

Variação de registro no alemão contemporâneo: estudo preliminar

Register variation in contemporary German: preliminary analysis

COSTA, Andressa⁸⁵

Resumo: Este artigo apresenta um estudo preliminar sobre variação de registro no alemão contemporâneo. Trata-se de um estudo com base em um corpus de multirregistros e que tem como método a análise multidimensional, uma abordagem metodológica da linguística de corpus que usa ferramentas da linguística computacional e procedimentos estatísticos para identificar grupos de características que coocorrem em alta frequência em textos. Após a identificação dos grupos, é feita uma análise qualitativa para determinar as funções comunicativas ou dimensões que as características coocorrentes desempenham nos textos. Nesse estudo, foram usadas 47 variáveis linguísticas e identificadas cinco dimensões. A interpretação dessas dimensões, no entanto, tem caráter provisório e servirá de base para análise principal.

Palavras-chave: Linguística de corpus; variação de registro; análise multidimensional, alemão

Abstract: This article presents a preliminary study on register variation in contemporary German. The study is based on a multi-register corpus and uses the multidimensional analysis, a methodological approach from corpus linguistics which uses computational linguistic tools and statistical procedures to identify groups of characteristics that co-occur in high frequency in the texts. After the identification of the groups, a qualitative analysis is performed to determine the communicative functions or dimensions, that the co-occurring characteristics have in the texts. In this study, 47 linguistic variables were used and five dimensions were identified. The interpretation of these dimensions, however, is provisional and will serve as basis for the main analysis.

Keywords: Corpus linguistics; register variation; multidimensional analysis; German

⁸⁵ PUC São Paulo/GELC. E-mail: acosta.andressa@gmail.com

1 Introdução

O presente estudo se propõe a investigar a variação entre registros no alemão através da análise multidimensional. Essa abordagem metodológica de cunho empírico foi desenvolvida por BIBER (1988) e aplicada primeiramente para o inglês. Ao desenvolver a abordagem multidimensional, BIBER (1988, p. 13) parte do princípio de que padrões de características linguísticas que coocorrem consistentemente em textos sinalizam dimensões funcionais latentes. O objetivo da análise multidimensional é, portanto, identificar as dimensões linguísticas que emergem da análise quantitativa e qualitativa de padrões linguísticos coocorrentes em um corpus. Com base na revisão da literatura especializada, são identificadas características linguísticas consideradas relevantes e que contribuem para a diferenciação entre registros em uma língua específica.

Um aspecto importante nesse estudo multidimensional da variação textual é que texto é considerado sob a perspectiva de registro. Isso implica que: 1) A análise pode ser feita tanto em textos completos, quanto em trechos; 2) Qualquer característica léxico-gramatical pode servir de base para a análise, desde que ela seja frequente e recorrente nos textos de uma dada variedade; 3) A interpretação dos resultados é de base funcional, que atribui a cada fator uma ou mais funções comunicativas compartilhadas pelas características linguísticas que o compõem.

Esse trabalho tem como base um corpus composto por 49 registros das modalidades escrita, falada, discurso roteirizado e internet. A abordagem multidimensional não é de todo desconhecida dos linguistas alemães. No entanto, ainda não se tem conhecimento na literatura especializada de estudos com esse método aplicado ao alemão. Logo, esse estudo se propõe a preencher essa lacuna respondendo às seguintes perguntas: 1) Que características linguísticas coocorrem frequentemente nos registros estudados? 2) Que função comunicativa elas exercem nesses registros? 3) Quais são as dimensões de variação de registro no alemão?

Trata-se aqui de uma análise preliminar com base em 47 características linguísticas que puderam ser extraídas do corpus sem necessidade de uma preparação maior do material, trazendo, portanto, algumas contribuições importantes: por um lado, ele contribuiu para uma melhor definição do conjunto de características linguísticas a serem usadas na análise principal; por outro lado, ele revelou um primeiro conjunto de dimensões textuais do alemão que podem ser confirmadas e melhor desenvolvidas

futuramente. Os procedimentos metodológicos da análise multidimensional, bem como o corpus design, são descritos na seção 2. Na seção 3 são apresentados os resultados das análises, especificando quais características linguísticas compõem cada dimensão identificada e quais registros são mais marcados em cada dimensão. Além disso, é feita uma primeira tentativa de rotular cada dimensão de acordo com a função ou as funções comunicativas compartilhadas pelas características linguísticas que compõem as dimensões. Por fim, na seção 4, são abordadas as implicações do estudo para a pesquisa sobre variação textual no alemão, suas limitações e o desenvolvimento futuro da pesquisa.

2 Método e dados da análise

Estudos baseados na análise multidimensional apresentam as seguintes características metodológicas (BIBER, 1988, p. 63): a) uso de corpora como base de dados; b) uso de programas de computador para contar a frequência das características linguísticas estudadas em uma grande variedade de textos; c) uso de técnicas estatísticas multivariadas, especialmente a análise fatorial, para identificar padrões de coocorrência de características linguísticas nos textos; d) uso de técnicas de análise microscópicas para interpretar os padrões de coocorrência subjacentes aos textos e identificar a função ou as funções comunicativas desses padrões.

A análise multidimensional é realizada em três etapas (BIBER, 1988, p. 64): a) Na primeira etapa, é feita uma revisão de pesquisas anteriores e gramáticas para a identificação de características linguísticas importantes. Em seguida, procede-se à compilação de um corpus em formato digital composto por uma gama variada de registros. Após a limpeza e etiquetagem do corpus, faz-se a contagem da frequência de ocorrência das características linguísticas nos textos através de programa de computador. b) Na segunda etapa, é realizada uma análise fatorial para identificar os grupos de características linguísticas que coocorrem em alta frequência nos textos (fatores). A seguir, os fatores são interpretados como dimensões textuais a partir da definição da função comunicativa mais amplamente compartilhada pelas características linguísticas que os compõem. c) Na terceira etapa, computam-se escores de fator para cada texto, situando-os nas dimensões (mais marcado, menos marcado e não marcado) e escores de fatores médios para os textos dentro de cada gênero, indicando quais textos são bons

representantes das suas categorias textuais. A seguir são descritos os procedimentos seguidos para a realização desse estudo:

2.1 Seleção das características linguísticas

A partir da revisão da literatura especializada, foram identificadas características linguísticas importantes do alemão. Esse estudo preliminar se baseou em características mais gerais, que foram possíveis de se extrair com as ferramentas disponíveis. O intuito da análise preliminar é testar o modelo com uma constelação determinada de características linguísticas e identificar que aspectos precisam ser mais desenvolvidos. As características analisadas foram: a) adjetivo (atributivo e predicativo), b) partícula (resposta, intensidade, negação e verbal), c) advérbio, d) artigo (definido, indefinido, demonstrativo e possessivo), e) número cardinal, f) conjunção (comparativa, coordenada, subordinada com verbo finito e subordinada com infinitivo), g) preposição, h) interjeição, i) substantivo (nome próprio e substantivo comum), j) pronome (1.^a pessoa, 2.^a pessoa, 3.^a pessoa, indefinido, demonstrativo, possessivo, relativo, reflexivo, interrogativo), k) advérbio pronominal (regular e interrogativo), l) verbo finito (auxiliar, modal e cheio no indicativo presente; auxiliar, modal e cheio no indicativo passado; auxiliar, modal e cheio no *Konjunktiv I* e auxiliar, modal e cheio no *Konjunktiv II*), m) imperativo, n) infinitivo com *zu*, o) particípio passado.

2.2 Compilação de um Corpus

Um corpus de multirregistros (*Koder – Korpus deutscher Register*) foi construído especialmente para o projeto (COSTA, 2019). Ele é composto por 50 registros, 2.772 textos e mais de 12 milhões de tokens. Os textos coletados foram produzidos entre 1990 e 2019 e representam o alemão contemporâneo. Após a coleta, os textos foram limpos semiautomaticamente e etiquetados com o RFTagger (SCHIMD, LAWS, 2008).

2.3 Contagem da frequência das características linguísticas nos textos

Após a construção e preparação do corpus (limpeza e etiquetagem), fez-se a contagem normalizada da frequência das características linguísticas nos textos com programa de computador. A frequência das características linguísticas foi normalizada

para um texto com total de 1.000 palavras para possibilitar a comparação da frequência de ocorrência das características estudadas em textos com tamanhos diferentes.

2.4 Análise Fatorial

A contagem normalizada da frequência das características linguísticas é usada na análise fatorial para identificar grupos de características que coocorrem nos textos em alta frequência. Essa técnica estatística utiliza métodos multivariados para investigar correlações subjacentes a grupos de variáveis observadas e determinar o menor número de variáveis para explicar a variância nos dados (LOEWEN, GONULAL, 2015, p. 182). A análise fatorial foi conduzida nesse estudo com SAS (Statistical Analysis System) seguindo vários passos, descritos a seguir:

a) Fatorabilidade dos dados

A adequação dos dados para a análise fatorial foi checada através de testes estatísticos, isto é, se o tamanho da amostra e o nível de correlação das variáveis independentes são apropriados. Tanto a correlação alta, como a correlação baixa entre as variáveis podem prejudicar os resultados da análise estatística. O coeficiente de correlação de Pearson foi usado para medir a existência de correlação alta entre as variáveis, enquanto que o teste de esfericidade de Bartlett foi usado para checar a existência de correlação baixa. O teste Kaiser-Meyer-Olkin (KMO) foi usado para medir se o tamanho da amostra era apropriado para a análise.

b) Análise Fatorial Inicial

Após checar a adequação dos dados, procedeu-se à análise fatorial inicial não rotacionada com o método principal para extração dos fatores estatisticamente importantes. O número de fatores a serem considerados na interpretação dos dados foi definido a partir observação do *Scree plot* gerado na análise fatorial e do *eigenvalue* de cada fator. Foram considerados apenas fatores com *eigenvalues* maiores que 1.

c) Determinação do número de fatores

De acordo com *Scree plot* abaixo, quatro ou cinco fatores podem ser analisados. Como se pode notar no gráfico, a capacidade de um fator de explicar a variância nas variáveis observadas, medida pelo *eigenvalue*, diminui a partir do sexto fator. Mas ela é relativamente grande entre o quarto e o quinto fator, o que sugere uma solução de quatro ou cinco fatores. Decidiu-se por uma solução com cinco fatores para a análise.

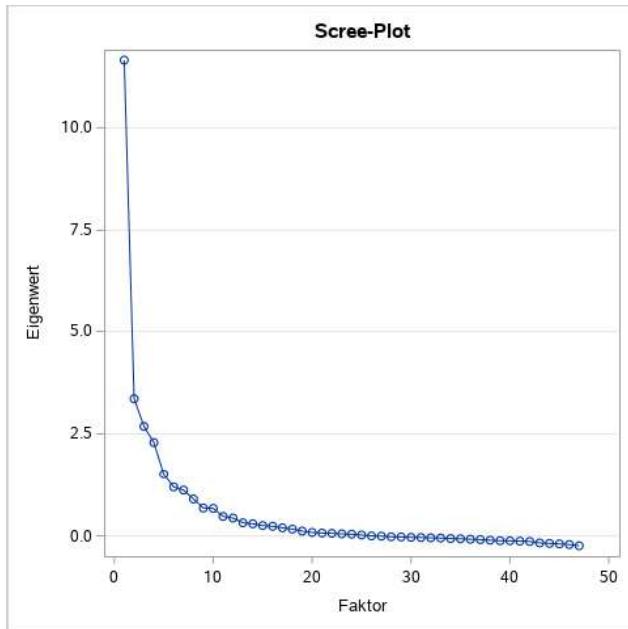


Gráfico 1: *Scree plot* de *Eigenvalues*

d) Análise Fatorial Rotacionada

Definido o número de fatores, procedeu-se à análise fatorial rotacionada com o método promax. Na extração inicial não rotacionada dos fatores, a maioria das variáveis carrega no fator 1, o que dificulta a interpretação, já que os outros fatores apresentam uma quantidade muito reduzida de variáveis. Essa distribuição desproporcional das variáveis é compensada pela rotação dos fatores, fazendo com que cada variável carregue em poucos fatores. Desse modo, cada fator é caracterizado por um grupo pequeno, porém significante de variáveis que melhor representam a parcela de contribuição do fator para explicar a variância compartilhada. Seguindo a recomendação em BIBER (1988, p. 88), foram excluídas da análise variáveis com carga fatorial abaixo de 0,30.

e) Interpretação dos fatores com base nas cargas fatoriais

Nesse passo, cada fator é interpretado qualitativamente com base nas variáveis que o compõe, para identificar pelo menos uma função comunicativa que explique a coocorrência das características linguísticas nos textos. A interpretação dos fatores com base na carga fatorial das características linguísticas considera apenas a relação destas variáveis com o fator no qual elas carregaram. Para incluir os textos na análise qualitativa, é necessário um procedimento complementar à análise fatorial, a saber, a computação dos escores de fator.

2.5 Escores de fator

O escore de fator caracteriza cada texto em relação a cada fator. Desse modo, é possível observar quais textos são mais marcados em relação a um fator, quais textos são menos marcados e quais textos não são marcados. O escore de fator é computado para cada texto em relação a cada fator. A computação dos escores de fator para esse estudo foi feita do seguinte modo: 1) Padronização das frequências para uma média de 0,0 e desvio padrão de 1,0. Isso previne que características que ocorrem com muito mais frequência exerçam maior influência na computação dos escores de fator do que aquelas com menor frequência (BIBER, 1988, p. 94); 2) Computação dos escores de fator para cada texto somando o número de ocorrências das características linguísticas em cada texto que compõem cada fator.

2.6 Interpretação – Dimensões textuais de variação

Nessa segunda fase da interpretação dos resultados, o texto é a base de análise. A primeira interpretação feita considerando apenas as cargas fatoriais é revista. As funções comunicativas subjacentes aos textos identificadas através da interpretação das características linguísticas coocorrentes são definidas como dimensões de variação textual.

3 Resultados e Discussão

A análise fatorial, com base em 47 variáveis linguísticas, identificou uma solução de cinco fatores. A estrutura fatorial final é descrita na Tabela 1. A carga fatorial indicada entre parênteses indica o peso de cada característica linguística no fator. Como se pode

observar na tabela, o fator 1 é o único que apresenta dois grupos (polo positivo e polo negativo) de variáveis linguísticas que coocorrem nos textos em uma relação de distribuição complementar. Todos os outros fatores apresentam apenas um grupo.

Fator 1	
Polo positivo	advérbio (0.88237), pronome demonstrativo (0.77100), pronome indefinido (0.76813), interjeição (0.68043), adjetivo predicativo (0.57211), conjunção (0.48049), <i>Konjunktiv II</i> (0.44978), verbo auxiliar no presente do indicativo (0.34306), negação (0.39394), pronome interrogativo (0.32757), pronome de 1. ^a pessoa (0.44982), partícula de resposta (0.43663), pronome adverbial interrogativo (0.36611)
Polo negativo	substantivo (-0.86557), preposição (-0.74208), adjetivo atributivo (-0.59709), artigo definido (-0.58177), preposição com artigo (-0.49817), número cardinal (-0.38726), palavra composta truncada (-0.38515)
Fator 2	
Polo positivo	verbo auxiliar no presente do indicativo (0.69169), pronome de 2. ^a pessoa (0.66535), verbo modal no presente do indicativo (0.65417), negação (0.60934), pronome interrogativo (0.60923), pronome de 1. ^a pessoa (0.58638), imperativo (0.56392), partícula de resposta (0.54154), pronome adverbial interrogativo (0.50392), verbo cheio no presente do indicativo (0.35728)
Fator 3	
Polo positivo	partícula <i>zu</i> (0.65732), pronome relativo (0,55), artigo demonstrativo (0.46413), pronome adverbial (0.42733), conjunção com infinitivo (0.42031), infinitivo com <i>zu</i> (0.40929), pronome indefinido atributivo (0.35718), conjunção (0.43404), pronome reflexivo (0.393)
Fator 4	
Polo positivo	verbo passado do indicativo (0.74443), verbo auxiliar <i>Konjunktiv I</i> (0.56104), participípio II (0.49956), nome próprio (0.45568), verbo modal <i>Konjunktiv I</i> (0.37053), verbo modal passado do indicativo (0.35527)
Fator 5	
Polo positivo	pronome possessivo (0.66807), pronome de 3. ^a pessoa (0.5798), verbo cheio presente do indicativo (0.48115), partícula verbal (0.47756), pronome reflexivo (0.44332)

Tabela 1: Estrutura Fatorial

A interpretação dos fatores é um processo de análise qualitativa, que tem como objetivo identificar a função comunicativa que as características linguísticas significativamente coocorrentes exercem nos registros estudados. Nesse processo, observam-se não só a composição de cada fator como também os registros mais marcados neles. Identificada a função comunicativa compartilhada pelas características linguísticas que compõem os fatores, estes recebem um rótulo que descreve essa função compartilhada, passando assim cada fator a constituir uma dimensão de variação textual. Foram identificadas as seguintes cinco dimensões. Ressalte-se que os rótulos abaixo são provisórios:

Dimensão 1: Discurso oral vs. discurso letrado

Dimensão 2: Oralidade simulada

Dimensão 3: Discurso com foco explicativo

Dimensão 4: Discurso narrativo

Dimensão 5: (sem rótulo)

3.1 Dimensão 1: Discurso oral vs. discurso letrado

3.1.1 Polo Positivo: Discurso oral

O polo positivo da dimensão 1 contém 12 variáveis que coocorrem com alta frequência nos registros estudados. A variável com maior carga fatorial é o advérbio. No entanto, ao observar a ocorrência de advérbios nos dados, nota-se que essa categoria abrange também partículas, tais como *ja*, *also*, *doch*, *auch*, *nur*, *so* etc., que são particularmente frequentes na comunicação oral (cf. DUDEN, 2009, p. 588). Isso se deve ao fato de que o etiquetador RFTagger não distingue entre advérbios e partículas. De fato, os etiquetadores existentes foram treinados em textos escritos e por isso não incluem características típicas da língua falada. O problema da não adequada anotação de características da língua falada nos corpora do arquivo foi abordado em WESTPFAHL, SCHMIDT (2013). Os cinco primeiros registros mais marcados nesse polo são registros falados provenientes do Banco de Dados do alemão falado do IDS (DGD) em Mannheim. Considerando que a variável advérbio no polo positivo da dimensão 1 inclui também partículas, e considerando que os registros mais marcados aqui são registros falados,

supõem-se que grande parte das ocorrências aqui descritas como advérbio são, na verdade, partículas. Além disso, ao observar as ocorrências de advérbios nos registros não marcados nesse polo, nota-se que neles se trata de fato de advérbios, e não partículas. Outro aspecto a se considerar são as interjeições que apresentam nessa dimensão uma carga fatorial alta. Interjeições são classificadas na gramática alemã também como um tipo de partícula. Isso implica que um grande número de partículas etiquetadas ora como advérbio, ora com interjeição coocorrem com outras características linguísticas no polo positivo da dimensão 1. Essa é, no entanto, uma questão a ser melhor observada na análise principal.

Outra característica linguística importante no polo positivo do fator 1 são os pronomes demonstrativos. Primos demonstrativos são usados para referir-se a objetos, pessoas e eventos situando-os num espaço e/ou tempo, no caso dos pronomes demonstrativos dêiticos, ou no discurso, no caso dos pronomes demonstrativos anafóricos (BECHARA, 2004, p. 187). Primos demonstrativos em alemão são: *der/die/das*, *dieser*, *jener*, *derjenige*, *derselbe* (DUDEN, 2009, p. 280). Entre eles, os tipos que mais ocorrem nos principais registros do polo positivo do fator 1 são *der/die/das*. Ao comparar com a ocorrência de demonstrativos nos registros mais marcados do polo negativo, percebe-se que há mais casos da forma genitiva (*dessen*, *derem*, *derer*) e da forma *dies-*. Um estudo interessante a ser realizado a partir dessa observação é investigar se há diferença na distribuição dos pronomes demonstrativos de acordo com o registro, isto é, se *der/die/das* são mais comuns em registros falados, enquanto que as outras formas são mais comuns em registros escritos, pois essa diferença de uso não é abordada nas gramáticas do alemão.

A categoria adjetivo predicativo abrange adjetivos que se referem a frases nominais e são termos oracionais independentes que constituem complemento de certos verbos (cf. DUDEN, 2009, p. 352). A variável oração subordinada com verbo finito abarca tanto conjunções que introduzem orações subordinadas substantivas (*Inhaltssatz*), quanto conjunções que introduzem orações subordinadas adverbiais. Para a análise principal, é importante diferenciar semanticamente as orações subordinadas. Contudo, nesse estudo preliminar, considerar-se-á apenas o fato de que a alta frequência de orações subordinadas e complemento predicativo indica que os textos marcados no polo positivo da dimensão 1 apresentam um forte estilo verbal.

Por fim, temos a variável verbo no passado do subjuntivo que descreve na verdade o modo *Konjunktiv II (KII)*. Ao observar as ocorrências dessa variável nos textos mais

marcados, nota-se que em praticamente 95% dos casos trata-se dos verbos auxiliares *sein*, *haben*, *werden* (usado para formar o *KII* da maioria dos verbos cheios) e verbos modais. O modo *Konjunktiv II* é usado no alemão para expressar irrealidade, pedidos/perguntas/propostas com cortesia (sobretudo com os verbos modais) e discurso reportado na língua falada e informal especialmente a forma *würde* + *Infinitiv*. O exemplo abaixo é um trecho retirado do subcorpus *Interview*, o registro mais marcado no polo positivo da dimensão 1. O exemplo ilustra o uso das características coocorrentes nos registros típicos desse polo:

- (1) MF *ich* schalte (.) das mal an (.) und höre da mal rein
 MF ja
 MF scheint zu funktionieren

- (2) BWS2 dann geht er halt weg]
 MF ach so hmhm na gut
 MF ja dass das so unangenehm sein könnte wenn man na ja

**[Pronome 1^a pessoa - Pronome demonstrativo - Advérbio/Partícula – Advérbio
- Conjunção subordinada – Adjetivo predicativo - Konjunktiv II]**

Como se observa no exemplo acima, as partículas (*ja*, *so*) são classificadas como advérbio. Além dessa característica, há também os pronomes demonstrativos, verbo no presente do indicativo e pronome de 1.^a pessoa que são marcas de referência dêitica e situam o processo comunicativo no aqui e agora dos interlocutores, adjetivo predicativo e conjunção subordinada que indicam um estilo verbal e *Konjunktiv II*. Os 10 registros mais marcados no polo positivo da dimensão 1 são registros de fala (tanto de fala espontânea, quanto de discurso roteirizado) e constituem na sua maioria interação oral: *Interview*, *Alltagsgespräch*, *Prüfungsgespräch*, *Institutionelle Kommunikation*, *Wissenschaftlicher Vortrag*, *TED Talk*, *Krimiserie*, *Fernsehserie*, *Komödie*, *Filmdrama*.

Faltam, todavia, outras marcas importantes de interação nesse fator como pronome de não 2.^a pessoa e imperativo que aparecem na dimensão 2. Além disso, as características com mais peso sugerem como principal função comunicativa do polo positivo da dimensão 1 a marcação de oralidade, sobretudo quando se compara esse polo com o polo negativo.

3.1.2 Polo Negativo: Discurso letrado

O polo negativo da dimensão 1 é composto por características linguísticas que marcam um estilo nominal: substantivo com elementos que o acompanham como preposição, artigo, número cardinal e adjetivo em função atributiva. Além disso, há a presença de uma variável que parece ser típica do polo negativo dessa dimensão, já que ela só aparece aqui, enquanto as outras variáveis também ocorrem em outras dimensões. Trata-se de uma forma de apresentação de palavras compostas que apresentam uma parte em comum, como por exemplo: *Mittel- und Westeuropa* (Europa central e ocidental). A palavra *Europa* é o elemento comum nas duas palavras (*Mitteleuropa, Westeuropa*). O exemplo (3) ilustra um trecho de texto marcado nesse polo:

(3) Besser als **gewöhnliche Wortschatzentlehnungen** sind sie daher geeignet, die Interaktion der Sprachkontakt- mit der Sprachstrukturgeschichte (...) der beteiligten Einzelsprachen zu beleuchten. (agw01.txt)

[Adjetivo atributivo - Substantivo comum - palavra composta truncada - Preposição - Artigo definido]

Os 10 registros mais marcados contêm em sua maioria textos escritos e um texto de discurso roteirizado (notícia de telejornal). Características comuns a esses textos são linguagem altamente elaborada e objetiva: *Gesetzliche Texte, Stellenangebot, Business Kommunikation, Rezept, Gesprochene Nachricht, Dissertation Leben-, Ingenieur- und Naturwissenschaft, Wikipedia Artikel, Gebrauchsanleitung, Artikel Geisteswissenschaft, Dissertation Geisteswissenschaft.*

3.2 Dimensão 2: Oralidade simulada

A dimensão 2 apresenta uma estrutura fatorial bastante interessante. Por um lado, o seu polo positivo contém seis variáveis que também carregaram no polo positivo do fator 1, porém, elas obtiveram aqui maior carga fatorial, isto é, sua contribuição é maior na dimensão 2 do que na dimensão 1. Isso indica que há uma relação muito próxima entre as duas dimensões. Os 10 registros mais marcados na dimensão 2 são em sua maioria registros de discurso roteirizado, ou seja, textos escritos para serem falados: *Fernsehserie, Komödie, Filmdrama, Krimiserie, Liedertext, Theaterstück, Dokumentarfilm, Institutionelle Kommunikation, Professioneller Chat, TED talk.*

Além disso, a dimensão 2 compõe-se de várias características linguísticas que sinalizam interação. Pronomes de 1.^a e 2.^a pessoa, sobretudo no singular, tem função dêitica (cf. DUDEN, 2009, p. 263). Eles referem-se às pessoas participantes do processo de interação verbal e por isso presentes no contexto comunicativo. Pronomes interrogativos, partículas de resposta e imperativo são também marcas de interação. Eles pressupõem um eu que fala com um tu/você que por sua vez reage.

Formas verbais no presente do indicativo são uma constante na dimensão 2. Nesse caso, trata-se dos verbos auxiliares *sein* e *haben*, usados para compor o *Präsensperfekt*, para referir-se a acontecimentos passados em relação ao momento de fala atual e verbos modais. Segundo DUDEN (2009, p. 1111), os tempos do presente do indicativo são usados em uma situação que é típica da comunicação oral direta, na qual o ouvinte tem a possibilidade de interagir com o falante. Os exemplos (4) e (5) mostram como essas características são usadas nos textos mais típicos dessa dimensão:

(4) **Wie kommt's, dass du mich besuchst?** - Ach, **ich bin** zufällig in Berlin. (fs01.txt)

(5) **Komm, lass uns** hochgehen. – **Ok.** (fs01.txt)

[Pronome adverbial interrogativo – Presente do Indicativo – Pronome 2.^a pessoa – Pronome 1.^a pessoa – Imperativo – Partícula de resposta]

À primeira vista, conclui-se que pronomes pessoais de 1.^a e 2.^a pessoas, pronomes interrogativos, partículas de resposta, verbo no imperativo e no presente do indicativo coocorrem em textos para produzir situações de interação oral. Essa hipótese é parcialmente confirmada por alguns dos registros mais marcados na dimensão 2. Contudo, há também entre os registros marcados alguns que não tem interação como textos de canção, documentário e TED talk. Além disso, chat não é comunicação oral, mas mediada por computador. Outro aspecto relevante nos registros principais diz respeito ao interlocutor: enquanto nos registros com interação direta, dialógica há uma certa familiaridade do sujeito falante com seu interlocutor, essa familiaridade falta em registros como texto de canção, documentário, TED talk pois o enunciador nesses casos se dirige a um público que ele desconhece. O elemento comum a esses textos é a oralidade, no aspecto conceitual, pois no aspecto medial chat seria excluído. Todavia, a oralidade aqui é distinta da oralidade da dimensão 1. A oralidade presente nos registros

prototípicos da dimensão 2 é simulada no sentido de não ser fala espontânea e sim fala preparada, ensaiada. De fato, os registros prototípicos dessa dimensão compõem-se em sua maioria de textos escritos para serem falados, com exceção de chat e comunicação institucional.

3.3 Dimensão 3: Discurso com foco explicativo

A dimensão 3 caracteriza-se pela recorrência de orações complexas indicada pela alta frequência de conjunções subordinadas. Segundo KOCH (2008, p. 68), conjunções estabelecem “entre orações, enunciados ou partes do texto, diversos tipos de relações semânticas e/ou pragmáticas.” Relações lógico-semânticas são estabelecidas entre orações que compõem um enunciado, enquanto as relações discursivas ou pragmáticas são estabelecidas entre dois enunciados distintos (cf. KOCH, 2008, p. 68; 71). Temos, portanto, na dimensão 3 diversas relações do tipo lógico-semânticas. Os registros mais marcados provêm de diferentes esferas comunicativas: *Plenarprotokolle, Populärwissenschaftlicher Artikel, Politische Rede, Leserbrief, Blog, Meinung, Leserkommentar, Leitartikel, Zeitungsinterview, Sachbuch*.

A partir do grupo de variáveis coocorrentes que compõe o fator 3, observa-se que construções com verbo no infinitivo precedidos pela partícula *zu* são bastante frequentes nos registros marcados nessa dimensão. Esse tipo de construção é representado nos dados pelas variáveis partícula *zu*, conjunção com infinitivo e *zu + Infinitiv*. Analisando as ocorrências das variáveis partícula *zu* e *zu + Infinitiv* no corpus, nota-se que se trata do mesmo tipo de característica linguística: verbo no infinitivo precedido por *zu*. A diferença é que as ocorrências etiquetadas com partícula são verbos simples (*zu interpretieren*), enquanto as ocorrências etiquetadas com *zu + Infinitiv* são verbos compostos com prefixo separável nos quais a partícula *zu* é acrescentada entre o prefixo e o verbo (*nachzudenken*). Isso, porém, não é uma diferença significativa em alemão, já que ela não é de natureza semântica, funcional, mas apenas formal. A diferença relevante no uso de construções infinitivas em alemão é de ordem sintática. Infinitivo com *zu* exerce em alemão as seguintes funções sintáticas: a) complemento oracional com função de sujeito ou objeto de certos verbos e locuções verbais; b) locução conjuntiva subordinativa formada pelas preposições *um (zu)*, *ohne (zu)* e *(an)statt (zu)* usadas para expressar

finalidade (*um ... zu*), substituição (*(an)statt ... zu*) e modalidade (*ohne ... zu*) (cf. WÖLLSTEIN-LEISTEN et al., 2006, p. 50-1).

As ocorrências identificadas como partícula *zu* e *zu + Infinitiv* englobam, no entanto, tanto complemento oracional (a) quanto locução conjuntiva subordinativa (b). Para uma análise mais acurada, é necessário observar a distribuição dos grupos sintáticos separadamente. Além disso, é importante diferenciar o tipo de locução conjuntiva, já que *um...zu*, *ohne...zu* e *anstatt...zu* não estabelecem o mesmo tipo de relação entre as orações que elas conectam. *Um...zu* estabelece uma relação de finalidade, *ohne...zu*, uma relação modal e *anstatt...zu*, de substituição (cf. WÖLLSTEIN-LEISTEN et al 2006: 50-1).

Um segundo grupo de variáveis coocorrentes nessa dimensão compartilha função atributiva, que por sua vez forma duas categorias de acordo com sua realização sintática: a) pronomes demonstrativos e pronomes indefinidos em função atributiva são usados em frases nominais; b) pronomes relativos introduzem orações relativas. Além disso, vale ressaltar que construções com *zu + Infinitiv* também são usados como atributo oracional (cf. WÖLLSTEIN-LEISTEN et al., 2006, p. 43).

Pronomes demonstrativos exercem função de referênciação anafórica. Porém, o modo de referenciar dos demonstrativos difere da referênciação feita por pronomes de 3.^a terceira pessoa. Enquanto estes podem retomar também objetos-de-discurso que foram introduzidos em um momento bem anterior no texto, aqueles referem-se apenas a objetos-de-discurso que aparecem em um contexto imediatamente anterior. Isso se deve à natureza dêitica desses pronomes. Por isso se fala em dêixis anafórica (anadeixis) ao se referir à função dos pronomes demonstrativos. (DUDEN, 2009, p. 280; p. 1107).

Os pronomes indefinidos (*Indefinita*) são usados para referir-se a algo que não é identificável de modo preciso e têm via de regra função anafórica, sobretudo aqueles pronomes que designam quantificação (DUDEN, 2009, p. 309; 1106-7), que são as ocorrências mais frequentes dessa categoria nos registros mais marcados na dimensão 3. Os pronomes relativos também exercem função de referênciação anafórica. Eles conectam duas orações, introduzindo uma oração atributiva na qual fazem referência a um elemento da oração principal ou a toda oração (DUDEN, 2009, p. 302; p. 1068).

Por fim, há os advérbios pronominais (*Pronominaladverb*) ou advérbios preposicionais (*Präpositionaladverb*). Eles são compostos pelo elemento dêítico *da-/dar-* seguidos de preposição. Esse tipo de advérbio pode exercer diferentes funções no texto (cf. GRAMMIS): como pronome anafórico, contribui para a progressão temática; como conector, ele estabelece uma relação semântica específica entre dois enunciados (por

exemplo causa - *deswegen*, consequência – *deshalb*); como correlato de complemento preposicional realizado em forma de oração subordinada. Um aspecto importante a se ressaltar no uso dos pronomes adverbiais com função anaforica diz respeito ao tipo de elemento ao qual ele faz referência: eles podem referir-se a objetos, acontecimentos, ao próprio enunciado anterior, porém, eles não fazem referência a pessoas.

Além dessas características principais, há também grande recorrência de conjunções subordinadas com verbo finito, que também tem a função de estabelecer relações específicas entre orações e pronomes reflexivos, que exercem função de referenciação anaforica. Em (6) temos um enunciado composto pelo encadeamento de várias orações no qual uma ação a ser tomada é apresentada e logo justificada. Em (7) é apresentada uma explicação para algo dito anteriormente e que é retomado no enunciado através do pronome *dabei*. O pronome adverbial *darum* serve como correlato da oração infinitiva seguinte que explica a intenção por trás do enunciado do falante.

(6) Was wir, die Bundesregierung und die Länder, gemeinsam tun sollten, ist, **zu zeigen**, dass das Prinzip des Föderalismus ein Strukturprinzip Europas bleiben muss, und zwar mit mehr Chancen, **um** die kulturellen Identitäten unterhalb der Ebene des Nationalstaates sich stärker ausprägen **zu** lassen, als man es bisher für möglich hielt. (pp01.txt)

[**Infinitivo com zu** - *Conjunção com verbo finito* - Pronome indefinido atributivo - **Conjunção com infinitivo** – Pronome reflexivo]

(7) **Dabei** möchte Jenuwein nun ganz und gar nicht die herausragende Leistung der Genetiker schmätern, ihm geht es **darum**, auf deren Ergebnissen aufbauend die Biologie *voranzutreiben* und *zu klären*, was jenseits der Gene vererbt wird. (plinw10.txt)

[**Pronome adverbial** – *Infinitivo com zu*]

Os enunciados em (6) e (7) são exemplos de progressão textual do tipo explicativo no qual um enunciado esclarece ou justifica o conteúdo de outro enunciado. Esses enunciados podem estar conectados através de conjunções como em (6), ou serem relacionados através de pronome adverbial como em (7). Uma descrição mais detalhada sobre os mecanismos de textualização/discursivização de explicação encontra-se em JAHR (2000). A partir do exposto, conclui-se que os registros que são marcados na dimensão 3 caracterizam-se por uma alta recorrência de enunciados explicativos, o que não significa que os textos sejam essencialmente explicativos, mas que apenas tenham em comum a alta frequência de tais tipos de enunciados.

3.4 Dimensão 4: Discurso narrativo

A dimensão 4 contém alta incidência de verbos no passado do indicativo e verbos no *Konjunktiv I*, além de nomes próprios e particípio passado. Uma análise das ocorrências de particípio passado nos registros marcados na dimensão 4 mostra que se trata do tempo passado *Perfekt*. Os registros mais característicos nessa dimensão abrangem textos de registros jornalísticos, textos fictícios e textos informativos: *Gesprochene Nachrichten, Meldung, Zeitungsartikel, Belletristik, Wikipedia Artikel, Krimiserie, Sachbuch, Dokumentarfilm, Dissertation Lebens-, Ingenieur- und Naturwissenschaft, Zeitungsinterview*.

A partir da observação do grupo de características linguísticas mais frequentemente coocorrentes e dos textos mais marcados, pressupõe-se que esse grupo de características é usado em textos para expressar relatos de eventos. De fato, os principais registros dessa dimensão são textos jornalísticos, que têm função de informar sobre fatos e eventos ocorridos. Outros tipos textuais no grupo com função de informar são documentários e artigos da Wikipédia. Literatura de ficção e série policial compõe um grupo específico, cuja principal função é contar uma história. Livros de não ficção e teses (no caso das áreas de biologia, engenharia e ciências naturais) têm a função de informar para instruir, relatando nesse processo também o que outros autores disseram/escreveram sobre o tema em questão.

Além disso, a dimensão 4 apresenta uma grande recorrência de formas verbais no passado do indicativo. Estas formas verbais englobam verbos cheios, auxiliares e modais usados no tempo *Präteritum*. O *Präteritum* é usado em alemão para narrar eventos acontecidos em um momento de referência anterior ao momento da narração/fala (cf. GRAMMIS). Na perspectiva da linguística textual proposta por WEINRICH (2001), que agrupa as formas verbais nas categorias tempos do comentário (*besprochenen Tempora*) e tempos do relato (*erzählende Tempora*), o *Präteritum*, assim como *Präteritumperfekt* e *Futur des Präteritum/Futurperfekt des Präteritum*, são tempos do mundo narrado (*erzählte Welt*). Eles caracterizam uma situação discursiva relaxada e indicam ao interlocutor que o que está sendo dito se trata de um relato ao qual ele não precisa reagir (DUDEN, 2009, p. 1110-1).

Formas verbais no modo *Konjunktiv I* também são recorrentes nessa dimensão. Elas são usadas em alemão para relatar enunciados proferidos em outro contexto de modo indireto. O modo *Konjunktiv I* nessa função é tipicamente usado em textos jornalísticos para sinalizar distanciamento do autor em relação à veracidade do enunciado que ele está

relatando (cf. GRAMMIS). A variável particípio passado é usada para formar o tempo composto *Präsensperfekt*, usado para se referir a um evento passado, cujo momento de referência é o momento da narração/fala. Ele é composto pelos auxiliares *sein* ou *haben* conjugados no presente do indicativo. No discurso reportado, *sein* e *haben* são conjugados no modo *Konjunktiv I* para sinalizar discurso indireto. Finalmente, a variável nome próprio é usada para se referir a indivíduos como pessoas, lugares, países, regiões etc. específicas que podem ser tanto o objeto-de-discurso sobre o qual é relatado algo, quanto o sujeito enunciador do enunciado relatado. O exemplo (8) ilustra um relato indireto com o conteúdo do relato no modo *Konjunktiv I*, o sujeito enunciador é referido com nome próprio e seu ato de enunciação com verbo no *Präteritum*:

- (8) Der deutsche Außenminister **Frank-Walter Steinmeier** *erklärte*, der Sicherheitsrat habe noch einmal bestätigt, ... (gn01.txt)

[Nome próprio - Verbo Passado Indicativo - Verbo auxiliar Konjunktiv I - Partíciplio Passado]

Partindo dessas considerações, assume-se nesse estudo que a coocorrência frequente de *Präteritum*, *Konjunktiv I* e nome próprio nos textos conferem a eles um caráter narrativo sendo, portanto, a função comunicativa compartilhada por estas características linguísticas expressar narratividade.

3.5 Dimensão 5: (sem rótulo)

A dimensão 5 apresenta uma estrutura fatorial com apenas quatro variáveis (as partículas verbais são parte dos verbos) (cf. tabela 1). Os registros mais marcados abrangem sete textos de ficção e três textos não-ficção. O tipo textual mais marcado é horóscopo: *Horoskop*, *Theaterstück*, *Belletristik*, *Filmdrama*, *Liedertext*, *Fernsehserie*, *Krimiserie*, *Predigt*, *Komödie*, *Gebrauchsleitung*.

Por se tratar de uma dimensão pouco clara no que diz respeito às variáveis que a compõem, prescindiu-se nesse estudo de uma interpretação para determinar a função comunicativa compartilhada pelas características linguísticas. Contudo, essa dimensão foi mantida, pois como se pode observar no gráfico 1 (*scree plot*), ela consegue capturar um alto grau de variância em relação ao fator 4. Na análise principal, será usado um grupo maior de variáveis do que nesse estudo, o que pode melhorar a composição do fator 5, contribuindo para a sua interpretação.

4 Considerações finais

Esse estudo preliminar identificou 5 dimensões de variação textual com base em um número reduzido de características. Ainda assim, ele apresentou resultados consistentes, que poderão ser confirmados na análise principal. Várias características linguísticas importantes não puderam ser incluídas nesse estudo por demandarem mais tempo de preparo do corpus. A realização dessa análise foi importante, contudo, para definir melhor não só quais variáveis linguísticas são de fato interessantes, como também quais aspectos de algumas dessas variáveis devem ser analisadas. Por exemplo, a análise indicou que considerar as conjunções de modo geral pode ser tão interessante quanto especificar os seus tipos semânticos.

Além disso, concluiu-se a partir dos resultados que não há necessidade de especificar todas os tempos verbais com presente, *Perfekt*, *Präteritum*, mais-que-perfeito futuro I e II, pois a oposição significativa em relação às formas temporais é marcada apenas pelo presente do indicativo e *Präteritum*. As outras formas são tempos compostos no alemão, cujo verbo auxiliar está conjugado no presente ou no *Präteritum*. Outro aspecto importante mostrado aqui é a necessidade de separar advérbio de partícula. Essas duas categorias têm distribuição e função distintas. Há outras diferenças na distribuição de características pertencentes a uma mesma categoria que foram salientadas nesse estudo e que serão consideradas na análise principal.

Referências bibliográficas

- BECHARA, Evanildo. *Moderna Gramática Portuguesa*. Rio de Janeiro: Lucerna, 2004.
- BIBER, Douglas. *Variation across speech and writing*. Cambridge: Cambridge University Press, 1998.
- COSTA, Andressa. Koder - A multi-register corpus for investigating register variation in contemporary German. *Research in Corpus Linguistics*, 7, 2019, 69-83. <https://doi.org/10.32714/rcl.07.04>
- DUDEN: *Die Grammatik*. Band 4. 8., überarb. Aufl. Mannheim/Wien/Zürich: Dudenverlag, 2009.
- JAHR, Silke. Vertextungsmuster: Explikation. In: BRINKER, Klaus et al. (Hrsg.) *Text- und Gesprächslinguistik HSk 16.1*, Berlin; New York: de Gruyter, 2000. S. 385-397.
- KOCH, Ingodore. *Coesão textual*. São Paulo: Contexto, 2008.
- GRAMMIS: Präpositionaladverb. In: Institut für Deutsche Sprache: "Wissenschaftliche Terminologie". Grammatisches Informationssystem grammis. DOI: 10.14618/terminologie. Permalink: <https://grammis.ids-mannheim.de/terminologie/207>
- GRAMMIS: Präteritum. In: Institut für Deutsche Sprache: "Systematische Grammatik". Grammatisches Informationssystem grammis. DOI: 10.14618/grammatiksystem Permalink: <https://grammis.ids-mannheim.de/systematische-grammatik/258>

GRAMMIS: Anaphorisches Personalpronomen. In: Institut für Deutsche Sprache: "Systematische Grammatik". Grammatisches Informationssystem grammis. DOI: 10.14618/grammatiksystem Permalink: <https://grammis.ids-mannheim.de/systematische-grammatik/394>

GRAMMIS: Konjunktiv in Text und Diskurs. In: Institut für Deutsche Sprache: "Systematische Grammatik". Grammatisches Informationssystem grammis. DOI: 10.14618/grammatiksystem Permalink: <https://grammis.ids-mannheim.de/systematische-grammatik/683>

LOEWEN, Shawn; GONULAL, Talip. Exploratory factor analysis and principal components analysis. In: PLONSKY, Luke (ed.). *Advancing quantitative methods in second language research*. New York/London: Routledge, 2015, 182-212.

SCHMID, Helmut; LAWS, Florian. *Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging*, 2008 Retrieved from: <https://www.cis.uni-muenchen.de/~schmid/papers/Schmid-Laws.pdf>.

WESTPFAHL, Swantje; SCHMIDT, Thomas. POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. In: *Journal for Language Technology and Computational Linguistics*, H. 1, 2013, S. 139-156.

WÖLLSTEIN-LEISTEN, Angelika; HEILMAN, Axel; STEPAN Peter; VIKNER, Sten. *Deutsche Satzstruktur: Grundlagen der syntaktischen Analysen*. Tübingen: Stauffenburg, 2006.