal. Esse número de fatores é definido através do gráfico de escarpa que é o resultado dessa análise não rotacionada em conjunto com a tabela de quantidade de variação compartilhada, conforme mostram as figuras 1 e 2. A partir dessas duas figuras, podemos fazer duas observações importantes: o primeiro fator possui um número altíssimo, maior que o triplo do segundo, ou seja, a maior parte da variação do *corpus* está nele, consequentemente, um número maior de variáveis aloca-se nele.

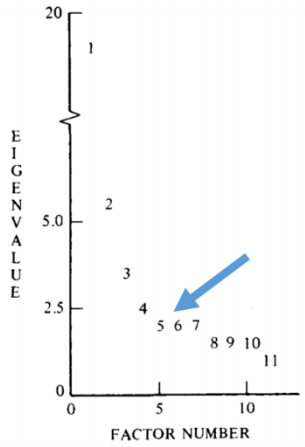


Figura 1. Gráfico de escarpa contendo *eigenvalues*[6] e o número de fatores. **Fonte:** Adaptado de Biber (1988: p.83).

Essa situação é muito comum em análise fatorial; podemos dizer que neste tipo de análise o primeiro fator é sempre o 'mais importante'. A segunda observação é que há uma estabilidade no valor dos *eigenvalues* a partir do fator 5, mas em estudos exploratórios como uma análise fatorial, recomenda-se testar a segunda rotação com

4 Etiquetamento é uma técnica que consiste em inserir em cada palavra do *corpus* uma etiqueta (um código). No caso do Biber Tagger cada etiqueta representa uma característica morfossintática. Exemplo disso é a etiqueta PRO1 que significa pronome de primeira pessoa.

5 Para a realização de uma análise fatorial faz-se necessário a utilização de pacotes estatísticos. A autora sugere SPSS (pago), SAS ou R (ambos gratuitos). O resultado obtido nos três pacotes é muito bom e confiável, a escolha fica a critério da familiaridade com o pacote estatístico e escolha pessoal do pesquisador.

6 *Eigenvalues* ou autovalores possuem o objetivo de explicar o máximo de variabilidade dos dados.

mais de uma quantidade de número de fatores. Neste trabalho, Biber testou com 07 fatores e acabou ficando com 5 em que, onde realmente pode-se ver uma estabilidade com 5 fatores, como pode ser visto nas figuras 1 e 2, que também mostram a alta porcentagem da variação captada pelo fator 1 e uma pequena diferença de variação a partir do fator 5. A primeira análise (não rotacionada) também é necessária para fornecer informações importantes para a análise, tais como a adequação das variáveis para uma análise fatorial, que é feita através do teste KMO[7] e a correlação das variáveis através do teste de esferecidade de Barlett[8].

Factor number	Eigenvalue	% of shared variance
1	17.67	26.8%
2	5.33	8.1%
3	3.45	5.2%
4	2.29	3.5%
5	1.92	2.9%
6	1.84	2.8%
7	1.69	2.6%
8	1.43	2.2%
9	1.32	2.0%
10	1.27	1.9%
11	1.23	1.9%

Figura 2. *Eigenvalues* da análise fatorial não rotacionada. **Fonte:** Adaptado de Biber (1988, p.83).

Com o número de fatores em mãos, uma segunda análise fatorial foi realizada, agora rotacionada. Na segunda análise, as variáveis foram distribuídas nos 07 fatores (no caso da pesquisa de Biber (1988)) de acordo com o peso de cada variável. A figura 3 ilustra os passos de uma análise fatorial completa (não rotacionada e rotacionada), onde podemos notar que essa análise se divide em 03 etapas:

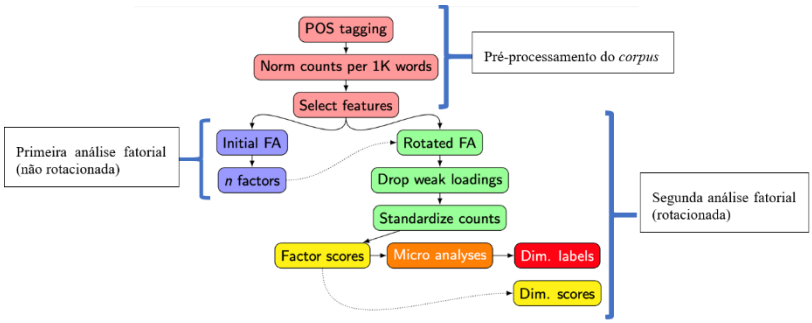


Figura 3. Passos de uma análise fatorial. **Fonte:** Adaptado de Berber Sardinha (2017).

7 KMO – medida de adequacidade da amostra de Kayser-Meyer-Olkin: medida que varia de 0 a 1, onde valores aceitáveis encontram-se entre 0,5 e 1. (KAUFFMAN, 2005)

8 Teste de Barlett – confirma se há correlação entre as variáveis, sendo que o nível de significância (Sig) deve ser inferior a 0,05. (KAUFFMAN, 2005).

1. Pré-processamento do *corpus*:
- o primeiro passo é a etiquetagem dos textos (POS tagging);
- a seguir, esses textos devem ser normalizados por 1000[9], para que textos de tamanhos diferentes consigam ser comparados sem que um texto com muitas palavras sobressaia em relação a um texto com menos palavras;
- a seguir a seleção das variáveis é realizada a partir do valor das comunalidades[10] das variáveis;

2. Primeira análise fatorial (não rotacionada):
- roda-se a primeira análise fatorial e o número de fatores é então determinado (figuras 1 e 2);

3. Segunda análise fatorial (rotacionada):
- roda-se então a segunda análise fatorial, retirando-se variáveis cujo peso é muito baixo;
- calculam-se os escores dos textos e determina-se os escores dos fatores (figura 4);
- a partir de uma análise qualitativa dos textos, juntamente com os resultados dos escores de cada variável em cada fator, as dimensões são nomeadas.

A figura 4 contém a distribuição de algumas variáveis nos 07 fatores extraídos no trabalho de Biber (1988) e é interessante analisar alguns dados. O primeiro deles é que todas as variáveis carregam em todos os fatores. O primeiro passo a ser tomado quando o pesquisador se depara com uma tabela dessas é alocar cada variável em um único fator, o fator onde ela possui maior peso absoluto. Por exemplo, as variáveis *PRO1* e *PRO2* (pronomes de primeira e segunda pessoa, respectivamente) possuem maior peso absoluto no fator 1 (F1)[11]. Elas podem ser consideradas nos outros fatores para a análise qualitativa, mas seu 'lugar principal' é no fator 1. Ao olharmos a variável *N* (substantivo), notamos que ela possui um peso negativo, mas o seu maior valor absoluto também é no fator 1.

9 Textos pequenos, como letras de música, tweets, postagens de redes sociais em geral podem ter sua normalização feita por 100 ao invés de 1000.

10 Comunalidade é a proporção de variabilidade de cada variável que é explicada pelo fator. Quanto maior a comunalidade, maior será o poder de explicação daquela variável pelo fator. Em estudos de AMD considera-se retirar variáveis cuja comunalidade esteja abaixo de 0,15.

11 Fator vem a ser o grupo de variáveis que co-ocorrem significativamente do ponto de vista estatístico. (BERBER SARDINHA, 2004).

LX FEATURE	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6	FACTOR7
PRO1	0.744	0.088	0.025	0.026	-0.089	0.008	-0.098
PRO2	0.860	-0.043	-0.018	0.016	0.007	-0.168	-0.064
PRO3	-0.053	0.727	-0.074	-0.018	-0.167	-0.076	0.138
PANY	0.618	0.046	0.011	0.085	-0.094	-0.085	-0.032
PDEM	0.756	-0.166	-0.001	-0.108	0.004	0.306	-0.077
PERFECTS	0.051	0.480	0.049	-0.016	-0.101	0.146	0.143
PASTTNSE	-0.083	0.895	0.002	-0.249	-0.049	-0.052	0.021
N	-0.799	-0.280	-0.091	-0.045	-0.294	-0.076	-0.213
N_NOM	-0.272	-0.237	0.357	0.179	0.277	0.129	-0.019
N_VBG	-0.252	-0.127	0.216	0.177	0.087	-0.052	0.052
PREP	-0.540	-0.251	0.185	-0.185	0.234	0.145	-0.008
ADVS	0.416	-0.001	-0.458	-0.020	-0.156	0.053	0.314
CONJNCTS	-0.141	-0.160	0.064	0.108	0.481	0.180	0.217
SUB_COS	0.661	-0.080	0.110	0.023	-0.061	0.078	-0.076
SUB_CON	0.006	0.092	0.100	-0.071	0.010	-0.056	0.300
SUB_CND	0.319	-0.076	-0.206	0.466	0.120	0.103	-0.007
SUB_OTHR	-0.109	0.051	-0.018	0.008	0.388	0.102	0.109
INF	-0.071	0.059	0.085	0.760	-0.274	-0.005	-0.074

Figura 4. Variáveis distribuídas nos 07 fatores da análise original de Biber. **Fonte**: Adaptado de Biber (1988, p.86).

Isso ocorre porque em estudos de análise multidimensional levamos em consideração o valor numérico, o sinal positivo ou negativo é levado em consideração para definirmos o polo da dimensão, não possuindo peso de valor ou carga semântica, ou seja, não é dizer que uma variável que está no polo positivo possui conotação positiva, essa marcação positiva e negativa é dada pelo programa estatístico e não pela significado da variável. Após as variáveis serem computadas nos fatores estabelecidos, os textos que fazem parte do *corpus* analisado possuem seus escores computados. Os escores consistem em somas relativas às quantidades das variáveis existentes em cada fator. Cada texto é computado em cada dimensão. A partir desta tabela completa (a figura 4 mostra apenas algumas variáveis), Biber interpretou o conjunto de características funcional e discursivamente, o que levou ao estabelecimento das dimensões.

Cada dimensão, nada mais é do que uma escala onde são dispostos todos os registros incluídos na análise, de acordo com seus escores de dimensão e que pode conter dois polos opostos, de tal modo que as dimensões são geralmente descritas como 'polo A *versus* polo B'. Quanto mais distantes os registros encontram-se na escala, mais diferentes esses registros são. Na terminologia da AMD, os termos 'positivo' e 'negativo' são empregados para se referir a esses polos, sendo que o polo A recebe o nome de 'positivo' e o polo B de 'negativo'. Entretanto, tal denominação não significa dizer que um polo é mais importante do que o outro, pois na verdade eles se complementam. Os termos 'positivo' e 'negativo' apenas refletem a análise fatorial que traz resultados positivos e negativos, significando que em um mesmo texto quando uma variável positiva ocorre, uma negativa tende a não ocorrer ou ocorrer em menor número e vice-versa. A figura 5 traz a estrutura fatorial dos 03 primeiros fatores da pesquisa de Biber (1988), demonstrando como variáveis presentes no polo positivo não estão presentes no negativo e vice-versa. Como exemplo, observemos o fator 1, que possui variáveis com peso positivo (*private verbs, that deletion, contractions,*

present tense verbs, 2nd person pronouns[12], entre outras) e variáveis com peso negativo (*nouns, word length, prepositions, type/token ratio, attributive adjs., place adverbials*[13], entre outras). Isso significa dizer que em textos que contêm grande quantidade de *verbos privados*[14] e *apagamento do that,* há uma tendência de aparecimento também de *contrações* e *verbos no presente.*

Por outro lado, nos textos em que existem *verbos privados, apagamento do that, contrações e verbos no presente* há uma tendência a escassez ou ausência de *substantivos* e *palavras longas*. Também pode-se notar que algumas das variáveis apresentam-se distribuídas nos fatores entre parênteses. O pesquisador se utiliza disso para demonstrar que elas estão presentes no polo apenas para efeito de interpretação qualitativa, mas seu valor absoluto é maior em outro fator. Exemplo disso é a variável *present tense verbs*, que está alocada nos fatores 1 positivo, com peso .86 e no fator 2 negativo, com peso -.47. Como seu valor absoluto é maior no fator 1 positivo, ela faz parte desse fator neste polo, mas pode ser levada em consideração na hora de interpretar a função comunicativa que o polo 2 negativo possui[15].

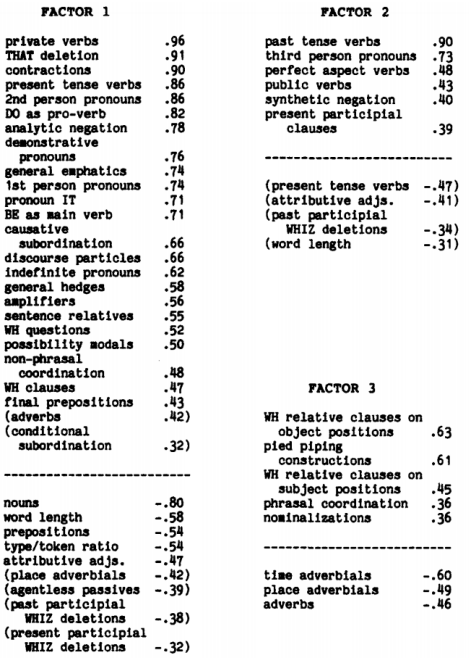
A partir da análise dessas variáveis em cada fator e, voltando aos textos para inspecionar como essas variáveis ocorriam em termos de função compartilhada, foi observado no fator 1 que as variáveis com peso positivo tinham como parâmetro subjacente o que se convencionou chamar de 'produção interativa'. No outro lado do *continuum*, o conjunto de características com peso negativo revelavam um parâmetro que se convencionou chamar 'produção informacional'. Assim, o rótulo para a dimensão 1 da língua inglesa foi 'produção interativa *versus* produção informacional'.

12 *Private verbs, that deletion, contractions, present tense verbs, 2nd person pronouns* – Verbos privados, apagamento do that, contrações, verbos no presente, pronomes de segunda pessoa (respectivamente).

13 *Nouns, word length, prepositions, type/token ratio, attributive adjs., place adverbials* – Substantivos, tamanho da palavra, preposições, relação entre o número total de palavras e o número único de palavras, adjetivos atributivos, advérbio de lugar

14 Verbos privados são verbos que expressam estados intelectuais. Os verbos privados mais comuns são *assume, believe, calculate, conclude, consider, deduce, think (*assumir, acreditar, calcular, concluir, considerar, deduzir pensar).

15 Observação da autora: O polo de um fator que, ao ser interpretado qualitativamente se tornará o polo de uma dimensão, deve preferencialmente conter no mínimo de 3 a 5 variáveis que carreguem prioritariamente nele. Levando-se em consideração o princípio de que uma dimensão mostra a função compartilhada por várias características linguísticas que co-ocorrem em vários textos, não faz sentido um polo de dimensão com pouquíssimas características. Elas podem co-ocorrer nos mesmos textos mas não são suficientes para compartilhar uma função linguística.



FACTOR 1	
private verbs	.96
THAT deletion	.91
contractions	.90
present tense verbs	.86
2nd person pronouns	.86
DO as pro-verb	.82
analytic negation	.78
demonstrative pronouns	.76
general emphatics	.74
1st person pronouns	.74
pronoun IT	.71
BE as main verb	.71
causative subordination	.66
discourse particles	.66
indefinite pronouns	.62
general hedges	.58
amplifiers	.56
sentence relatives	.55
WH questions	.52
possibility modals	.50
non-phrasal coordination	.48
WH clauses	.47
final prepositions	.43
(adverbs	.42)
(conditional subordination	.32)
---	
nouns	-.80
word length	-.58
prepositions	-.54
type/token ratio	-.54
attributive adjs.	-.47
(place adverbials	-.42)
(agentless passives	-.39)
(past participial WHIZ deletions	-.38)
(present participial WHIZ deletions	-.32)

FACTOR 2	
past tense verbs	.90
third person pronouns	.73
perfect aspect verbs	.48
public verbs	.43
synthetic negation	.40
present participial clauses	.39

(present tense verbs	-.47)
(attributive adjs.	-.41)
(past participial WHIZ deletions	-.34)
(word length	-.31)

FACTOR 3	
WH relative clauses on object positions	.63
pied piping constructions	.61
WH relative clauses on subject positions	.45
phrasal coordination	.36
nominalizations	.36
---	
time adverbials	-.60
place adverbials	-.49
adverbs	-.46

Figura 5. Estrutura fatorial dos 03 primeiros fatores do estudo de Biber (1988). **Fonte**: Adaptado de Biber (1988, p.89).

A figura 6 traz como os registros foram distribuídos na dimensão 1. Lembrando que os números que vemos na escala da dimensão 1 representam a média dos escores dos textos que compõem cada registro do *corpus* em estudo. Como se percebe na figura 6 os registros mais fortemente associados com interação (na parte superior da escala), são as conversas, tanto por telefone como face a face; já os registros mais diretamente ligados à informação (na parte inferior) são documentos oficiais (como relatórios) e a escrita acadêmica e jornalista.

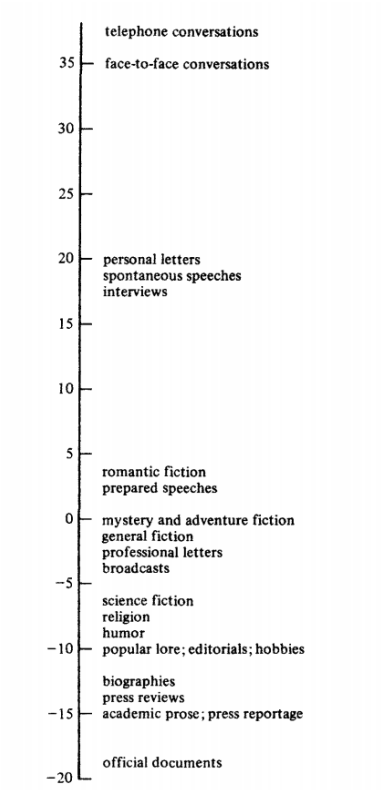


Figura 6. Dimensão 1 de Biber (1988). **Fonte**: Biber (1988, p.128).

Biber valeu-se do mesmo procedimento para a descrição e nomeação das outras dimensões, que receberam os seguintes nomes e possuíram os seguintes registros mais tipicamente presentes em cada uma:[16]

- Dimensão 1 – Produção interativa *versus* produção informacional – polo interativo (positivo): conversas telefônicas e face a face; polo informacional (negativo): documentos oficiais, reportagem jornalística e prosa acadêmica.
- Dimensão 2 – Preocupações narrativas *versus* não narrativas – polo narrativo (positivo): ficção; polo não narrativo (negativo): rádio e TV, passatempos e documentos oficiais.
- Dimensão 3 – Referências explícitas *versus* referências dependente do contexto – polo referências explícitas (positivo): documentos oficiais, cartas profissionais, resenhas jornalísticas e prosa acadêmica; referências dependentes do contexto (negativo): rádio e TV, conversas telefônicas e face a face e ficção romântica.

16 A autora escolheu listar apenas as 05 primeiras dimensões do trabalho seminal de Biber (1988), que foram as que ele determinou posteriormente como as dimensões da língua inglesa.

- Dimensão 4 – expressão explícita de persuasão *versus* não explícita – polo persuasivo (positivo): cartas profissionais, editoriais e ficção romântica; polo não persuasivo (negativo): rádio e TV, resenhas jornalísticas e ficção de aventura.
- Dimensão 5- informação abstrata *versus* não-abstrata – polo abstrato (positivo): acadêmico, documentos oficiais e religiosos; polo não-abstrato (negativo): conversas telefônicas e face a face e ficção romântica.

Este estudo proporcionou o mapeamento da língua inglesa nestas 05 dimensões e, portanto, um registro pode ser caracterizado funcionalmente/gramaticalmente, a partir da sua distribuição nestas 05 dimensões. Peguemos, por exemplo, o registro *conversas face a face.* Ao observarmos este registro nas 05 dimensões da língua inglesa, ele é caracterizado como: interativo (Dim. 1), não narrativo (Dim. 2), dependente do contexto (Dim. 3), não persuasivo (Dim. 4) e não-abstrato (Dim. 5).

1. ANÁLISE MULTIDIMENSIONAL ADITIVA

Nem todos os registros foram estudados por Biber (1988), então trabalhos como o de Delfino (2016), que estudou um *corpus* de letras de música pop, um registro não estudado por Biber em 1988, precisam ter seu escore descoberto para que possa haver a adição desse registro aos registros que Biber trabalhou. Nesta modalidade da análise não é necessário proceder-se à extração fatorial (análise rotacionada), que é típica de uma AMD completa. O *corpus* precisa estar etiquetado com as mesmas etiquetas do estudo base, ou seja, aquele ao qual será 'adicionado', além de ter as frequências das variáveis do *corpus* a ser adicionado padronizadas segundo a média e desvio padrão do estudo base.

Após a etiquetagem com o Biber Tagger e contagem das etiquetas com o programa Biber Tag Count[17] (exatamente como no trabalho de Biber (1988)), tem-se os escores dos textos e, por conseguinte do novo registro a ser adicionado. O *corpus* do referido trabalho citado como exemplo (DELFINO, 2016) foi desmembrado em 04 *subcorpora*[18] (cada banda ou artista foi considerado um subcorpus) e, tanto o *corpus* como um todo como cada uma das bandas e do artista foram alocados nos registros de Biber (1988). A figura 7 mostra como esses registros foram inseridos na Dimensão 1 de Biber (1988). Como pode ser observado o *corpus* CoEL e cada um de seus subcorpora estão posicionados entre cartas

17 Vale salientar que o programa Biber Tagger Count já normaliza os valores absolutos por 1.000

18 O *corpus* CoEL (Corpus of English Lyrics) era formado pelas letras de música de 03 bandas: Beatles, Bon Jovi e Maroon 5 e do cantor Bruno Mars.

pessoais (a banda Bon Jovi é a que mais se aproxima deste registro) e conversas face a face (as bandas Beatles e Maroon 5 são as que mais simulam uma conversa com o público).

A autora mapeou o *corpus* CoEL como um registro adicionado às cinco dimensões da língua inglesa (BIBER, 1988), da seguinte maneira: envolvido (Dim. 1), não narrativo (Dim. 2), dependente do contexto (Dim. 3), persuasivo (Dim. 4) e não abstrato (Dim. 5).

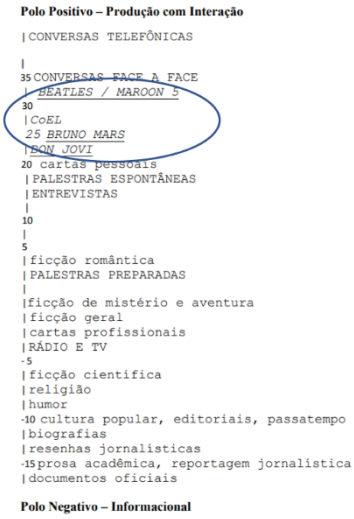


Figura 7. Inserção do corpus CoEL na Dimensão de Biber (1988). **Fonte:** adaptado de Delfino (2016, p.72).

2. ANÁLISE MULTIDIMENSIONAL LEXICAL

Berber Sardinha, em 2014, propôs um novo modelo para a Análise Multidimensional apresentada por Biber em 1988 que se baseia não na interpretação funcional, mas na interpretação lexical dos fatores, com o intuito de encontrar dimensões de variação lexical, identificadas por meio da interpretação dos campos temáticos subjacentes à coocorrência do léxico mais saliente. Para tanto, o autor investigou o uso dos adjetivos *American* e *Brazilian*, bem como seus colocados – palavras que ocorrem perto do nódulo – para identificar os parâmetros de representação de identidade nacional e cultural por meio do qual os EUA e o Brasil são representados nas produções textuais em inglês a partir do século XIX disponibilizados pelo *Google Books*.

Na perspectiva lexical, a AMD empregada no referido estudo considerou como variáveis (unidade de análise) apenas as palavras de conteúdo[19] ou multipalavras para a identificação das dimensões de variação, ou seja, a diferença fundamental entre a análise

19 Palavras de conteúdo são substantivos, verbos, advérbios e adjetivos.

multidimensional funcional/gramatical (proposta por Biber em 1988) e a lexical/temática (proposta por Berber Sardinha em 2014) recai justamente sobre as variáveis selecionadas e o etiquetador utilizado. No estudo de Biber (1988) e naqueles que se seguiram (BIBER, 2006; BIBER e TRACY-VENTURA, 2007; BERBER SARDINHA, KAUFFMANN e ACUNZO 2014, entre outros) o objetivo é a análise da variação gramatical/funcional; para tanto, o etiquetador Biber Tagger é utilizado e as variáveis são palavras de cunho léxico-gramatical (pronomes, substantivos, verbos, preposições, adjetivos, advérbios,...).

Já para estudos lexicais/temáticos, o etiquetador recomendado é o Tree Tagger que etiqueta por lema, ou seja, pela forma da palavra conforme ela está grafada no dicionário. A figura 8 ilustra os aspectos similares e distintos entre a AMD funcional/gramatical e lexical/temática. No estudo de Berber Sardinha (2014), o autor fez os mesmos passos de uma análise multidimensional, mudando apenas as variáveis e o etiquetador (ou seja, a mudança ocorre no pré-processamento do *corpus*, mantendo o objetivo da análise em si (vide figura 3). O autor realizou duas análises multidimensionais, uma onde a palavra *American* era destacada e os colocados mais frequentes dela eram utilizados como variáveis e outra análise onde a palavra *Brazilian* foi destacada e seus colocados mais frequentes foram utilizados como variáveis. As dimensões encontradas por Berber Sardinha (2014) foram:

	Funcional	**Lexical**
Objetivo	Identificar parâmetros subjacentes de variação nos textos de um *corpus*	
Unidade de observação	Textos ou segmentos de texto	Palavras, colocações
Traços linguísticos	Léxico-gramaticais	Lexicais

Figura 8. Aspectos da AMD funcional/gramatical e lexical/temática. **Fonte:** Adaptado de Berber Sardinha (2017).

Para *American:* Dimensão 1 – *Superpower versus regional status;* Dimensão 2 – *The people, the flag and the institutions;* Dimensão 3 – *Individuals, community and culture;* Dimensão 4 – *The military, slavery and ideals;* Dimensão 5 – *Literate expression versus revolution and the new nation.*[20] Para *Brazilian:* Dimensão 1 - *The economy and politics;* Dimensão 2 – *Traditional art, sciences, the people and the land;* Dimensão 3 – *Raw materials and the landscape;* Dimensão 4 – *New artistic forms, women and men, religion and the environment;* Dimensão 5 – *The monarchy, steam transportation, and the wilderness.*[21]

20 Dim. 1 – Super poder *versus* status regional; Dim. 2 – As pessoas, a bandeira e as instituições; Dim. 3 – Indivíduos, comunidade e cultura; Dim. 4 – As forças armadas, escravidão e ideais; Dim. 5 – Expressão literária *versus* revolução e a nação.

21 Dim. 1 – Economia e política; Dim. 2 – Arte tradicional, ciências, as pessoas e a terra; Dim. 3 – Matéria-prima e a paisagem; Dim. 4 – Novas formas artísticas, mulheres e homens, religião e meio ambiente; Dim. 5 – A monarquia, transporte a vapor, e região selvagem.

3. ANÁLISE MULTIDIMENSIONAL COLOCACIONAL

Em 2017, Berber Sardinha revelou a análise multidimensional colocacional sob uma perspectiva de registro, usando o *Corpus of Contemporary American English* (COCA), com 450 milhões de palavras (Davies, 2012). A metodologia também seguiu os passos de Biber em seu estudo seminal de 1988, com o mesmo objetivo de determinar as dimensões ou os parâmetros de variação subjacente presentes no *corpus* estudado. Sabendo que a análise fatorial, a base da análise multidimensional, tem como premissa a correlação de variáveis que 'andam juntas', faz muito sentido uma AMD colocacional, já que a colocação nada mais é do que palavras que 'andam juntas'. O relacionamento entre o nódulo e seus colocados já havia sido apresentado por autores como Williams (1998), Mollet et al. (2011) e Brezina et al. (2015), mas o relacionamento entre os colocados e uma série de nódulos não era levado em consideração até estudos multidimensionais como o de Berber Sardinha (2017) e de Zuppardi (2020).

As diferenças entre a AMD funcional/gramatical e a colocacional são as seguintes:

- as unidades de análise não foram os textos e sim pares de palavras, sendo uma o nódulo e outra o colocado (selecionados entre as palavras mais frequentes em cada registro do COCA);
- as medidas utilizadas não foram as contagens dos textos (na AMD funcional essas contagens são fornecidas pelo programa Biber Count em uma planilha de Excel que é inserida no programa estatístico). Aqui as medidas são o log-dice, a força de atração que duas palavras possuem entre si (RYCHLY, 2008);
- os escores de fator foram calculados para os colocados de cada palavra nódulo no *corpus* ao invés de cada texto (como nos estudos funcionais/gramaticais);
- a base para interpretação dos fatores neste tipo de AMD foram as características lexicais reveladas pela sua preferência semântica, pacotes lexicais, campos de palavras, 'aboutness'[22], tópicos e assuntos e não na base funcional/comunicativa, como na AMD funcional/gramatical.

Berber Sardinha identificou 09 dimensões de colocação da língua inglesa: Dimensão1 – *Literate discourse*, com os seguintes exemplos: *issue* + *relate, fator* + *relate, seem* +

22 O termo 'aboutness' abrange, além do tópico dos textos, a representação construída através das características lexicais. Nesse sentido, pode-se determinar o posicionamento do(s) autor(es) do(s) texto(s) a partir de sua análise lexical.

appropriate[23]*;* Dimensão 2 – *Oral discourse*, com os seguintes exemplos: *want* + *know, people* + *know, want* + *say*[24]; Dimensão 3 - *Objects, people, and actions*, com os seguintes exemplos: *stare* + *window, stare* + *ceiling, slide* + *open*[25]; Dimensão 4 – *Colloquial and informal language use*, com os seguintes exemplos: *afraid* + *lose, mama* + *papa, mama* + *daddy*[26]; Dimensão 5 – *Organizations and the government*, com os seguintes exemplos: *protection* + *agency, oficial* + *say, international* + *monetary*[27]; Dimensão 6 – *Politics and current affairs*, com os seguintes exemplos: *other* + *politician, decline* + *interview, police* + interview[28]; Dimensão 7 – *Feelings and emoticons*, com os seguintes exemplos: *feel* + *shame, feel* + *guilt, feel* + *rage*[29]; Dimensão 8 – *Cooking*, com os seguintes exemplos: *mix* + *bowl, mix* + *ingredient, cup* + *sugar*[30]; Dimensão 9 – *Education*, com os seguintes exemplos: *student* + *benefit, rate* + *scale, expose* + *student*.[31]

4. ANÁLISE MULTIDIMENSIONAL SEMÂNTICA

Em 2021, Delfino e Berber Sardinha trouxeram um quarto modelo de análise multidimensional, em que o objetivo era entender a variação semântica que ocorria em um *corpus* de letras de música pop. A metodologia também seguiu aquela introduzida por Biber em 1988, com o mesmo objetivo de determinar as dimensões ou os parâmetros de variação subjacente presentes no *corpus* estudado. As diferenças entre a AMD funcional/gramatical e a semântica são as seguintes:

- as variáveis são os campos semânticos (anatomia, roupas) e não características de cunho estrutural (pronomes, verbos) como nos estudos funcionais/gramaticais;
- o etiquetador utilizado é o UCREL, o etiquetador de análise semântica desenvolvido pela universidade de Lancaster[32] e que possui como etiquetas principais os seguintes campos semânticos que, em alguns casos, possuem subdivisões: *General & Abstract*

23 Dim. 1 – Discurso letrado, com os seguintes exemplos: assunto + relacionar, fator + relacionar, parecer + apropriado.
24 Dim. 2 – Discurso oralizado, com os seguintes exemplos: querer + saber, pessoas + saber, querer + dizer.
25 Dim. 3 - Objetos, pessoas e ações, com os seguintes exemplos: encarar + janela, encarar + teto, deslizar + abrir.
26 Dim. 4 – Uso coloquial e informal da língua, com os seguintes exemplos: medo + perder, mamãe + papai, mamãe + paizinho.
27 Dim. 5 – Organizações e o governo, com os seguintes exemplos: proteção + agência, oficial + dizer, internacional + monetário.
28 Dim. 6 – Política e assuntos atuais, com os seguintes exemplos: outro + político, recusar + entrevista, polícia + entrevista.
29 Dim. 7 – Sentimentos e emoções, com os seguintes exemplos: sentir + vergonha, sentir + culpa, sentir + fúria.
30 Dim. 8 – Culinária, com os seguintes exemplos: misturar + tigela, misturar + ingrediente, xícara + açúcar.
31 Dim. 9 – Educação, com os seguintes exemplos: aluno + benefício, avaliar + escala, expor + aluno.
32 Etiquetador disponível no site: http://ucrel-api.lancaster.ac.uk/usas/tagger.html

> *Terms; The Body & The Individual; Arts & Crafts; Emotional Actions, States & Processes; Food & Farming; Goverment & The Public Domain; Architecture, Buildings, Houses & The Home; Money & Commerce; Entertainment, Sports & Games; Life & Living Things; Movement, Location, Travel & Transport; Numbers & Measurement; Substances, Materials, Object and Equipment; Education; Language and Communication; Social Actions, States & Processes; Time; The World & Environment; Psychological Actions, States & Processes; Science & Technology; Names & Grammatical Words.*[33]

Os autores (DELFINO e BERBER SARDINHA, 2021) identificaram as 05 dimensões semânticas do corpus CoLiE (*Corpus of Lyrics in English*): Dimensão 1 – *Society, emotions, and actions*; Dimensão 2 – *Evaluation of people in different places;* Dimensão 3 – *Necessities among people*; Dimensão 4 – *Age and sanity*; Dimensão 5 – *Planning and making decisions in in life and love*.

5. CONSIDERAÇÕES FINAIS

A Análise Multidimensional é uma metodologia cujo objetivo é revelar a variação textual em *corpora* eletrônicos por meio de procedimentos estatísticos, mais especificamente a análise fatorial. Variação essa dada em forma de escala, auxiliando no entendimento das variantes linguísticas estudadas, por meio de suas propriedades comunicativas, funcionais e discursivas compartilhadas por determinados registros e que estão subjacentes aos textos. Biber em 1988 trouxe ao mundo essa técnica que possibilitou revelar as dimensões funcionais/gramaticais da língua inglesa, porém a língua é formada não apenas por características estruturais. Faltava uma abordagem de variação da língua do ponto de vista lexical/temático para podermos chegar na lexicogramática, que vem a ser um dos pilares da linguística sistêmico-funcional (BERBER SARDINHA, 2020) e, que atribui à semântica, juntamente com o léxico e a gramática uma visão holística da língua, que é vista como um *continuum*, com a gramática numa ponta e o léxico em outra, conforme podemos ver na figura 9.

33 Termos Gerais e Abstratos; O Corpo e O Indivíduo; Artes e Ofícios; Ações Emocionais, Estados e Processos; Comida e Agricultura; Governo e o Domínio Público; Arquitetura, Prédios, Casas e Lar; Dinheiro e Comércio; Entretenimento, Esportes e Jogos; Vida e Coisas Vivas; Movimento, Localização, Viagem e Transporte; Números e Medidas; Substâncias, Materiais, Objeto e Equipamento; Educação; Línguas e Comunicação; Ações Sociais, Estados e Processos; Tempo; O Mundo e o Meio Ambiente; Ações Psicológicas, Estados e Processos; Ciência e Tecnologia; Nomes e Palavras Gramaticais.

Ao enxergarmos a língua pelo pólo da gramática, a consideraremos um sistema estrutural, fechado, enquanto se nos movermos para o lado lexical, estaremos vendo a língua em função de seu significado, como um sistema aberto. A Linguística de *Corpus* também trabalha com a lexicogramática, já que se dedica ao estudo do léxico e da gramática e do relacionamento entre os dois a partir da análise de dados presentes em um *corpus*, dados esses que muitas vezes chamamos de colocação, coligação, fraseologia, padrão lexical, *chunk, lexical bundle*, linguagem formulaica, *lexical frame,* multi-palavras, entre outros. Sinclair, ao falar do termo lexicogramática, afirma que ela 'não integra os dois tipos de padrão, como o nome pode sugerir – ela basicamente vem a ser o estudo da gramática com um olhar atento aos padrões lexicais dentro dos parâmetros gramaticais'[34] (SINCLAIR, 1970/2004, p.164).

Figura 9. O continuum gramática – léxico. **Fonte:** Adaptado de Halliday e Mathiessen (2004, p. 43).

Biber tem trabalhado com uma grande variedade de unidades lexicogramaticais e, para tanto, define o termo lexicogramática em Biber, Conrad e Reppen (1998, p.84) como 'associações entre palavras e estruturas gramaticais'[35] e tais padrões podem ser usados para diferenciar palavras que são aparentemente sinônimos. Na verdade, a pesquisa de Biber, ao longo de sua carreira, tem mostrado sistematicamente que o 'registro sempre importa'[36] (GRAY, 2013) e que o léxico e a gramática estão juntos nas 'dimensões de variação de registro', que nada mais são do que agrupamentos de características lexicogramaticais correlacionadas que co-ocorrem frequentemente nos textos e que possuem uma função compartilhada (BIBER, 1988). Porém, os trabalhos de Biber e os de autores que o seguiram, priorizaram as características funcionais/gramaticais dos textos. Se fôssemos colocar os estudos de análise multidimensional funcional/gramatical na escala da figura 9, eles estariam mais próximos do polo gramatical.

Berber Sardinha, a partir de 2014, se valeu dos pressupostos da AMD para investigar o polo do léxico na dada escala. Podemos dizer que as AMDs lexical/temática, colocacional e semântica estão intrinsicamente relacionadas ao léxico, mas cada uma com uma especificidade. O pesquisador deve ter em mente qual é o objetivo da sua pesquisa enfocando o léxico para saber qual das 03 análises faz mais sentido para a sua pesquisa:

34 *...does not integrate the two types of patter as its name might suggest – it is fundamentally grammar with a certain amount of attention to lexical patterns within the grammatical frameworks.*
35 *...associations between words and grammatical structures.*
36 *Register always matters!*

tanto a lexical/temática como a colocacional são recomendadas para estudos temáticos, de representações ou identidades, com a diferença de que na lexical/temática as palavras mais frequentes no *corpus* que darão origem a uma lista de palavras que serão analisadas, enquanto que na colocacional, apesar da lista de palavras nódulo + colocados serem geradas a partir das palavras mais frequentes, o que vai ser levado em consideração é a força de atração entre as palavras, ou seja, na AMD colocacional as variáveis serão pares de palavras que ocorrem juntas com uma alta frequência.

Já a AMD semântica lida com campos semânticos pré-definidos pelo etiquetador. Assim como a AMD funcional/grammatical, qualquer estudo realizado valendo-se da AMD semântica, as variáveis serão as mesmas, enquanto nas AMDs lexical/temática e colocacional, as variáveis diferirão de *corpus* para *corpus*. Gostaria de terminar este capítulo com duas figuras esquemáticas (figuras 10 e 11), demonstrando como os 04 tipos de estudos multidimensionais podem ser entendidos. A figura 10 traz os 04 tipos de análises multidimensionais completas que conhecemos até o momento[37] e, podemos notar que não importa qual o estudo de AMD que o pesquisador realizará, tal estudo tem como base a análise fatorial.

A partir do pré-processamento do *corpus,* o pesquisador deve decidir se irá trabalhar com variáveis lexicais ou gramaticais, de acordo com os objetivos da sua pesquisa. Se o objetivo do estudo for entender a variação funcional/gramatical que ocorre no seu *corpus*, as variáveis selecionadas devem ser as gramaticais e, a partir daí seguir os passos de uma AMD, conforme relatamos no capítulo. Por outro lado, se o objetivo do pesquisador for entender a variação do léxico no seu *corpus*, as variáveis selecionadas devem ser lexicais, que podem ser as palavras de conteúdo mais frequentes do *corpus*, que originará a AMD lexical/temática, os nódulos + colocados mais frequentes do *corpus*, que originará a AMD colocacional ou os campos semânticos, que originará a AMD semântica.

37 A autora não incluiu a análise multidimensional aditiva por não conter todos os passos de uma AMD completa como as outras 04 AMDs presentes nas figuras 10 e 11.

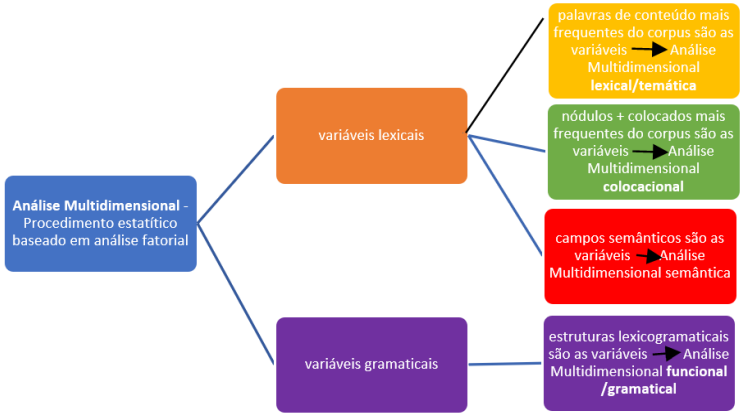


Figura 10. Tipos de Análises Multidimensionais. **Fonte:** Delfino (2021) para esta pesquisa.

Portanto, o pré-processamento do *corpus* deve estar intimamente ligado com o objetivo da pesquisa em estudos envolvendo análises multidimensionais. Em relação à replicabilidade das análises, todo estudo de AMD pode ser replicado, mas o pesquisador precisa ter em mente que análises funcionais/gramaticais e semânticas possuem um número limitado de variáveis que se repetem independentemente do *corpus* que está sendo analisado. Por outro lado, análises lexicais/temáticas e colocacionais terão suas variáveis escolhidas de acordo com a frequência que elas aparecem no *corpus* e tais variáveis não serão as mesmas em todos os estudos, conforme podemos ver na figura 11.

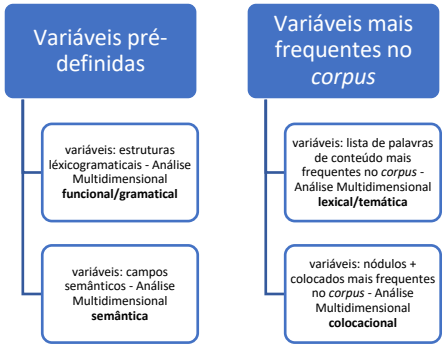


Figura 11. Tipos de variáveis**. Fonte:** Delfino (2021) para esta pesquisa.

As línguas são vistas como sistemas probabilísticos e as escolhas de palavras que os falantes de uma língua fazem, mesmo que inconscientemente, podem não ocorrer ao acaso (BERBER SARDINHA, 2004). Antes do advento da AMD, a Linguística de *Corpus* valia-se dos padrões presentes nas colocações, coligações e preferência semântica. A AMD vem a ser um passo além, na medida em que se vale da variação presente nesses padrões, mas numa quantidade de textos muito maior e, mostrando padrões de co-ocorrência em diferentes registros, ou seja, o que é comum em um registro como *horóscopo* pode estar presente em

um *artigo científico* (BERBER SARDINHA, KAUFFMAN, ACUNZO, 2014). O que à primeira vista pode parecer estranho pode ser comprovado através da co-ocorrência de características presentes em vários textos, mostrando funções compartilhadas em diferentes tipos de registros.

A partir da análise multidimensional foi possível a descrição de registros muito parecidos e de registros diferentes, mostrando funções comunicativas que os une e que os separa. Biber trouxe ao mundo as dimensões funcionais/gramaticais da língua inglesa e os trabalhos que se seguiram focaram nessa perspectiva. A partir de 2014, Berber Sardinha apropriou-se do outro lado da escala da figura 9, o léxico que, por ser um sistema aberto, é confuso e necessita de um olhar por várias perspectivas, daí a necessidade de 03 tipos de MDA para identificar padrões de variações lexicais, seja temático, colocacional ou semântico. Hoje em dia, para poder se fazer uma descrição de uma língua, variedade linguística ou mesmo de um registro, faz-se necessário cada vez mais trabalhos que unam o quantitativo e o qualitativo, levando o estudo linguístico para além do texto, afinal Sinclair já dizia na década de 90 que 'a língua parece muito diferente quando você olha para um monte de textos ao mesmo tempo'.

REFERÊNCIAS

BERBER SARDINHA, T. *Linguística de Corpus*. Barueri, Manole, 2004.

____. On being American and Brazilian, 2014 (Não publicado).

BERBER SARDINHA, T.; KAUFFMANN, C.H.; ACUNZO, C. A multidimensional analysis of register variation in Brazilian Portuguese. *Corpora*, v.9, n.2, p. 239-271, 2014.

BERBER SARDINHA, T. *A corpus-based history of Applied Linguistics*. Apresentação de Trabalho. Associação Internacional de Linguística Aplicada (AILA), Rio de Janeiro, 2017.

____. Lexicogrammar. In: CHAPELLE, C.A. (org). *The concise encyclopedia of Applied Linguistics*. Wiley Blackwell, pp. 701-706, 2020.

BERBER SARDINHA, T. A historical characterization of American and Brazilian cultures based on lexical representations. In: *Corpora*, v.14, n.2, pp. 183-212, 2020(b).

BERBER SARDINHA, T. e VEIRANO PINTO, M. (org.) *Multidimensional Analysis*. Bloomsburry Academic, 2019.

BIBER, D. Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics* 23, pp. 337-360, 1985.

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. New York: Cambridge University Press, 1998.

BIBER, D. *University language*: A corpus-based study of spoken and written registers. Amsterdam/Philadelphia: John Benjamins, 2006

BIBER, D.; TRACY-VENTURA, N. Dimensions of register in Spanish. In: PARODI, G. (Org.) *Working with Spanish corpora*. London: Continuum, 2007.

BREZINA, V., MCENERY, T., & WATTAM, S. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139-173, 2015.

DAVIES, M. *Corpus of Contemporary American English*. Available at corpus.byu.edu/full-text., 2012.

DELFINO, M. C. N. *Uso de música para o ensino de inglês como língua estrangeira em um ambiente baseado em corpus*. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem), Pontifícia Universidade Católica de São Paulo. São Paulo, 2016.

DELFINO, M.C.N. e BERBER SARDINHA, T. *A multi-dimensional view of pop music*: a metadata-rich semantic analysis. Apresentação de Trabalho. American Association of Applied Linguistics (AAAL). Online conference, 2021.

GRAY, B. Interview with Douglas Biber. In: Journal of English Linguistics, 41(4), pp. 359-379, 2013.

HALLIDAY, M. A. K., e MATTHIESSEN, C. M. I. M. *An introduction to functional grammar* (3rd ed.). London, England: Edward Arnold., 2004.

KAUFFMANN, C. H. *O corpus do jornal*: variação linguística, gênero e dimensões da imprensa diária escrita. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem), Pontifícia Universidade Católica de São Paulo. São Paulo, 2005.

MOLLET, E., WRAY, A., e FITZPATRICK, T. Accessing second-order collocation through lexical co-occurrence networks. In: T. Herbst, S. Faulhaber & P. Uhrig (Eds.), *Phraseological View of Language: A Tribute to John Sinclair*. Berlin: Mouton de Gruyter, pp. 87-122, 2011.

RYCHLY, P. A lexicographer-friendly association score. In P. Sojka & A. Horák (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN* (pp. 6-9). Brno: Masaryk University, 2008.

SINCLAIR, J. *Corpus, concordance and collocation*. Oxford University Press, 1991.

SINCLAIR, J. M., JONES, S., e DALEY, R. English lexical studies: Report to OSTI on project C/LP/08. In R. Krishnamurthy (Ed.), *English collocation studies*: The OSTI report (pp. 2–138). London, England: Continuum., 1970/2004.

WILLIAMS, G. Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3(1), 151–171. https://doi.org/10.1075/ijcl.3.1.07wil, 1998.

ZUPPARDI, M.C. *Collocation dimensions in academic English.* Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem), Pontifícia Universidade Católica de São Paulo. São Paulo, 2020.

LEITURA RECOMENDADA

Para um maior aprofundamento da análise multidimensional, a autora sugere a leitura do livro *Multidimensional Analysis*, de Berber Sardinha e Veirano Pinto (org.), de 2019, que engloba em mais detalhes todos os passos de uma análise multidimensional além de pesquisas recentes na área.