Pontifícia Universidade Católica de São Paulo

PUC-SP

Elaine Cristina de Oliveira

A linguagem verbal das TED Talks: uma análise multidimensional

Mestrado em Linguística Aplicada e Estudos da Linguagem

São Paulo

2021

Elaine Cristina de Oliveira

A linguagem verbal das TED Talks: uma análise multidimensional

Mestrado em Linguística Aplicada e Estudos da Linguagem

Dissertação apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de MESTRA em Linguística Aplicada e Estudos da Linguagem, sob a orientação do Prof. Dr. Antonio Paulo Berber Sardinha.

São Paulo

2021

Autorização

Na qualidade de autora, autorizo, exclusivamente para fins acadêmicos e científicos, a reprodução parcial ou total desta dissertação por processos fotocopiadores ou eletrônicos.

Assinatura:

São Paulo, 30 de julho de 2021.

e-mail: elainecoliveira.mail@gmail.com

Currículo Lattes: http://lattes.cnpq.br/8632168898360848

OLIVEIRA, Elaine Cristina de.

A linguagem verbal das TED Talks: uma análise multidimensional / Elaine Cristina de Oliveira. - São Paulo: 2021.

pp. XVII + 170.

Orientador: Professor Doutor Antonio Paulo Berber Sardinha.

Dissertação (Mestrado em Linguística Aplicada e Estudo da Linguagem) – Pontifícia Universidade Católica de São Paulo, Programa de Pós-Graduação em Linguística Aplicada e Estudo da Linguagem, 2021.

Área de concentração: Linguística Aplicada e Estudos de Linguagem.

1. Linguística de Corpus. 2. Análise Multidimensional.

Elaine Cristina de Oliveira

A linguagem verbal das TED Talks: uma análise multidimensional

Aprovada em: ____/____/____

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

Banca Examinadora:

_____

Prof. Dr. Antonio Paulo Berber Sardinha – Orientador

_____

Profa. Dra. Sandra Madureira

_____

Profa. Dra. Simone Vieira Resende

A Deus, por seu amor incondicional.

À minha mãe Vilma e ao meu irmão Fernando, pelo apoio, presença, exemplo e cuidado, que são os mais belos atos de amor.

Agradecimento à CAPES

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Programa:
Número do Processo: 88887.372160/2019-00
Período: 01/08/2019 a 31/07/2021
Instituição: Pontifícia Universidade Católica de São Paulo

Agradecimentos

Primeiramente, gostaria de agradecer a Deus por essa grande oportunidade em fazer um curso de mestrado na renomada Pontifícia Universidade Católica de São Paulo, ainda mais por poder estudar uma grande paixão de minha vida que é a língua inglesa. Agradeço muito a Ele por ter sido contemplada com a bolsa de estudos da CAPES, sem a qual essa conquista não teria sido possível. E também, agradeço por ter colocado pessoas tão importantes e especiais em todo esse meu percurso acadêmico.

Agradeço pela Profa. Dra. Simone Vieira Resende (sou sua fã!), pessoa crucial não somente na descoberta do mundo da tradução, que se tornou uma nova paixão, como na descoberta do mundo multidimensional nos estudos da linguagem do Prof. Dr. Tony Berber Sardinha. Agradeço por tê-lo tido como orientador, me possibilitando conhecer uma mente tão brilhante e instigante. Agradeço pela Profa. Dra. Sandra Madureira, que gentilmente aceitou ser parte da minha banca de qualificação. E também agradeço pela amizade da Maria Lúcia Reis, funcionária dedicada não somente ao seu trabalho mas a nós alunos do LAEL.

Agradeço por todos os integrantes do grupo de estudos GELC, sem exceção. Agradeço pela acolhida, pelo apoio, pela amizade, pela paciência, pela ajuda, pela generosidade, pelas mensagens, pelo ombro amigo, enfim, agradeço por toda essa experiência inesquecível ao lado deles nesses dois anos de mestrado.

Agradeço pelo apoio que recebi de minha mãe Vilma e de meu irmão Fernando, pessoas que não somente amo como admiro muito.

Por fim, agradeço por ter me dado saúde para concluir mais essa jornada.

Obrigada! Amém!

“Tudo tem o seu tempo determinado, e há tempo para todo o propósito debaixo do céu.”

Eclesiastes 3:1

Resumo

O objetivo da presente pesquisa foi analisar a linguagem verbal dos vídeos popularmente chamados de TED Talks. Para isso, foram consideradas transcrições de 3.411 vídeos TED Talks em inglês dos anos 1984 até o final de 2019, coletados via site oficial da TED – formando o corpus chamado CoTED. A escolha do objeto de pesquisa se deu porque, apesar de encontrarmos pessoas fazendo referências, usando, estudando, pesquisando, analisando e coletando a linguagem verbal das TED Talks, ainda não existem pesquisas sobre essa linguagem sob a perspectiva da Análise Multidimensional Funcional (AMD) (BIBER, 1988; 2009) – que se baseia na pesquisa de corpora utilizando ferramentas computacionais especializadas. A AMD pode nos fornecer os parâmetros subjacentes das características linguísticas presentes em um corpus ou corpora, ou seja, identifica quais características gramatico-funcionais estão coocorrendo nos textos que fazem parte de um determinado corpus ou corpora. Com a AMD Funcional Aditiva, os textos do CoTED foram mensurados segundo as dimensões da língua inglesa encontradas por Biber (1988); e, com a AMD Funcional Completa, foi feita a extração fatorial completa do CoTED, com o intuito de se revelar quais são as dimensões de variação gramático-funcionais presentes no corpus. Também, por meio do procedimento ANOVA de Modelo Linear Geral (*General Linear Model* – GLM), foi possível verificar se/e como ocorre a variação multidimensional funcional em termos das variáveis independentes "apresentador" e "evento". Os resultados da Análise Multidimensional Aditiva e Completa demonstraram que, por exemplo, a linguagem verbal das TED Talks é adaptável tanto para a linguagem falada quanto para a linguagem escrita, e que fatores externos a ela exercem grande influência – trazendo a hipótese de podermos considerar as TED Talks como um registro híbrido. Deste modo, ao entender o funcionamento da linguagem usada pelas TED Talks, espera-se contribuir para que possamos conhecer de modo detalhado o funcionamento da linguagem verbal dessa influente modalidade de comunicação contemporânea, e contribuir de forma direta ou indireta com futuras pesquisas e possíveis aplicações.

Palavras-chave: TED Talks, linguagem verbal, Linguística de Corpus, Análise Multidimensional.

Abstract

The aim of this research was to analyze the verbal language of the popularly called TED Talks videos. For this, we considered the transcripts of 3,411 TED Talks videos in English from 1984 to the end of 2019, collected via TED's official website – forming the corpus called CoTED. The choice of the object of research occurred because, although we find people making references, using, studying, researching, analyzing and collecting the verbal language of the TED Talks, there are still no research on this language from the perspective of the Multidimensional Functional Analysis (MDA) (BIBER, 1988; 2009) – which is based on corpora research using specialized computational tools. The MDA can provide us the underlying parameters of the linguistic characteristics present in a corpus or corpora, that is, it identifies which grammatical-functional characteristics are co-occurring in texts that are part of a particular corpus or corpora. Considering the Additive Functional MDA, the CoTED texts were measured according to the dimensions of the English language found by Biber (1988); and, with the Complete Functional MDA, a complete factor extraction of the CoTED was performed, in order to reveal the dimensions of grammatical-functional variation present in the corpus. Also, through the General Linear Model (GLM) ANOVA procedure, it was possible to verify whether/and how the multidimensional functional variation occurs in terms of the independent variables "presenter" and "event". The results of the Additive and Complete Multidimensional Analysis demonstrated that, for example, the verbal language of the TED Talks is adaptable to both spoken and written language, and that factors external to it exert great influence – bringing the hypothesis that we can consider the TED Talks as a hybrid register. Thus, by understanding the functioning of the language used by the TED Talks, it is expected to contribute to the understanding in detail of the functioning of the verbal language of this influential modality of contemporary communication and contribute directly or indirectly to future studies and possible applications.

Keywords: TED Talks, Verbal Language, Corpus Linguistics, Multidimensional Analysis.

Lista de tabelas

Lista de quadros

Lista de figuras

Sumário

1. Introdução

Fundada em 1984, nos Estados Unidos, pelo arquiteto e designer gráfico Richard Saul Wurman, juntamente com o designer de televisão Harry Marks, TED[1] surgiu como uma conferência de um dia sobre tecnologia, entretenimento e design – formando o acrônimo TED (*Technology, Entertainment and Design*). Na época, foram convidadas pessoas influentes de tais áreas para participar como palestrantes, cujos ouvintes também precisavam ser exclusivamente convidados. Contudo, apesar de ter surgido no início dos anos 1980, a ideia TED deslanchou somente em 1990. A partir de então, a conferência se tornou um evento anual (sempre pago) ocorrendo em Monterey, na Califórnia (EUA), atraindo um público crescente e influente das mais variadas áreas do conhecimento.

Atualmente, TED é uma organização sem fins lucrativos dirigida por Chris Anderson (desde 2001), adquirida por meio de sua organização sem fins lucrativos *The Sapling Foundation*[2], cujo moto era "*fostering the spread of great ideas*", ou seja, "promovendo a disseminação de grandes ideias"; que originou o moto original da TED "*ideas worth spreading*", ou seja, "ideias que merecem ser espalhadas/disseminadas". Ao explorar a programação dos eventos, é possível perceber que, com o passar dos anos, as conferências passaram a ter aspectos cada vez mais distintos das conferências consideradas tradicionais. Ademais, o tempo estipulado, os temas, os ensaios, o uso de tecnologias etc. passaram a ser uma marca registrada da TED. Desta forma, é perceptível perceber que os eventos TED se tornaram cada vez maiores e até globais – com o TEDGlobal[3], a partir de 2005; e o TEDx[4], a partir de 2009 –, retrato do crescente interesse apresentado por muitas pessoas ao redor do mundo em quererem participar tanto como palestrante quanto como ouvinte.

Todavia, talvez, o grande salto da TED tenha sido a globalidade ou a acessibilidade dos seus vídeos on-line (de forma gratuita). Inicialmente, Anderson (2016, p. 184-185) pretendia disponibilizá-los via televisão, mas não houve aceitação por parte dos produtores de programas de TV. Porém, devido à crescente explosão da internet, além do lançamento do YouTube[5] (em 2005) – com a opção de se assistir a vídeos on-line –, Chris Anderson e sua equipe tiveram a ideia considerada por eles radical de também postar os vídeos TED na internet.

[1] https://www.ted.com/about/our-organization/history-of-ted
[2] https://www.ted.com/about/our-organization; https://www.ted.com/about/our-organization/how-ted-works
[3] https://www.ted.com/attend/conferences/tedglobal
[4] https://www.ted.com/participate/organize-a-local-tedx-event/before-you-start/what-is-a-tedx-event
[5] https://pt.wikipedia.org/wiki/YouTube / https://www.youtube.com/watch?v=SWjBd0yWqeg

Dia 22 de junho de 2006, foram postados seis vídeos em seu site oficial[6]. De 1.000 visualizações, rapidamente passaram a ter 10.000 e, em três meses, chegaram a 1 milhão de visualizações. Com tal resultado, em março de 2007, passaram a disponibilizar mais cem vídeos em seu site (muitos de seus vídeos também estão disponíveis no próprio YouTube) e não pararam mais, sobretudo, considerando todo o retorno recebido. Contudo, é importante ressaltar que nem todas as apresentações feitas nos eventos TED são postadas on-line, pois todos os vídeos passam por um crivo de qualidade. Também, como outro episódio relevante, a partir de março de 2012, surgiu o TED-Ed[7], que traz vídeos educacionais animados feitos segundo o estilo TED[8], como ferramenta de auxílio para professores e alunos – tais vídeos estão inclusos no site oficial da TED juntamente com os demais vídeos TED.

Desta forma, a linguagem analisada neste trabalho é a linguagem verbal em uso dos vídeos popularmente chamados de TED Talks. Para isso, foram consideradas as transcrições das falas dos vídeos TED – em inglês – dos anos 1984 até o final de 2019, coletados via o site oficial da TED[9]. É importante ressaltar que, como o foco está na linguagem verbal, alguns dos vídeos coletados foram desconsiderados – principalmente os que eram exclusivamente de performance artística. Porém, ainda restaram 3.411 vídeos para serem utilizados na análise, inclusive os vídeos que contêm a combinação de palestras com performances artísticas.

Na área da pesquisa científica da Linguística Aplicada (LA) sobre a linguagem verbal das TED Talks, geralmente, encontramos um ou poucos aspectos analisados por vez – sendo eles, muitas vezes, aspectos tradicionais como o ensino de idiomas. Por tal motivo, a proposta desta presente pesquisa é de trazer um estudo mais abrangente dentro dos estudos de LA ao analisar um grande número de vídeos TED, de forma a entender e enxergar como as pessoas se comunicam e compartilham suas ideias por meio da linguagem verbal TED.

O conceito de LA adotada nesta pesquisa não é como uma disciplina inserida na Linguística, mas sim como "uma ponte [ou pontes] com tráfego em dois sentidos, uma encruzilhada" (CELANI, 1998, p. 116). Isto posto, a presente pesquisa foi feita sob a perspectiva da Linguística de Corpus (LC) (BERBER SARDINHA, 2004; BIBER, CONRAD, REPPEN, 1998) e da Análise Multidimensional (AMD) de Biber (1988). a LC "trabalha dentro de um quadro conceitual formado por uma abordagem empirista e uma visão da linguagem como sistema probabilístico" o qual "pressupõe que, embora muitos traços linguísticos sejam

[6] https://www.ted.com/playlists/168/the_first_6_ted_talks_ever
[7] https://ed.ted.com/
[8] https://ed.ted.com/educator?user_by_click=educator
[9] https://www.ted.com/talks

possíveis teoricamente, não ocorrem com a mesma frequência", ou seja, "as possibilidades da estrutura não se realizam todas com a mesma frequência", e "[o] mais importante da diferença de frequências entre os traços é não serem aleatórias" – em outras palavras, "a variação não é aleatória". Além disso, "há uma correlação entre as características linguísticas e situacionais (os contextos de uso)" (BERBER SARDINHA, 2004, p. 30-31). E para poder estudar essa variação não aleatória da linguagem verbal das TED Talks, foi utilizada a abordagem da AMD Funcional nas versões Aditiva e Completa (BIBER, 1988, 2009; BERBER SARDINHA, 2019) – que se baseiam na pesquisa de corpora utilizando ferramentas computacionais especializadas.

Concentrando-se na área da LC, também existem estudos bastante válidos sobre as TED Talks, visando principalmente ao ensino da língua inglesa, à análise do discurso ou ao estudo da prosódia (seção 2.1). Contudo, dentro do arcabouço da AMD, ainda não existem pesquisas sobre a linguagem verbal das TED Talks. Deste modo, as pesquisas precursoras da área servem como fontes não somente teóricas como também de inspiração na presente pesquisa. Diante disso, o primeiro pesquisador em AMD a ser mencionado é Douglas Biber (1988), cujo trabalho pioneiro nos trouxe as dimensões de variação linguística da língua inglesa – que são a base desta pesquisa.

Isto posto, entender o funcionamento da linguagem usada pelas TED Talks significa entender a criação de um modo contemporâneo de pensar o mundo que, além de disseminar ideias, se multiplicou, criou comportamentos e meios de agir. E por ser considerado como um molde em uma nova onda de influenciadores, existe sim uma lacuna nos estudos sobre sua linguagem verbal. Há, portanto, uma necessidade premente de sabermos quais são as características linguísticas – gramático-funcionais – usadas, e como essas características coocorrem para criar o discurso típico das TED Talks. Para tal empreitada, foram elencados três objetivos principais: 1) Verificar se as TED Talks podem ser classificadas como um registro distinto composto por três sub-registros: TED Tradicional, TEDx e TED-Ed; 2) Comparar as TED Talks e seus três sub-registros com os parâmetros de variação da língua inglesa encontrados por Biber (1988); e 3) Verificar se há variação na linguagem verbal das TED Talks e, se houver, quais são os parâmetros que movem tal variação.

Para alcançar tais objetivos, foram levantadas três perguntas: 1) Como o corpus das TED Talks (CoTED) se encaixa nas dimensões de variação da língua inglesa encontradas por Biber (1988)? 2) Quais são as dimensões de variação do corpus das TED Talks (CoTED) sob a perspectiva da AMD Funcional Completa? 3) Como se dá a variação multidimensional

funcional em termos das variáveis independentes "apresentador" e "evento"[10] do corpus das TED Talks (CoTED)?

Em síntese, a presente dissertação está organizada em cinco seções, a seguir. Na primeira seção, temos um apanhado geral sobre a fundamentação teórica que embasa a presente pesquisa. Nela, é feito um breve panorama sobre alguns estudos previamente feitos sobre as TED Talks. Também, é feito um sucinto panorama histórico sobre as TED Talks, além de ser discutida a escolha e definição desse objeto de estudo. Do mesmo modo, são apresentados um breve panorama histórico e os conceitos nucleares da Linguística de Corpus e dos pressupostos da Análise Multidimensional de Biber (1988). Na segunda seção, são discutidos os procedimentos metodológicos adotados, abordando as etapas de planejamento, compilação e organização do corpus CoTED, bem como as etapas do processamento de dados, segundo a premissa da AMD. Na quarta seção, são apresentados os resultados e suas análises correspondentes, isto é, são apresentadas as interpretações dos dados obtidos com os resultados das análises multidimensionais funcionais da linguagem verbal das TED Talks. Por fim, na quinta e última seção, são apresentadas as considerações finais sobre a presente pesquisa.

2. Fundamentação Teórica

O foco desta pesquisa encontra-se em uma das vertentes da Linguística Aplicada (LA), no caso, a Linguística de Corpus (LC) (BERBER SARDINHA, 2004; BIBER, CONRAD, REPPEN, 1998), sob a tutela da abordagem metodológica da Análise Multidimensional (BIBER, 1988, 2009; BERBER SARDINHA, VEIRANO PINTO, 2019). Para tanto, foi feito um apanhado geral sobre a fundamentação teórica por trás de toda a pesquisa por meio da conceituação e descrição do alicerce teórico e metodológico aqui empregados. Desta forma, a seguir, serão apresentadas as seguintes subseções: a primeira traz um panorama sobre alguns estudos previamente feitos sobre as TED Talks; a segunda traz um breve panorama histórico sobre as TED Talks; a terceira traz a estrutura encontrada e considerada das TED Talks; a quarta traz uma breve discussão sobre a escolha e definição do objeto de estudo, que é a linguagem verbal das TED Talks; a quinta traz um breve panorama histórico e os conceitos nucleares da Linguística de Corpus (LC); a sexta fala sobre a tipologia e design de corpus; a sétima traz algumas considerações adicionais sobre a LC; a oitava traz a contextualização e o embasamento

[10] As variáveis independentes "apresentador" correspondem aos 2.845 apresentadores/ autores dos vídeos/textos das TED Talks analisadas; e "evento" correspondem aos 412 títulos de eventos TED encontrados (seção 3.1.1).

teórico da Análise Multidimensional (AMD); a nona fala sobre a análise fatorial; a décima fala sobre o cálculo estatístico univariado (ANOVA); a décima primeira fala sobre a interpretação de fatores; e a última, a décima segunda subseção, traz algumas considerações adicionais sobre a AMD.

2.1 TED Talks – Estudos da linguagem

Conforme previamente dito, concentrando-se na área da LC, existem estudos bastante válidos sobre as TED Talks, mas que visam principalmente ao ensino da língua inglesa, à análise do discurso ou ao estudo da prosódia. No Brasil, temos alguns exemplos como: 1) Silva *et al.* (2018) – com o intuito de trazer diferentes modelos pedagógicos na criação de atividades didáticas, com a utilização de corpora no processo de ensino-aprendizagem de línguas, Silva *et al.* (2018) criaram atividades em inglês baseadas em um corpus oral – utilizando 127 TED Talks – que foram posteriormente aplicadas em um minicurso de 12 horas, ministrado em uma faculdade de tecnologia no noroeste do estado de São Paulo. 2) Silva (2017) selecionou 10 TED Talks para a elaboração de atividades didáticas de ensino da língua inglesa voltadas a estudantes de cursos técnicos e de cursos superiores que possuem a disciplina de ESP (*English for Specific Purposes* – Inglês para Fins Específicos) de forma a buscar evidências da importância do ensino de vocabulário de áreas específicas, aliado ao ensino de estratégias que desenvolvam a percepção oral dos estudantes; e as características presentes nos vídeos TED seriam fontes úteis quanto a isso. 3) Sabota e Almeida Filho (2017) enfocaram no uso das TED Talks como uma das dez ferramentas tecnológicas classificadas como potenciais mediadoras no processo de aprimoramento da competência teórica do professor de idiomas – pesquisa que foi aplicada em um curso de extensão, cujo público-alvo era professores de inglês em formação, no caso, alunos de uma universidade pública no Estado de Goiás. 4) Miyamoto (2017) trouxe uma discussão sobre o uso de novas tecnologias em aulas de inglês como LE (língua estrangeira) em cursos tecnológicos, mencionando as palestras TED Talks como um desses recursos. Segundo a autora, tais recursos se demonstraram efetivos como auxiliadores no desenvolvimento de habilidades e competências dos estudantes. 5) Sátiro e Silva (2018) propuseram o uso das TED Talks em aulas de Física, aplicando atividades em aulas de um curso pré-vestibular ministrado na Universidade do Ceará. Segundo os autores, foi possível observar o aumento no comparecimento e do engajamento dos alunos com tais aulas. 6) Franco Silva *et al.* (2020) utilizaram as transcrições de 60 TED Talks relacionadas à área da engenharia de forma a elaborar atividades que utilizam *lexical bundles* (pacotes lexicais), de forma a contribuir

na aprendizagem do idioma inglês por alunos de graduação da UNESP. De acordo com os autores, os pacotes lexicais de fato ajudaram os alunos a lembrar e entender melhor os textos estudados em comparação com o ensino de palavras individuais. 7) Miranda (2016), por sua vez, ao analisar 3 palestras, argumenta que as TED Talks podem ser classificadas como um novo gênero discursivo. Segundo ela, as TED Talks são fruto da sociedade contemporânea cujos traços apontam para o surgimento de um novo gênero do discurso e que, ao mesmo tempo, instigam uma reflexão a respeito da possibilidade de pensá-las como um hipergênero[11].

No contexto internacional, temos exemplos como: Rousseau *et al.* (2012), que coletaram um corpus (TED-LIUM) de 774 TED Talks (118 horas de transcrições) por meio de um sistema de reconhecimento de falas (*Automatic Speech Recognition* – ASR) desenvolvido por eles pelo programa chamado LIUM. Tal sistema é baseado na iteração de forma a aprimorar o alinhamento entre os dados do áudio com o texto das legendas. Segundo os autores, os dados coletados tiveram WER (*word error rate* – taxa de erro de palavra) de 17,4%. O intuito dos pesquisadores foi o de disponibilizar um corpus desenvolvido como uma ferramenta que fosse útil para o público em geral – atualmente já existe o TED-LIUM corpus release 3[12], com 2.351 TED Talks.

Chang e Huang (2015) examinaram a estrutura retórica de 58 TED Talks, de forma a explorar a possibilidade de incorporar as palestras TED como um recurso pedagógico para auxiliar nas apresentações orais de alunos de língua inglesa como língua estrangeira. De acordo com os autores, foram identificados sete movimentos nas TED Talks, com seus respectivos passos: 1) Direcionamento ao ouvinte – cumprimentar o público e buscar engajamento por meio do discurso meta-nível (falar do evento e das pessoas envolvidas); 2) Introdução sobre o assunto (obrigatório) – definir o cenário com algumas informações extras de contextualização, introduzir o tópico e delimitar os assuntos que serão tratados dentro do tópico; 3) Apresentação do palestrante – falar sobre sua história, formação, conhecimento e competência na área discutida e seu posicionamento a respeito do tópico apresentado; 4) Desenvolvimento do tópico (obrigatório) – apresentar argumentos e explicações além de descrever os processos ou eventos envolvidos; 5) Encerramento (obrigatório) – sinalizar de modo a preparar o público para o término da apresentação, trazendo um resumo ou respondendo às questões anteriormente levantadas; 6) Mensagem final (obrigatório) – convidar os ouvintes à ação ou fazer generalizações e especulações sobre as conclusões do tópico discutido; 7) Agradecimentos

[11] Hipergênero é definido como um gênero maior, formado por um agrupamento ordenado (conjunto) de gêneros, compondo uma macrounidade discursivo-textual. (BONINI, 2003).
[12] https://lium.univ-lemans.fr/en/ted-lium3/

(obrigatório) – agradecer ao público e pela oportunidade de se apresentar em um evento TED. Os autores ressaltam que, apesar de encontrarmos padrões entre os movimentos encontrados, existe uma flexibilidade em sua ordem e estrutura. Segundo eles, tais padrões retratam traços de reprodução ou adaptação de outros gêneros[13] como discursos de formatura, apresentações em conferências, discursos políticos e apresentações comerciais. Contudo, os autores afirmam que o cenário e o propósito característicos das TED Talks são o que as diferenciam dos demais gêneros. Assim sendo, segundo eles, as TED Talks apresentam uma natureza heterogênea, podendo ser resultado de sua missão proposta de "informar, inspirar, surpreender e encantar" seus ouvintes (CHANG; HUANG, p. 50, 2015). Por conclusão, os autores afirmam que seus achados podem sim auxiliar os professores com seus alunos no aprimoramento de suas apresentações orais, sendo um bom recurso pedagógico.

Por meio do projeto *The Pisa Audio-visual Corpus Project*[14], Camiciottoli e Bonsignori (2015) propuseram coletar um corpus audiovisual de vídeos em inglês de variados gêneros que representam, segundo elas, determinada relevância para alunos ESP (*English for Specific Purposes* – Inglês para Fins Específicos). Cada gênero foi classificado em uma escala que vai desde os mais autênticos, monológicos, formais e científicos até os mais ficcionais, popularizados, informais e conversacionais. Em tal escala, as TED Talks foram classificadas juntamente com as palestras acadêmicas de disciplinas específicas, que se encontram no lado da escala entre os mais autênticos, monológicos, formais e científicos. Porém, as autoras destacam que alguns gêneros apresentam diferentes colocações na escala, e as TED Talks são um exemplo disso, pois são descritos como discursos de curta duração realizados por especialistas de variadas áreas do conhecimento e em um formato popularizado. Desta forma, diante de sua pesquisa proposta, Camiciottoli e Bonsignori (2015) afirmam que as TED Talks são reconhecidamente um recurso importante em contextos instrucionais, principalmente no campo do ensino de línguas.

Caliendo e Compagnone (2014) compararam um corpus de 207 TED Talks (TED_ac) com um corpus de 35 palestras proferidas na Universidade de Michigan (MICASE_lect) a fim

[13] O termo adotado neste trabalho será "registro". Registros, por sua vez, são variedades de texto definidas situacionalmente, isto é, por meio do contexto em que ocorrem na sociedade. Eles podem ser variedades amplas, como "escrita acadêmica" , ou específicas, como "editoriais de jornal". Em ambos os casos, o contexto de uso é o fator predominante em sua identificação: escrita acadêmica é um conjunto de práticas de produção textual pertinentes ao contexto da academia (universidade, ciência, saber etc.), chamado e conhecido como tal por seus pares; e editoriais são escritos e lidos como tais pelos seus usuários (jornalistas, editores, leitores etc.). O que diferencia os dois registros é sua abrangência; escrita acadêmica engloba vários registros distintos, como o artigo, a resenha e a tese, enquanto editorial é um rótulo que se aplica somente a esse registro específico (BERBER SARDINHA, 2013 p. 55).

[14] Apesar de o artigo ter sido publicado, não foi encontrado o corpus do projeto.

de investigar a maneira como os acadêmicos transmitem uma postura epistêmica, ou seja, a maneira como transmitem uma imagem de especialistas no assunto tratado. De acordo com os autores, em ambos os casos, existem uma recorrência dos verbos lexicais epistêmicos – como *see, show, know, think* – e de pronomes de primeira e de segunda pessoas. Desta forma, Caliendo e Compagnone (2014) explicam que, de fato, as palestras TED também possuem um caráter informativo, assim como as palestras de universidades. Porém, segundo eles, o que as diferencia das demais palestras é o fato de as palestras TED exercerem um papel de espaço pragmático alternativo, onde acadêmicos constroem sua imagem ao dar ênfase na sua afiliação a uma comunidade de especialistas e ao promover suas pesquisas, e seus resultados, como algo tangível e extremamente confiável.

Focando na questão do discurso utilizado em contextos autênticos, Ratanakul (2017) analisou 50 TED Talks, por meio da Análise dos Movimentos Retóricos, de forma a contribuir na concepção de materiais e práticas em sala de aula para cursos de apresentação oral e outros eventos relacionados. A autora identificou três estágios nas apresentações orais TED: 1) A abertura; 2) O corpo; e 3) O encerramento; e que cada estágio possui movimentos secundários (que totalizam em 35 e que podem ser compartilhados), como introdução (abertura), informações básicas (corpo), tese (abertura e corpo), motivo do problema (corpo), solução (corpo), avaliação positiva da sugestão de solução (corpo e encerramento), agradecimentos (encerramento) etc. Na fase substancial de cada palestra TED, o corpo, quatro movimentos são empregados para transmitir a mensagem: 1) O movimento situação – informações de fundo/base iniciais (8,16% do corpo); 2) O movimento problema – obstáculos, necessidades, restrições, empecilhos e dilemas (44,27% do corpo); 3) O movimento resposta – soluções para o problema (35,92% do corpo), e 4) O movimento avaliação – consequências positivas e negativas e vantagens e desvantagens das soluções propostas (11,65% do corpo). Segundo a autora, os resultados encontrados em sua pesquisa demonstram que a natureza das características dos movimentos que os falantes (apresentadores TED) empregam são dirigidas por um propósito e vinculados a um contexto, e tais movimentos, além de suas características, podem ser reconstruídos e ensinados.

Por meio de comentários coletados sobre 405 TED Talks, Tsou *et al.* (2014) analisaram a reação do público referente às características dos apresentadores e sua relação com as plataformas TED e YouTube. De acordo com os autores, os comentários sobre as características dos apresentadores eram mais propensos a aparecerem no YouTube, ao passo que, os comentários sobre o conteúdo dos vídeos eram mais recorrentes na TED. Ademais, eles

também ressaltaram que as pessoas tendiam a fazer comentários mais emotivos (das formas positivas e negativas) quando era um apresentador do sexo feminino. Ainda assim, segundo eles, isso não anula o fato de ambas as plataformas serem veículos de divulgação muito válidos na disseminação de ideias propostas pela TED.

Tsai (2015) coletou 391 TED Talks, com apresentadores do sexo masculino, para analisar quais são as características que separam um palestrante TED de outros palestrantes – no caso, professores universitários – de forma a identificar e entender as diferenças prosódicas entre ambos os grupos. Segundo o autor, as TED Talks são o ápice do falar em público, pois conseguem comunicar um conteúdo atraente de forma impecável, demonstrando grande popularidade atestada pelas milhões de visualizações de seus vídeos. Ao utilizar o modelo discriminativo (*pitch* e *energy* – frequência e energia), os áudios foram classificados como pertencentes a uma palestra TED ou a uma palestra regular feita por um professor. Para os áudios de 5 minutos, foi possível fazer uma previsão de porcentagem <10% de erros. Para os áudios de 5 segundos, a porcentagem era <25% de erros. Com os dados obtidos, Tsai (2015) define as diferenças prosódicas mais marcantes para os falantes TED: as palestras são mais compactas (ou seja, menos espaço e silêncio), falam com uma voz mais densa e têm um fluxo de entrega mais consistente. O autor comenta que essas diferenças podem advir do resultado da diferença entre uma palestra longa e uma palestra mais curta, mas ressalta que, cumprir as características acima não é tarefa fácil. Segundo ele, um palestrante que passa todo o seu tempo em um discurso de alta energia, enquanto mantém uma entrega consistente, é certamente um orador que está muito bem preparado, pois tem algo a dizer e sabe como dizê-lo. Para Tsai (2015), talvez, seja essa a parte fundamental do que faz com que os apresentadores TED se destaquem.

Hasebe (2015) criou uma plataforma chamada *TED Corpus Search Engine* (TCSE)[15] com um corpus de todos os vídeos das TED Talks – sendo, segundo ele, frequentemente atualizado. A plataforma traz a possibilidade de analisar os n-gramas[16] de cada vídeo, ter acesso às listas de palavras, palavras-chave, nuvens de palavras e marcadores de discurso. Seu objetivo é de disponibilizar um corpus composto de exemplos retirados da linguagem em uso, sem que tenhamos que nos basear em exemplos meramente inventados. Entretanto, não é possível

[15] https://tcse.gitbook.io/doc/ / https://yohasebe.com/tcse/

[16] Um n-grama ´e uma sequência de n itens dentro de uma frase. Os itens podem ser palavras, letras, sílabas, classificação gramatical das palavras, ou qualquer outra base. Um n-grama de tamanho 1 ´e chamado de unigrama, de tamanho de 2, de bigrama, de tamanho 3 ´e chamado de trigrama, de 4 em diante ´e n-grama. Para uma sequência de palavras, por exemplo ”Ações da Petrobras sobem”, um bigrama de palavras seria: ”# Ações”, ”Ações da”, ”da Petrobras”, ”Petrobras sobem”, ”sobem #”. (VILELA, 2011, p. 23).

analisar todos os vídeos juntos ou em conjuntos, como pretendido neste projeto, mas sim em separado.

Com uma proposta similar ao de Hasebe (2015), Raine (2019) criou um corpus de 2.051 vídeos TED – o *Talk Corpus*[17] – para auxiliar professores e alunos de língua inglesa. Seus objetivos são de classificar os vídeos TED coletados de acordo com o nível de dificuldade, além de fornecer dados complementares como velocidade da fala, lista de palavras e n-gramas, de forma a facilitar a compreensão da relação entre os colocados[18] – considerado como necessária para o desenvolvimento da habilidade de se compreender e usar a língua inglesa de forma mais natural possível. Porém, nesse caso, também não é possível analisar todos os vídeos juntos ou em conjuntos, como pretendido neste projeto.

Tanveer *et al.* (2019) coletaram transcrições de 2.233 vídeos TED e de mais de 5 milhões de avaliações do público – as classificações feitas pelo público são 14, e na seguinte ordem (figura 1): bonito, confuso, audaz, fascinante, engraçado, informativo, genial, inspirador, "de cair o queixo", enfadonho, desagradável, OK, persuasivo e não convincente – e propuseram um modelo que prevê tais avaliações por meio da comparação entre arquiteturas de redes neurais e o Aprendizado Estatístico de Máquina. Segundo os pesquisadores, foi possível prever todas as 14 classificações, com um AUC (*Area Under the Curve* – área sob a curva[19]) de 0,83, ao se utilizar somente as transcrições e as características de prosódia – apesar de as características de prosódia não terem influenciado muito na previsão.

[17] https://www.apps4efl.com/tools/talk_corpus/

[18] "Colocado(s) (collocate(s)): palavra(s) que ocorre(m) ao redor do nódulo, em posições relativas (primeira à esquerda, segunda à esquerda). Difere de palavra de contexto porque esta é opcional , definida pelo usuário no momento da busca. Os colocados, contudo, são todas as palavras que ocorrem perto do nódulo, dentro do horizonte especificado, incluindo as palavras de busca que existirem." (BERBER SARDINHA, 2004, p. 188).

[19] A área sob a curva é um resumo estatístico utilizado na determinação da acurácia de um teste, onde o máximo valor possível é 1. (PINHEIRO, 2018).

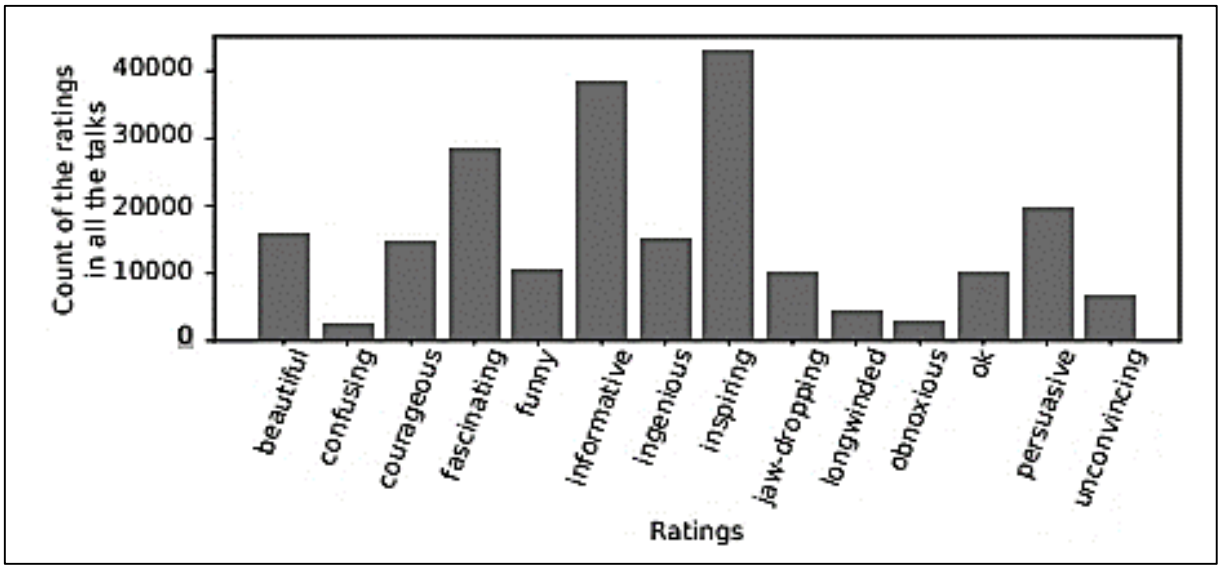


Figura 1: Classificações das TED Talks (Tanveer *et al*., 2019, p. 2).

Correia (2018) coletou 730 TED Talks (em inglês; e que foram reduzidos à 180 até o final de sua pesquisa, formando o corpus METATED) de forma a detectar e classificar automaticamente o uso de metadiscurso em contextos de apresentação oral. O autor traz teorias sobre a vertente oral de metadiscurso, focando numa taxonomia (de 16 categorias) que define os conceitos metadiscursivos de forma funcional – considerando diferentes conjuntos de características lexicais, sintáticas e semânticas –, ou seja, atribui uma função discursiva às ocorrências de metadiscurso em vez de analisar exclusivamente a sua forma. Segundo ele, a análise do desempenho dessa classificação pode ser aplicada como auxílio para tarefas de Processamento de Língua Natural (tais como sumarização e detecção de tópico), ou como parte de um currículo de técnicas de apresentação.

Contudo, apesar do grande interesse dentre os pesquisadores sobre as TED Talks, conforme já visto, ainda não existem pesquisas sobre sua linguagem verbal dentro do arcabouço da AMD. É por tal motivo que esta pesquisa se faz presente e espera contribuir ao trazer uma nova forma de se enxergar esse objeto de estudo.

2.2 TED Talks – Breve histórico

A conferência TED nasceu em 1984, nos Estados Unidos, criada a partir da idealização de Richard Saul Wurman, juntamente com Harry Marks, de que existia uma poderosa convergência entre os três campos: tecnologia, entretenimento e design – formando o acrônimo TED. Na época, eram convidadas somente pessoas influentes de tais áreas para participar como

palestrantes, cujos ouvintes também precisavam ser exclusivamente convidados – observando que o evento é pago. A primeira conferência TED incluiu uma demonstração do CD (disco compacto), do e-book e de gráficos 3D de última geração da Lucasfilm, além da demonstração do matemático Benoit Mandelbrot de como mapear os litorais usando sua teoria em desenvolvimento da geometria fractal. Mas, apesar de um começo grandioso, o evento perdeu dinheiro e somente seis anos depois que Richard e Harry tentaram novamente. A partir de 1990, a Conferência TED tornou-se um evento anual ocorrendo em Monterey, na Califórnia, atraindo um público cada vez maior e influente das mais diversas áreas do conhecimento – como cientistas, filósofos, músicos, líderes religiosos, empresários, filantropos etc. – que quisesse compartilhar uma descoberta considerada por eles empolgante[20]. Atualmente, TED é uma organização sem fins lucrativos dirigida por Chris Anderson (desde 2001), adquirida por meio de sua organização sem fins lucrativos *The Sapling Foundation*[21], cujo moto era "*fostering the spread of great ideas*", ou seja, "promovendo a disseminação de grandes ideias"; que originou o moto original da TED "*ideias worth spreading*", ou seja, "ideias que merecem ser espalhadas/disseminadas".

Segundo Anderson (2016, p. 182), o que pode ter motivado o surgimento dessa união entre tecnologia, entretenimento e design seriam as grandes mudanças tecnológicas que já estavam ocorrendo; além do crescente interesse por parte dos pesquisadores na área da tecnologia em tornar seus produtos mais atraentes e, por parte dos arquitetos, designers e donos de empresas de entretenimento em buscar novas áreas de atuação e compreender os avanços tecnológicos existentes.

Ao conhecer a conferência TED (em 1998), Anderson (2016, p. 181) afirma que tinha uma visão muito diferente do que iria encontrar no evento, pois achava as conferências – ou os ciclos de palestras –, de modo geral, como algo não muito atrativo ou um mal necessário. Porém, essa primeira experiência foi definida por ele como uma "revelação" de algo novo e que merecia sua atenção:

> Quando tive de ir embora, eu já compreendia por que o ciclo de palestras significava tanto para aqueles ali presentes. Estava empolgado com tudo o que havia aprendido. Senti-me tomado por uma sensação de poder demonstrar um "jogo de cintura" que vinha me faltando havia muito tempo. Era como se eu tivesse passado por uma revelação. (ANDERSON, 2016, p. 183).

[20] https://www.ted.com/about/our-organization/history-of-ted / https://www.ted.com/about/conferences
[21] https://www.ted.com/about/our-organization; https://www.ted.com/about/our-organization/how-ted-works

Foi a partir de sua primeira experiência ao assistir a um evento TED que Anderson (2016) afirma ter passado a enxergar o quanto tais palestras significavam para aqueles que participavam da conferência. Assim, quando soube que Rick Wurman queria vender a TED, Anderson decidiu adquiri-la, afirmando ser uma oportunidade de se construir algo maior (ANDERSON, 2016, p. 184).

Contudo, para Anderson (2016, p. 18-19, 182) a TED não seria o que é hoje sem toda a influência de Rick Wurman, inclusive no formato das palestras. O fundador da TED é descrito por ele como a "alma" das conferências e como alguém obcecado com a ideia de tornar acessíveis os conhecimentos "obscuros" por meio da chamada "Arquitetura da Informação", ou seja, ele tinha a habilidade de persuadir os palestrantes a encontrar formas mais atraentes de expor suas ideias para um público que não conhecesse o assunto. Também, Anderson (2016) explica que a impaciência de Rick Wurman foi fator crucial para determinar o tempo cada vez mais limitado das palestras e, além disso, foi Rick Wurman quem eliminou por completo as perguntas feitas pela plateia, alegando ser mais interessante encaixar um novo palestrante do que ouvir alguém fazendo uma pergunta para se promover. Para Anderson (2016, p. 182) isso "pode ter aborrecido alguns, mas para o público em geral a decisão foi uma dádiva. Contribuiu para uma sequência rápida de apresentações".

Na primeira vez em que Anderson promoveu o evento TED solo em 2003 – passando de oitocentas pessoas por ano à setenta –, ele percebeu que não seria fácil substituir a "alma" da TED. Ainda assim, ele declara que quis arriscar, motivado por ter tornado a TED em uma organização sem fins lucrativos, passando uma mensagem de que, quem participa em um evento TED, acaba ajudando a construir uma nova forma de descobrir e partilhar ideias (ANDERSON, 2016, p. 189-190). Todo o esforço de Anderson parece ter sido recompensado, considerando que os eventos se tornaram cada vez maiores e até globais – com o TEDGlobal, a partir de 2005, e o TEDx, a partir de 2009. Além disso, em 2005, foi criado o chamado TED Prize[22], que se tornou no *The Audacious Project*[23] em 2018, cujo objetivo é o de premiar ou financiar algum projeto apresentado em um evento TED. Todavia, por mais que seja uma organização sem fins lucrativos, existe um preço a ser pago para poder participar dos eventos TED. No próprio site da TED estão disponibilizados os preços praticados para cada tipo de evento, de adesão e como o lucro é distribuído[24].

[22] https://www.ted.com/about/programs-initiatives/ted-prize
[23] https://audaciousproject.org/about
[24] https://www.ted.com/attend/conferences/ted-conference / https://www.ted.com/about/conferences

2.3 TED Talks – Estrutura

No site oficial da TED podemos encontrar como são divididos os eventos presenciais[25] atualmente. A primeira divisão é a dos eventos ou conferências TED – eventos realizados pela própria TED, cujo "carro-chefe" são as palestras. Cada evento pode ter por volta de 50 palestras com duração de 1 semana ou até um encontro de algumas horas. Dentre eles, temos: Conferência TED, TEDGlobal, TEDWomen, TEDSummit e demais eventos, como TEDSalons, TED Institute e TED-Ed Weekend (figura 2):

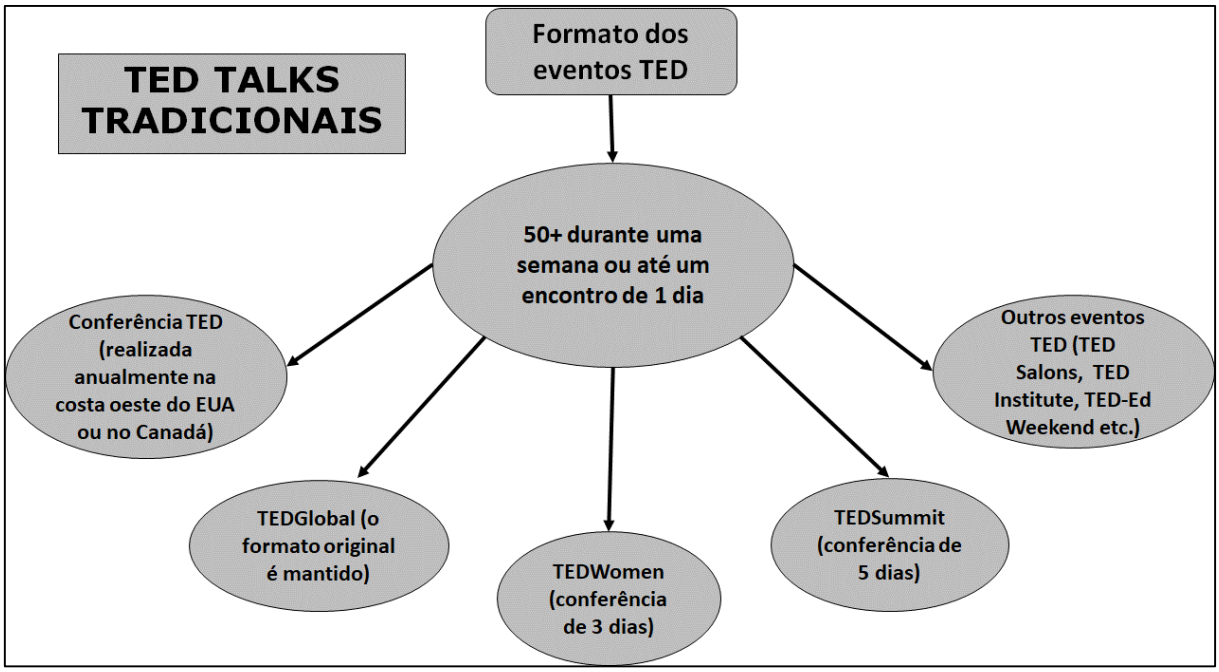


Figura 2: TED Talks tradicionais (elaborado pela autora - Fonte: https://www.ted.com/about/conferences).

A partir de 2003, na direção de Chris Anderson, cada evento TED passou a ter uma temática principal (com subtemas e pequenos eventos paralelos) como *The Future Belongs to Those Who Create It* (O futuro pertence àqueles que o criam – em 2003) e *The Pursuit of Happiness* (Em busca da felicidade – em 2004)[26]. E, a partir de 2005, os eventos passaram a ocorrer até duas vezes ao ano, deixando de serem exclusivamente feitos nos EUA e, cada vez mais, assumindo as características que são hoje conhecidas. O TED2009[27], por exemplo, que

[25] https://www.ted.com/about/conferences / https://www.ted.com/participate/organize-a-local-tedx-event/before-you-start/event-types
[26] https://www.ted.com/about/conferences/past-teds
[27] https://conferences.ted.com/TED2009/program/schedule.php.html

ocorreu em Long Beach (EUA), teve uma programação que incluía não somente as esperadas palestras como também performances artísticas, exibições, entrevistas, excursões, coquetéis, almoços, jantas, festas e até um piquenique; fora a presença e contribuição de patrocinadores.

Uma segunda divisão a ser considerada dos eventos TED presenciais seria a chamada TEDx[28], que são realizados por entidades autônomas em todo o mundo – com duração de até um dia –, cujo "carro-chefe" também são as palestras. Os TEDx são encontros "independentes, mas licenciados e orientados pela organização TED. O 'x' no final, é justamente para indicar que aquele acontecimento é realizado por entidades autônomas em todo o mundo"[29]. A licença TEDx é gratuita, com exceção do TEDx Business[30], e os organizadores são responsáveis pelos custos do evento – podendo procurar por patrocinadores – e por cobrar os ouvintes. Os eventos podem seguir um modelo padrão, ou serem direcionados para contextos específicos, como um ambiente universitário; ou temas específicos, como questões relacionadas a jovens ou mulheres. Segue, logo abaixo, uma lista dos possíveis tipos de eventos TEDx (figura 3):

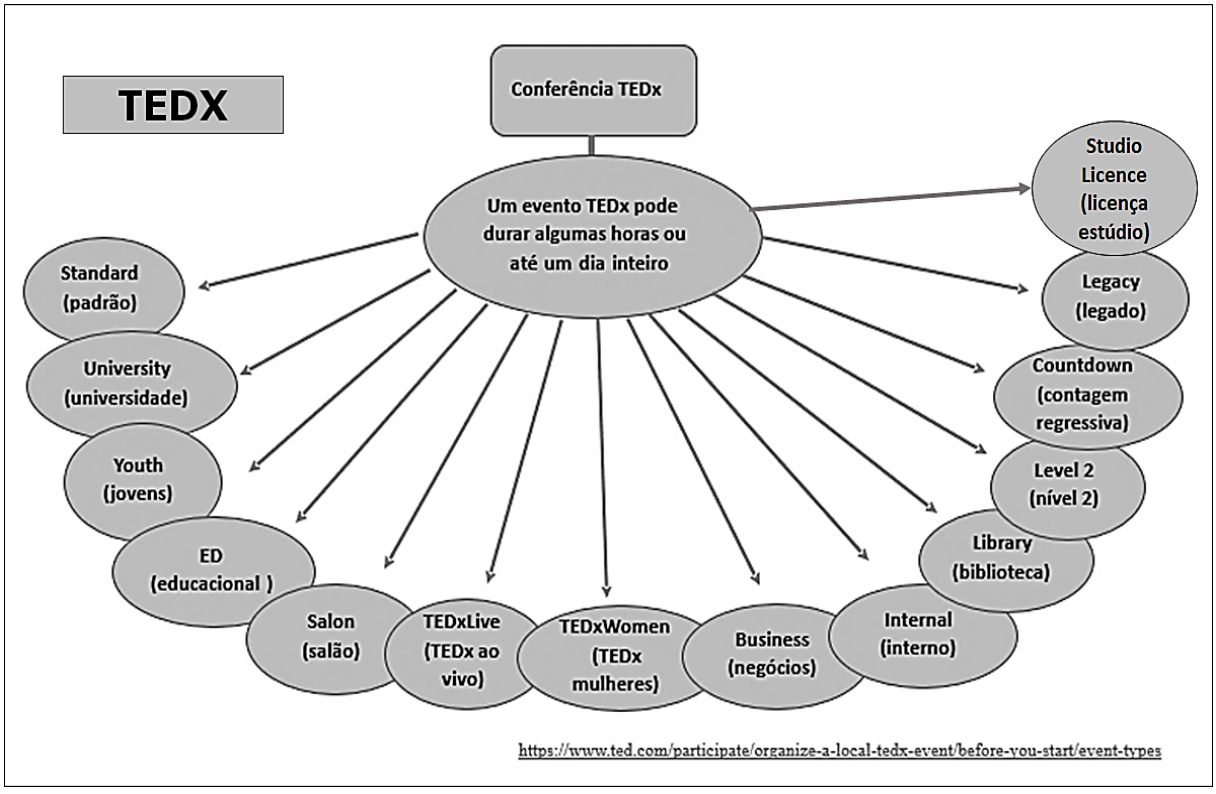


Figura 3: TEDx (elaborado pela autora - Fonte: https://www.ted.com/participate/organize-a-local-tedx-event/before-you-start/event-types).

[28] https://www.ted.com/participate/organize-a-local-tedx-event/before-you-start/what-is-a-tedx-event
[29] https://tedxsaopaulo.com.br/o-que-sao-os-tedx-talks-no-brasil/
[30] https://help.ted.com/hc/en-us/articles/360039157493-Is-there-a-license-fee-for-a-TEDx-event-

Fora dos eventos presenciais, mas que cabe como uma subdivisão das TED Talks, temos os vídeos TED-Ed[31] – vídeos educacionais animados feitos, segundo o estilo TED[32], por palestrantes, professores, designers, pesquisadores, jornalistas etc.–, já que eles são popularmente chamados de TED Talks, como também estão inclusos no site oficial da TED juntamente com os demais vídeos TED tradicionais e TEDx. Deste modo, a divisão dos tipos de vídeos TED aqui considerada é TED Geral – com todos os vídeos TED coletados; TED tradicional; TEDx; e TED-Ed – o que nos traz as três categorias ou (sub-registros) anteriormente citadas:

1. TED Tradicional – evento realizado pela própria TED, cujo "carro-chefe" são as palestras;
2. TEDx – evento realizado por entidades autônomas em todo o mundo, cujo "carro-chefe" são as palestras;
3. TED-Ed – vídeos educacionais animados feitos, segundo o estilo TED, por palestrantes, professores, designers, pesquisadores, jornalistas etc.
4. TED Geral – a soma do TED Tradicional, TEDx e TED-Ed.

De fato, no site oficial da TED, também é possível classificar os vídeos que desejar assistir. Porém, cada vídeo apresenta uma quantidade variada de tópicos (ou *tags*), não sendo possível definir um tópico único ou principal segundo essa forma de classificação. Além disso, não é possível analisar as TED Talks de modo geral ou agrupadas por meio da configuração do site, mas sim uma por vez. Dada essa delimitação, foi feita a coleta das transcrições das TED Talks em inglês e definida uma divisão em quatro grupos de TED Talks – acima descrita – para o presente trabalho: TED Geral, composto por TED tradicional, TEDx e TED-Ed.

2.4 TED Talks – Linguagem verbal

Talvez, o ano de 2005 foi o marco principal na história das TED Talks, pois foi quando começaram a ser disponibilizados os seus vídeos de modo on-line e de forma gratuita. A princípio, Anderson (2016, p. 184-185) pretendia disponibilizar as gravações das apresentações via televisão, mas não houve aceitação por parte dos produtores de programas de TV. Apesar disso, como houve uma crescente explosão da internet na época, possibilitando o lançamento

[31] https://ed.ted.com/
[32] https://ed.ted.com/educator?user_by_click=educator

do YouTube[33] e a opção de se assistir a vídeos on-line, Anderson (2016) e sua equipe resolveram também postar os vídeos TED na internet. Definida como o segundo maior impulso no renascimento da arte de falar em público (ou da oratória), a internet – em particular, os vídeos on-line – contribuiu para que a TED se tornasse um dos pioneiros de uma nova forma de compartilhar conhecimento (ANDERSON, 2016, p. 189-190). Desta forma, dia 22 de junho de 2006, foram postados seis vídeos no site oficial da TED. De 1.000 visualizações, rapidamente passaram a ter 10.000 e, em três meses, chegaram a 1 milhão de visualizações. Considerando todo o retorno recebido (muitos de seus vídeos também estão disponíveis no próprio YouTube), em março de 2007, passaram a disponibilizar mais de 100 vídeos em seu site oficial e não pararam mais (ANDERSON, 2016, p. 191).

Anderson (2016, p. 192-193) explica que a mais profunda implicação dos vídeos on-line é a criação de um ecossistema interativo mundial no qual todos podemos aprender uns com os outros, dando visibilidade para grandes talentos e o incentivo necessário para se melhorar o mundo em que vivemos. Tal fenômeno foi nomeado por ele como *crowd-accelerated innovation* (inovação motivada pelo público), já que a participação do público se tornou muito mais ampla. Para ele, nessa nova e atual era do conhecimento, não basta a tradicional ideia de que precisamos de pessoas cada vez mais especialistas em determinadas áreas do conhecimento, mas sim, de pessoas que saibam aproveitar o que o outro tem para dizer. Esse contexto é definido por ele como o primeiro grande impulso no renascimento da arte de falar em público (ANDERSON, 2016, p. 188).

É interessante notar que, tal conceito de que podemos aprender com as outras pessoas que não atuam no mesmo campo de conhecimento vem de encontro com a questão da transdisciplinaridade da LA. Segundo Celani (1998), a LA identifica, investiga e busca soluções para problemas com relevância social relacionados à linguagem na vida real, dentro ou fora do contexto escolar, cujo caráter transdisciplinar traz a possibilidade de um diálogo ou uma interação entre os diversos saberes (não simplesmente disciplinas ou ramos do conhecimento), independentemente de suas supostas similaridades, perspectivas, objetivos ou objetos de estudo. Assim, ao postular a ideia de que existe a possibilidade de um diálogo ou uma interação entre os diversos saberes (não simplesmente disciplinas ou ramos do conhecimento), independentemente de suas supostas similaridades, perspectivas, objetivos ou objetos de estudo, Celani (1998) explica que a LA procura identificar, investigar e buscar soluções para problemas com relevância social relacionados à linguagem na vida real. Diante

[33] https://www.youtube.com/

disso, as TED Talks também podem definidas como uma forma de se identificar, investigar e buscar por soluções para problemas com relevância social, podendo também ser relacionados à linguagem na vida real. Assim sendo, para Anderson (2016, p. 10 e 185), esse renascimento da arte de falar em público fez as TED Talks como as conhecemos hoje, a que fascina tanto seus espectadores, a que combina várias formas de conhecimento e a que alcança níveis globais. Perante o exposto, não é de se estranhar que o estilo ou formato TED esteja em voga. Isso, tanto que, a própria TED oferece cursos para ensinar o estilo ou formato TED Talks[34].

Também é interessante ver que o falar em público ou a arte da oratória tem sido de grande importância na história norte-americana. Baskerville (1979) descreve a cultura norte-americana como amante e sagaz dominante da oratória, cuja importância é também retratada por vários pesquisadores da área. Segundo ele, tem sido frequentemente afirmado que os americanos têm tradicionalmente demonstrado um gosto pela oratória e, fazendo referências a ensaístas, biógrafos, historiadores e outros pesquisadores, Baskerville (1979, p. 2-3) cita afirmações como: "assim que um bebê ianque pode sentar-se em seu berço, ele chama o berçário para ordenar e começar a se dirigir à assembleia"; "se alguma vez houve um país onde a eloquência era um poder, são os Estados Unidos"; "o povo americano sempre foi admirador ardente de uma oratória genuinamente grandiosa"; "o amor pela oratória é inerente aos americanos"; "em nenhum outro país tem oradores e uma oratória que desempenharam um papel tão evidente na formação de assuntos públicos, como na América" [35].

Diante disso, Baskerville (1979, p. 43) salienta que uma das explicações que pode ser atribuída à essa paixão pela oratória, ou pelo falar em público, seriam as analogias feitas com a Grécia e Roma antigas, pois era considerada uma honra ser comparado aos antigos oradores. E aprender a ser um bom orador era algo comumente ensinado nas escolas. Bower (1943, p. 308) nos conta que, o papel da comunicação oral sempre esteve presente em várias instituições americanas tais como a igreja, o tribunal de justiça, o poder legislativo e empresas comerciais. E, nesse contexto, a instituição maior influenciada é a escolar, principalmente durante o início

[34] Um deles é oferecido para alunos de 6 a 18 anos chamado *TED-Ed Student Talks Program* (https://ed.ted.com/student_talks); e um outro seria o *TED Masterclass*, destinado para o público em geral, empresas e organizações educacionais (https://masterclass.ted.com/).

[35] Original: It has frequently been asserted that Americans have traditionally displayed a keen appetite for oratory. Wendell Phillips contended that as soon as a Yankee baby could sit up in his cradle, he called the nursery to order and proceeded to address the house. "If there ever was a country where eloquence was a power," Emerson exclaimed in one of his lectures on the subject, "it is the United States." "The American people have always been ardent admirers of genuinely great oratory," said Warren C. Shaw in introducing his History of American Oratory in 1928. The theme is reiterated endlessly by essayists, biographers, historians, and especially by anthologists of speeches, who predictably introduce their collections with such statements as: "The love of oratory is inherent in Americans," or "In no other country have orators and oratory played so conspicuous a part in shaping public affairs, as in America."

do século XIX. Segundo Baskerville (1979, p. 166-167), "[d]urante os 'Anos Dourados' do início do século XIX, estudos em retórica (a arte do discurso efetivo) faziam parte dos currículos das principais faculdades americanas[36]. Podemos até dizer que a oratória não deixou de estar presente na comunicação oral estadunidense. Afinal, atualmente, é possível encontrar referências presentes em disciplinas, cursos e competições de oratória[37]. Contudo, com o advento do rádio e da televisão, muitas mudanças ocorreram em toda a concepção do falar em público. No caso do rádio, ele trouxe mudanças impactantes na história política dos EUA:

> Vimos como o rádio ampliou o alcance da voz do falante. Ele trouxe questões públicas para as casas por todo o país, tornando possível que todos ouvissem discursos públicos de importantes estadistas e sem ter que sair de sua poltrona. Durante os anos trinta e quarenta, o rádio criou uma assembleia virtual americana captada pelo ar. (BASKERVILLE, 1979, p. 213).[38]

A própria história do ex-presidente norte-americano Franklin Delano Roosevelt não teria sido a mesma sem a influência do rádio. Roosevelt passou a dirigir-se aos cidadãos por meio desse recurso em todas as noites de domingo, tornando-se muito mais presente na vida das pessoas do que qualquer outro presidente havia conseguido na história de seu país (BASKERVILLE, 1979, p. 177, 180, 182, 183)[39].

A televisão – a sucessora da rádio – também é retratada por Baskerville (1979, p. 213-214) como um instrumento utilizado para se falar ao público no mundo político norte-

[36] Original: During the "Golden Age" of the early nineteenth century, studies in rhetoric (the art of effective discourse) were part of the curricula of major American colleges.

[37] Exemplos: https://www.statesman.com/photogallery/TX/20200117/NEWS/117009998/PH/1 (alunos em uma competição de oratória no Museu George Washington Carver (16 janeiro, 2020); https://www.rhetoricsociety.org/aws/RSA/pt/sp/graduate_programs (cursos de pós-graduação).

[38] Original: We have seen how radio extended the range of the speaker's voice, brought public affairs into private homes all across the land, making it possible for Everyman to hear public addresses by eminent statesmen without moving from his easy chair. Radio during the thirties and forties created a virtual American town meeting of the air.

[39] Original: The president had established the practice of addressing the nation by radio on Sunday evenings, when the audience was largest.
[…]
The importance of the radio as an influence on the speaking of this period cannot be overestimated. This new medium of communication greatly increased the prominence and power of the president, while tending to diminish that of the Congress.
[…]
Radio came to maturity at precisely the right time for Franklin D. Roosevelt. It proved to be the perfect instrument to meet the national need for unity, and Roosevelt was admirably equipped to use it as a means of effective personal leadership.
[…]
Radio was also the principal medium for creating and conveying the presidential image to the nation, and eventually to the world.

americano. Mas seu impacto não alcançou as mesmas proporções, chegando a ser considerada como um instrumento de declínio do falar em público – declínio que também foi futuramente compartilhado com o rádio:

> E, como as gerações posteriores descobririam em sua experiência com o rádio e a televisão, o mercado se demonstra mais disposto a existir para o entretenimento e a diversão do que para a educação e o conhecimento. (BASKERVILLE, 1979, p. 102).[40]

Por ser um texto do final dos anos 1970, Baskerville não teve a oportunidade de comparar com o advento do computador e da internet – e dos vídeos on-line. Mas, conforme anteriormente mencionado, Anderson (2016) afirma que as palestras TED seriam um novo paradigma na comunicação oral, sendo parte da renascença do falar em público:

> De repente, uma arte antiga ganhou alcance global.
> Essa revolução levou ao renascimento da arte de falar em público. Muitos de nós já aguentamos anos de aulas maçantes na universidade, sermões intermináveis em igrejas e discurseiras políticas previsíveis. Mas as coisas não precisam ser assim. (ANDERSON, 2016, p. 10).

Em suma, atualmente, os eventos TED são aqueles eventos que abordam os mais variados tópicos em mais de 100 línguas, e que ocorrem em várias partes do mundo[41], como os Estados Unidos, Canadá, Brasil, Tanzânia, Escócia, Reino Unido, Índia, e Equador. Por sua vez, os vídeos TED, ou as populares TED Talks, são gravações de tais eventos – que são quase sempre editadas, podendo ser compostas por palestras, entrevistas, performances artísticas (contendo ou não contendo textos falados por seus apresentadores) –, além dos vídeos educacionais. No final de 2019, já havia mais de 4.000 vídeos postados com mais de 7 bilhões de visualizações no total. Deste modo, não é de se surpreender que a linguagem verbal das TED Talks têm causado instigação nas mentes de tantos pesquisadores, inclusive a pesquisadora deste trabalho. É por isso que, a presente pesquisa buscou analisar de forma multidimensional como a linguagem verbal das TED Talks funciona e molda essa maneira contemporânea de se espalhar ideias.

Assim sendo, para analisar a linguagem em uso, no caso, a linguagem verbal das TED

[40] Original: And, as later generations would discover in their experience with radio and television, a readier market seems to exist for entertainment and diversion than for education and enlightenment.
[41] https://www.ted.com/about/our-organization

Talks, foram consideradas as transcrições das falas dos vídeos TED em inglês feitas pelos voluntários TED[42] – os quais geralmente são estudantes ou especialistas em tradução, transcrição e legendagem. As transcrições seguem regras específicas[43] que, dentre elas são (figura 4): a duração de uma legenda tem que ser de no mínimo de 1 segundo e no máximo de 7 segundos; a duração máxima de caracteres por segundo é de 21 (CPS); pode-se ter no máximo 2 linhas por legenda; a quantidade máxima de caracteres por linha é de 42; a quantidade máxima de caracteres por legenda é de 84; a legenda deve iniciar no máximo 100 microssegundos antes da fala; não pode separar inteiros linguísticos (como sintagmas nominais) quando ocorre uma quebra de linha; uma linha não pode ser inferior a 50% do tamanho da outra; não é permitido unir o fim e o início de duas frases; não fazer a divisão das frases se não for estritamente necessário; usar parênteses para representar sons (e reações); e utilizar colchetes para textos que aparecem na tela.

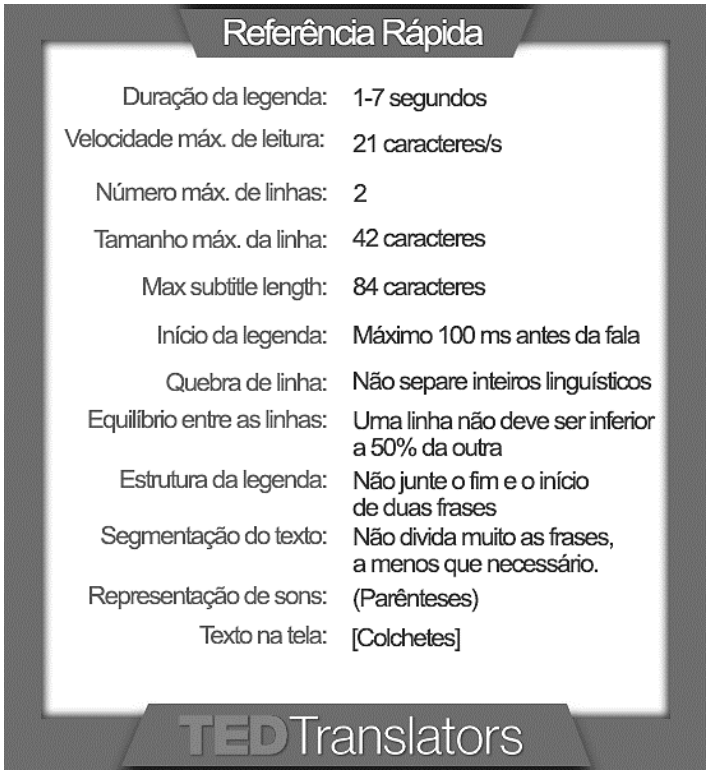


Figura 4: Referência rápida para tradutores e transcritores TED (fonte: https://translations.ted.com/Portuguese_(Brazil).

Os voluntários TED são orientados a transcrever as falas com a máxima fidelidade

[42] https://www.ted.com/participate/translate/transcribe
[43] https://translations.ted.com/How_to_Tackle_a_Transcript / https://translations.ted.com/Portuguese_(Brazil)

possível, sendo somente aceitas breves intervenções, caso seja preciso reescrever uma fala para que ela caiba no comprimento adequado de uma legenda, além de evitar transcrever sílabas vazias, repetições de palavras e erros desnecessários[44]. Desta forma, as transcrições são os textos utilizados na presente pesquisa, ou seja, se estamos falando sobre transcrição, estamos falando sobre texto.

Segundo Veirano Pinto (2013, p. 138), a linguagem verbal do cinema e da televisão possui três tipos de textos: o real (representado pela fala), o retratado (representado pelas legendas) e o planejado (representado pelos roteiros). Desta forma, se a linguagem verbal do cinema e da televisão possui três tipos de textos – o real, o retratado e o planejado –, podemos dizer o mesmo da linguagem verbal das palestras TED Talks. Afinal, em sua grande maioria, os textos TED são planejados (para representar a linguagem em uso), falados e legendados. Então, como as legendas ou as transcrições são um retrato da linguagem oral, mesmo que planejada, a linguagem TED pode ser considerada como oral retratada em formato de texto escrito.

Focando no caso específico das palestras – que compõem a maior parte das TED Talks –, a recomendação feita é de que seja utilizada a chamada linguagem conversacional, coloquial e falada, mesmo que memorizada para conseguir alcançar o público ouvinte (ANDERSON, 2016, p. 89 e 112). Deste modo, não somente é permitido a memorização dos textos das palestras, como a maioria dos apresentadores utiliza dessa técnica. Mas isso, diante de algumas condições para não deixar que a palestra soe memorizada e para garantir maior acessibilidade. Segundo o manual do palestrante TEDx[45], o palestrante tem seis meses para memorizar e se preparar para apresentar uma palestra, incluindo vários ensaios, que começam a partir do quarto mês antes do evento. Outros pontos destacados para garantir um "padrão TED" são o tempo estipulado para cada apresentação – 18 minutos –, período de tempo que nem sempre é cumprido; e a quantidade de palavras estipulada para cada 18 minutos de palestra são de 2.500 palavras (esses temas serão retomados na **Metodologia** deste trabalho). Outra característica interessante (anteriormente mencionada) é que nem tudo o que é produzido nos eventos TED é postado on-line, pois existe um controle de qualidade quanto a isso (ANDERSON, 2016, p. 151). Isso significa que, muitos dos vídeos postados podem ter sido previamente editados, ou seja, podem ocorrer cortes ou edições nas gravações e, consequentemente, nos textos originais.

[44] https://translations.ted.com/Guia_de_Estilo_para_Tradu%C3%A7%C3%B5es_e_Transcri%C3%A7%C3%B5es_em_Portugu%C3%AAs_(Brasil)_no_TED_Translators#Transcri.C3.A7.C3.B5es

[45] https://www.ted.com/participate/organize-a-local-tedx-event/tedx-organizer-guide/speakers-program/prepare-your-speaker/outline-script / https://storage.ted.com/tedx/manuals/tedxspeakerguide.pdf

Agora, focando no caso específico dos vídeos educacionais animados TED-Ed, não foi encontrado um guia específico para a sua elaboração, porém, é afirmado que eles são feitos segundo o estilo TED[46].

Isto posto, Anderson (2016, p. 45) nos traz um resumo em formato de perguntas do que se é esperado de uma TED Talk: O assunto me apaixona? Ele provoca curiosidade? Ele faz diferença para a plateia? Minha palestra é um presente ou um pedido? As informações são novas ou já são conhecidas? Eu consigo explicar o tema, com os exemplos necessários, no tempo concedido? Conheço o assunto o suficiente para que a palestra valha o tempo dos ouvintes? Tenho a credibilidade necessária para falar do assunto? Quais são as quinze palavras que resumem minha palestra? Essas quinze palavras fariam alguém se interessar por ouvir minha palestra?

Adicionalmente, segundo Anderson (2016, p. 161-162), um último item que pode ser adicionado à essa lista seria a inspiração, pois "[m]uitas ideias são arquivadas e provavelmente esquecidas em pouco tempo", porém, a inspiração "capta uma ideia e invade o núcleo da atenção de nossa mente: 'Alerta geral! Ideia nova relevante chegando! Preparar para ativar!'". Em suma, para ele, a TED é mais do que "uma receita para um ciclo de palestras mais interessante", mas uma visão do conhecimento e da compreensão como "a chave para sobrevivermos e crescermos no mundo novo que bate à nossa porta" (ANDERSON, 2016, p. 181 e 185).

Por fim, todos esses argumentos que foram até o momento expostos neste trabalho servem como base contextual, descritiva e especulativa sobre a estruturação, o funcionamento e a influência da linguagem verbal das TED Talks em áreas como a da pesquisa, da educação, de eventos, do entretenimento etc. Mas, apesar do grande interesse por entender e até copiar o chamado estilo ou formato TED, não existe uma pesquisa suficientemente abrangente que defina como e o quanto que a variação da linguagem verbal das TED Talks – mais especificamente, a variação gramatico-funcional dos textos das TED Talks que compõem o CoTED – exerce toda essa influência. Assim, sob uma visão mais empirista e multidimensional, a presente pesquisa pretende preencher essa lacuna.

2.5 Linguística de Corpus (LC) – Definição e histórico

A definição atual que podemos trazer da Linguística de Corpus (LC) é de uma área que

[46] https://ed.ted.com/educator?user_by_click=educator

trata do uso de corpus (singular) ou corpora (plural) computadorizados – que são coletâneas de textos, escritos ou de transcrições de fala, coletados criteriosamente e mantidos em arquivo de computador – a qual "questiona os paradigmas estabelecidos dos estudos lingüísticos e mostra novos caminhos para o lingüista, o professor, o tradutor, o lexicógrafo e muitos outros profissionais" (BERBER SARDINHA, 2004, p. XVII-XVIII). Deste modo, quando falamos da LC, nos valemos dessa imagem atual de pesquisadores e seus computadores e softwares especializados coletando e analisando corpus:

> Talvez, atualmente na mente dos linguistas, a linguística de corpus seja facilmente associada com a busca em tela após tela de linhas de concordância e listas de palavras geradas por software de computador, na tentativa de dar sentido a fenômenos em grandes textos ou em grandes coleções de textos. (O'KEEFFE; MCCARTHY, 2010, p. 3).[47]

Em resumo, também podemos dizer que um corpus (ou corpora) é composto por uma coletânea de dados linguísticos que seguem alguns critérios: textos autênticos, que não podem ser artificialmente criados, ou seja, precisam ser naturais (escritos ou falados) e autenticamente produzidos por humanos para, então, serem coletados e analisados via computador (BERBER SARDINHA, 2004).

Contudo, recorrendo à nossa história, o uso de corpus ou corpora era definido como simplesmente o uso de um conjunto de documentos (ou papeis). A coleta de corpus não é uma ideia tão nova, datada durante a Grécia Antiga – com o Alexandre "o grande" (356 a.C. e 323 a.C.) o qual definiu o chamado Corpus Helenístico[48] – como também durante a Antiguidade e a Idade Média, quando se produziam corpora de citações da Bíblia (BERBER SARDINHA, 2004, p. 3). O'Keeffe e McCarthy (2010) explicam que, durante o século XIII, estudiosos e seus ajudantes usavam a coleta de corpus, chamado por ele de método, na indexação das palavras da Bíblia, por exemplo. Como antes era um trabalho totalmente braçal, ou seja, os textos eram coletados manualmente pelos estudiosos da Bíblia e literários (e seus ajudantes), isso causava uma grande demora, demandando um número expressivo de pessoas no processo:

[47] Original: Corpus linguistics nowadays is perhaps most readily associated in the minds of linguists with searching through screen after screen of concordance lines and wordlists generated by computer software, in an attempt to make sense of phenomena in big texts or big collections of smaller texts.
[48] https://www.opengreekandlatin.org/

> Esse método de exegese, baseado em pesquisas detalhadas de palavras e frases em múltiplos contextos e em grandes quantidades de texto, remonta ao século XIII, quando estudiosos bíblicos e suas equipes de servos se debruçaram sobre página após página da Bíblia cristã e indexavam manualmente suas palavras, linha por linha, página por página. (O'KEEFFE; MCCARTHY, 2010, p. 3).[49]

Fora do escopo da Bíblia e da literatura, estudiosos também apresentavam um grande interesse pelos estudos da linguagem, em especial pela língua inglesa. Como exemplo, O'keeffe e Mccarthy (2010, p. 3) relatam que, somente após muitos anos de coleta de corpus em papel, com registros de uso da língua de 1560 a 1660, foi possível que o primeiro dicionário da língua inglesa pudesse ser criado. Outro exemplo, considerado por eles como o mais marcante, é o do dicionário da Oxford da década de 1880, o qual teve um corpus de mais de três milhões de tiras de papel:

> Talvez, o exemplo mais famoso de "corpus em tiras de papel" seja o de mais de três milhões de tiras, atestando o uso de palavras, que o projeto Oxford English Dictionary (OED) acumulou na década de 1880. Material armazenado no que hoje em dia poderia servir como um galpão de jardim. Esses milhões de pedaços de papel foram etiquetados ou classificados em categorias numa tentativa de organizá-los de modo significativo, de forma que o mundialmente famoso dicionário pudesse ser compilado. (O'KEEFFE; MCCARTHY, 2010, p. 3).[50]

Porém, esse trabalho braçal foi um dos grandes motivos para a desconfiança e críticas quanto a qualidade da coleta de corpus, pois sabe-se que "o ser humano não é talhado para tarefas desse tipo" (BERBER SARDINHA, 2004, p. 4). Mesmo assim, isso não impediu uma mudança significativa na coleta de corpus durante os anos 1950, com os estruturalistas e sua ideia de coletar dados reais ou a linguagem autêntica (linguagem realmente utilizada por falantes e escritores da(s) língua(s) e em situações reais). Segundo O'Keeffe e McCarthy (2010, p 4), enquanto o trabalho dos primeiros estudiosos bíblicos e literários fornecia o *modus operandi* de fundo na busca e indexação da palavra, os estruturalistas foram os precursores dos

[49] Original: This method of exegesis based on detailed searches for words and phrases in multiple contexts across large amounts of text can be traced back to the thirteenth century, when biblical scholars and their teams of minions pored over page after page of the Christian Bible and manually indexed its words, line by line, page by page.

[50] Original: And perhaps the most famous example of the 'corpus on slips of paper' is the more than three million slips attesting word usage that the Oxford English Dictionary (OED) project had amassed by the 1880s, stored in what nowadays might serve as a garden shed. These millions of bits of paper were, quite literally, pigeon-holed in an attempt to organise them into a meaningful body of text from which the world-famous dictionary could be compiled.

corpora como conhecemos hoje, não apenas no sentido de coleta de dados, mas em termos do compromisso de colocar dados de linguagem real no centro do que os linguistas estudavam[51]. Nota-se, desta forma, "a importância primordial de um corpus como fonte de informação, pois ele registra a linguagem natural realmente utilizada por falantes e escritores da língua em situações reais" (BERBER SARDINHA, 2004, p. 32). A partir de então, o conceito de corpus ou corpora passou a ser o de coletar a linguagem natural, ou seja, a linguagem realmente utilizada em contextos reais; isso, para que possamos chegar a conclusões baseadas em dados efetivos.

Contudo, a crítica quanto à coleta de corpus também ganhou força durante os anos 1950, pois foi nessa época quando o norte-americano Noam Chomsky (1957) trouxe um novo paradigma na Linguística com suas teorias racionalistas da linguagem, criticando o modo empirista dos trabalhos baseados em corpus como não confiáveis. Porém, um dos grandes problemas do racionalismo de Chomsky (1957) é que ele postula uma Gramática Universal inata, que independe da cognição e contém as regras de todas as línguas. Berber Sardinha (2004, p. 30) nos explica que, segundo Chomsky "o conhecimento provém de princípios, estabelecidos a *priori*" e "se fundamenta no estudo da linguagem por meio da introspecção [ou intuição], como forma de verificar modelos de funcionamento estrutural e processamento cognitivo da linguagem", isto é, a linguística "chomskyana gerativista enfatiza a determinação de quais agrupamentos sintáticos são possíveis (permissíveis) dado o conhecimento que um falante nativo possui de sua língua". Assim, conforme Petter (2002, p. 22) explica, segundo a Teoria Gerativa de Chomsky, todo ser humano possui uma gramática inata ou universal, pois se nasce com ela e a desenvolve independentemente de onde a pessoa nasça, e a sua intuição é o seu único guia para determinar o que é gramatical ou agramatical. De acordo com Chagas (2002, p. 149), outra questão a ser considerada é que, Chomsky – e Saussure – excluiu a estrutura da sociedade e sua história na análise da língua, pressupondo um falante idealizado em um mundo idealizado.

Em contrapartida, o britânico Michael Halliday (1985; 2014) foi o maior expoente contra a visão racionalista da linguagem de Chomsky. Halliday (1985; 2014) enxergava a linguagem humana como probabilidade e não possibilidade, ou seja, a linguística de Halliday segue a teoria sistêmica, que considera a probabilidade dos sistemas linguísticos segundo os

[51] Original: As Leech (1992) points out, it was in the 1950s, in the era of American structuralists such as Harris, Fries and Hill among others, when the notion of collecting real data came into its own. Where the work of the early biblical and literary scholars provides the background modus operandi of word searching and indexing, the structuralists were the forerunners of corpora not only in the sense of data gathering but in terms of the commitment to putting real language data at the core of what linguists study.

contextos em que os falantes se encontram; em outras palavras, as escolhas léxico-gramaticais feitas pelos falantes não são aleatórias, mas baseadas em relações sociais e contextuais – premissa base da LC:

> A teoria sistêmica recebe esse nome pelo fato de que a gramática de uma língua é representada na forma de redes de sistemas, não como um inventário de estruturas. É claro que a estrutura é uma parte essencial da descrição; mas ela é interpretada como uma forma externa tomada por escolhas sistêmicas, não como a característica definidora da linguagem. Uma língua é um recurso para construir significado, e o significado reside em padrões sistêmicos de escolha. (HALLIDAY, 1985, 2014, p. 23).[52]

Apesar de Halliday (1985; 2014) não ser definido como um linguista de corpus, pois sua Linguística Sistêmico-Funcional não faz uso de corpus ou de seus instrumentos de análise, a sua importância na estruturação da Gramática Sistêmico-Funcional e sua influência na LC são notadas.

Ainda assim, a LC é muitas vezes "acusada de apenas fazer *statement of facts*, ou seja, de apenas registrar as ocorrências lexicais e estruturais"; desta forma, para "deixar de ser um tipo de *contabilidade lingüística*, a Lingüística de Corpus necessita explicitar qual é o quadro teórico que lhe dá coerência e sustentação" (BERBER SARDINHA, 2004, p. 43). Segundo Tognini-Bonelli (2001), quando falamos que uma pesquisa é baseada em corpus (*corpus-based*), faz-se o uso de dados obtidos no corpus ou corpora estudados de forma a explorar uma ou mais teorias pré-existentes (antes da criação de corpus/corpora). Deste modo, com os dados obtidos, podemos validar, refutar ou aprimorar essa(s) teoria(s). Ademais, semelhante ao que a premissa da Linguística Sistêmico-Funcional de Halliday (1985; 2014) preconiza, o pesquisador precisa não somente se certificar de que esteja claro em sua análise qual teoria é utilizada como também considerar o papel que os contextos situacionais e culturais exercem nos textos que formam o corpus.

Brookes e McEnery, (2020, p. 380) afirmam que, outra crítica relacionada aos corpora é que o processo de conversão de textos em um corpus os divorcia dos contextos sociais em que foram originalmente produzidos e compreendidos. Por tal motivo, eles nos explicam que o

[52] Original: Systemic theory gets its name from the fact that the grammar of a language is represented in the form of system networks, not as an inventory of structures. Of course, structure is an essential part of the description; but it is interpreted as the outward form taken by systemic choices, not as the defining characteristic of language. A language is a resource for making meaning, and meaning resides in systemic patterns of choice.

analista de corpus deve trabalhar para garantir que sua análise retrate o papel que os contextos de situação e cultura desempenham na produção e consumo dos textos contidos em seu corpus[53].

Mas, vale dizer que, de fato, foram algumas mudanças nas tecnologias – em destaque a tecnologia da informação – que decididamente contribuíram para mudar esse panorama negativo contra a coleta de corpus e estabelecer essa imagem de pesquisadores e seus computadores e softwares especializados coletando e analisando textos (O'KEEFFE; MCCARTHY, 2010, p. 3). Sabe-se que os corpora tomaram a forma hoje conhecida por influência de um corpus não computadorizado chamado *Survey of English Usage* (SEU), compilado em Londres por Randolf Quirk e sua equipe, no ano de 1959 (BERBER SARDINHA, 2004, p. 3). Contudo, foi a partir dos anos 1960, com a invenção do computador, que passou a ser possível a criação e manutenção de grandes corpora, trazendo maior confiabilidade na coleta e na análise de grandes quantidades de dados.

O primeiro corpus eletrônico de linguagem escrita (inglês americano), o *Brown University Standard Corpus of Present-day American English* – ou simplesmente corpus *Brown* – foi criado em 1964 por Francis e Kucera e continha 1 milhão de palavras, sendo considerado como propulsor no desenvolvimento da Linguística de Corpus – como forma de abordagem que conhecemos hoje (BERBER SARDINHA, 2004, p 1). Quanto ao primeiro corpus eletrônico de linguagem falada (da língua inglesa) – com 220 mil palavras –, podemos atribuir ao britânico John McHardy Sinclair (1966), cujo trabalho pioneiro na área de léxico traçou os caminhos da maioria da pesquisa em Linguística de Corpus feita até hoje (BERBER SARDINHA, 2004, p 1 e 12). Contudo, foi a partir dos anos 1980, com o advento dos microcomputadores pessoais, que houve uma grande popularização de corpora e de ferramentas de processamento da linguagem natural, interesse que se estende até hoje (BERBER SARDINHA, 2004, p. 4-5). O'Keeffe e McCarthy (2010) resumem bem esse acontecimento tecnológico que, com o passar dos anos, revolucionou a coleta de corpus e, consequentemente, os estudos em Linguística de Corpus:

> No momento em que os computadores passaram a ser utilizáveis por qualquer pessoa que não fosse um pequeno grupo de especialistas, as tradições de (a) vasculhar textos

[53] Original: A related criticism of corpora is that the process of converting texts into a corpus divorces them from the social contexts in which they were originally produced and understood. The corpus analyst must therefore carry out more work to ensure that their analysis is cognizant of the role that contexts of situation and culture play in the production and consumption of the texts contained in their corpus.

> para encontrar todos os exemplos de uma determinada ocorrência de linguagem, (b) escrever dicionários baseados no uso, e (c) analisar a linguagem com base em dados reais de informantes foram todas consolidadas. Foi a revolução do hardware e do software nas décadas de 1980 e 1990 que realmente permitiu que a Linguística de Corpus como a conhecemos emergisse. (O'KEEFFE; MCCARTHY, 2010, p. 5).[54]

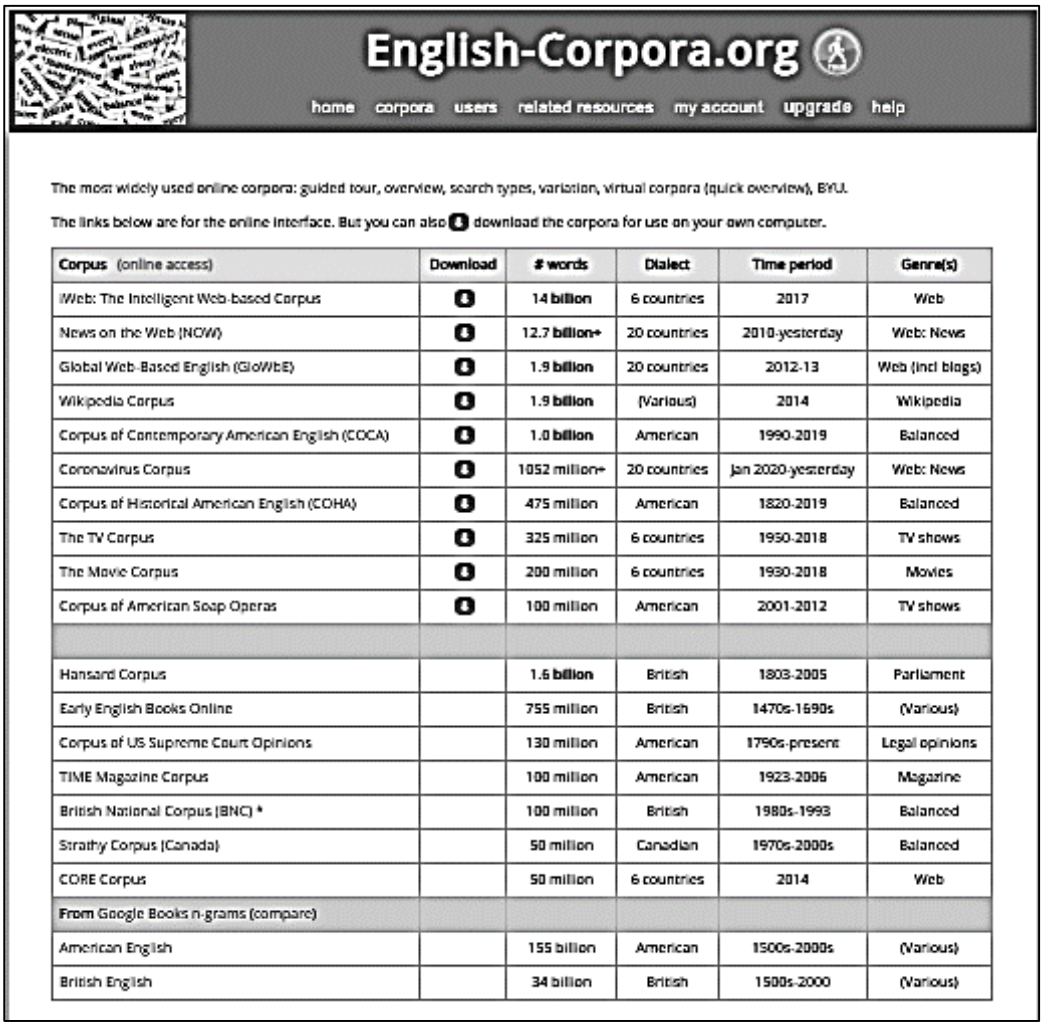
Certamente, equipamentos e softwares mais rápidos e com maior capacidade de processamento de dados possibilitaram a produção de pesquisas que não seriam tão praticáveis antes. Porém, a evolução tecnológica que ampliou em escala global o compartilhamento de dados entre pesquisadores das mais variadas áreas, assim como a da LC, foi a Internet – sem a qual, certamente, essa pesquisa não seria possível, não na proporção de dados – possibilitando a construção do CoTED. O'Keeffe e McCarthy (2010, p.5) comentam que, o crescimento paralelo da internet e da velocidade de download significavam que os dados e os resultados obtidos nas pesquisas agora poderiam ser facilmente transferidos de estudiosos para estudiosos[55]. E é por conta da internet que atualmente podemos contar com um número considerável de corpora on-line, como por exemplo, os corpora disponíveis no website criado por Mark Davies[56] (english-corpora.org[57]), que contém alguns dos corpora em língua inglesa mais utilizados por pesquisadores da LC, ilustrado na figura 5 a seguir:

[54] Original: By the time computers came to be usable by anyone other than a tiny group of specialists, the traditions of (a) trawling through texts to find all examples of a particular piece of language, (b) writing dictionaries based on attested usage, and (c) analysing language based on actual informant data were all well-established. It was the revolution in hardware and software in the 1980s and 1990s which really allowed corpus linguistics as we know it to emerge.

[55] Original: The parallel growth of the internet and fast download speeds meant that data and results could be transferred easily from scholar to scholar, while the role of the clumsy text scanners of the early 1980s some as big as household chest-freezers could be replaced by instant access to vast quantities of text already in electronic form. In tandem, heavy and cumbersome reel-to-reel tape recorders were replaced by manageable analogue cassette recorders in the 1970s and later by miniature digital recorders and small but high-powered video and DVD recorders, with a consequent positive effect on the ability of scholars to create spoken corpora.

[56] https://www.mark-davies.info/

[57] https://www.english-corpora.org/ (acessado em 10 de junho de 2021)



English-Corpora.org

home corpora users related resources my account upgrade help

The most widely used online corpora: guided tour, overview, search types, variation, virtual corpora (quick overview), BYU.

The links below are for the online interface. But you can also download the corpora for use on your own computer.

Corpus (online access)	Download	# words	Dialect	Time period	Genre(s)
iWeb: The Intelligent Web-based Corpus		14 billion	6 countries	2017	Web
News on the Web (NOW)		12.7 billion+	20 countries	2010-yesterday	Web: News
Global Web-Based English (GloWbE)		1.9 billion	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus		1.9 billion	(Various)	2014	Wikipedia
Corpus of Contemporary American English (COCA)		1.0 billion	American	1990-2019	Balanced
Coronavirus Corpus		1052 million+	20 countries	Jan 2020-yesterday	Web: News
Corpus of Historical American English (COHA)		475 million	American	1820-2019	Balanced
The TV Corpus		325 million	6 countries	1950-2018	TV shows
The Movie Corpus		200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas		100 million	American	2001-2012	TV shows
Hansard Corpus		1.6 billion	British	1803-2005	Parliament
Early English Books Online		755 million	British	1470s-1690s	(Various)
Corpus of US Supreme Court Opinions		130 million	American	1790s-present	Legal opinions
TIME Magazine Corpus		100 million	American	1923-2006	Magazine
British National Corpus (BNC) *		100 million	British	1980s-1993	Balanced
Strathy Corpus (Canada)		50 million	Canadian	1970s-2000s	Balanced
CORE Corpus		50 million	6 countries	2014	Web
From Google Books n-grams (compare)					
American English		155 billion	American	1500s-2000s	(Various)
British English		34 billion	British	1500s-2000	(Various)

Figura 5: english-corpora.org (website de corpora on-line criado por Mark Davies).

Cada corpus tem suas datas distintas de criação e a maior parte do trabalho de sua coleta foi feita pelo próprio Mark Davies. O COCA (*The Corpus of Contemporary American English*)[58], por exemplo, foi criado em 1990 e é constantemente atualizado, tendo já a marca de 1 bilhão de palavras, atualmente. No caso da língua portuguesa, temos *O corpus do português*[59] – composto por quatro corpora – também criado por Mark Davies, em 2004. Temos também o *Corpus Brasileiro*[60] do português contemporâneo com diversos registros da fala e da escrita, publicado por Berber Sardinha em 2011 (VEIRANO PINTO, 2013, p. 131). Vale ressaltar que corpora – de modo geral – não estão disponíveis para uso sem autorização prévia ou sua aquisição ocorre apenas via pagamento. Deste modo, conforme podemos ver, a história da LC está intrinsecamente ligada à história da tecnologia:

[58] https://www.english-corpora.org/coca/
[59] https://www.corpusdoportugues.org/xp.asp
[60] http://corpusbrasileiro.pucsp.br/cb/Inicial.html

> A história da Lingüística de Corpus está condicionada à tecnologia, que permite não somente o armazenamento de corpora, mas também a sua exploração e, por isso, está relacionada à disponibilidade de ferramentas computacionais para análise de corpus [...]. (BERBER SARDINHA, 2004, p. 15).

E foi assim que a LC surgiu e se desenvolveu. Desde os anos 1960, a LC – em especial a britânica – exerceu um papel fundamental na pesquisa de corpus, criando reconhecidos centros de pesquisa na área e incentivando outras vertentes em todo o mundo. Dentre essas vertentes, temos a norte-americana, que ganhou força especialmente durante os anos 1990 (BERBER SARDINHA, 2004, p. 5, 297-298). Todavia, apesar da reconhecida importância e influência da LC britânica, o presente trabalho utiliza como base a vertente norte-americana, impulsionada pelo linguista norte-americano Douglas Biber[61]:

> Atualmente [2004], nos EUA, o lingüista de corpus que mais se destaca é Douglas Biber, que impulsiona um tipo de metodologia de estudo de linguagem baseado em corpus chamado Análise Multidimensional [...]. A Lingüística de Corpus em desenvolvimento nos Estados Unidos, sob a liderança de Douglas Biber, tem um forte componente sociolingüístico, inspirado nos estudos de variação [...], com raízes nos estudos de William Labov. Douglas Biber é um lingüista de corpus versátil que, além de impulsionar a Análise Multidimensional, também trabalha com a descrição da padronização e da freqüência [...], conforme pode ser observado mais notadamente na gramática recente da língua inglesa organizada sob sua direção [...]. (BERBER SARDINHA, 2004, p. 299).

De fato, Biber é um dos maiores expoentes da Linguística de Corpus mundial e não somente norte-americana (BERBER SARDINHA, 2004, p. 5 e 13), sendo ele o responsável pela identificação das dimensões de variação de registro – parâmetros linguísticos coocorrentes subjacentes (variedades textuais definidas situacionalmente) – da língua inglesa, por meio de corpora; sendo ele também responsável pelo crescente interesse entre vários pesquisadores pelo texto representativo da linguagem humana em uso. Em suma, Biber se fundamenta na ideia de que a LC é mais do que um método, é uma abordagem – dentro da LA – baseada em corpus, utilizada para se enxergar a linguagem em uso (BERBER SARDINHA, 2004, p. 37).

[61] https://directory.nau.edu/person/biber

Por fim, atualmente, no contexto brasileiro, o pesquisador e professor da Pontifícia Universidade Católica de São Paulo, Tony Berber Sardinha[62], vem trazendo grandes contribuições na pesquisa de corpus, principalmente relacionada à compilação, exploração e análise de corpora para a investigação da variação linguística. Como exemplo de sua atuação, em parceria com Kauffman e Acunzo (em 2014), Berber Sardinha identificou as dimensões de variação do português brasileiro.

2.6 Linguística de Corpus (LC) – Tipologia e design

Uma das características que podemos atribuir a um corpus (ou corpora) é que ele não se resume a um só tipo. Veirano Pinto (2013) afirma que "os tipos mais comuns de corpora utilizados em análises linguísticas são o corpus especializado, o corpus que representa uma determinada língua, os corpora comparáveis, os corpora paralelos, o corpus de aprendiz, o corpus pedagógico, o corpus histórico ou diacrônico e o corpus de monitoramento", além do adicionado na lista, o "corpus audiovisual" (VEIRANO PINTO, 2013, p. 142). Segundo a autora (2013, p. 166), o corpus audiovisual é composto de transcrições ortográficas de eventos comunicativos orais acompanhadas pelas gravações em áudio ou vídeo de tais eventos. A autora acrescenta que, o propósito da construção desse tipo de corpus é possibilitar o estudo da linguagem oral em sua totalidade, isto é, considerar os aspectos linguísticos e extralinguísticos do evento comunicativo. Assim, certamente podemos classificar o corpus deste trabalho, o CoTED, como um corpus audiovisual.

Berber Sardinha (2004), por sua vez, explora algumas possíveis classificações ou tipologia dos corpora, segundo os critérios modo, tempo, seleção, conteúdo, autoria, disposição interna e finalidade. O quadro 1 a seguir resume bem esses conceitos:

MODO	Falado	Composto de porções de fala transcritas.
	Escrito	Compostos de textos escritos, impressos ou não.
TEMPO	Sincrônico	Compreende um período de tempo.
	Diacrônico	Compreende vários períodos de tempo.
	Contemporâneo	Representa o período de tempo corrente.
	Histórico	Representa o período de tempo passado.
SELEÇÃO	De amostragem[8]	Composto por porções de textos ou de variedades textuais, planejado para ser uma amostra finita da linguagem como um todo.

[62] http://lattes.cnpq.br/6940454346543706

	Monitor	Sua composição é reciclada para refletir o estado atual de uma língua. Opõe-se ao de amostragem.
	Dinâmico ou orgânico	É um corpus monitor cujo crescimento ou diminuição são permitidos.
	Estático	Corpus de amostragem. Fixo em sua composição.
	Equilibrado[9]	Os componentes são distribuídos em quantidades semelhantes.
CONTEÚDO	Especializado	Os textos são de um tipo específico.
	Regional ou dialetal	Os textos são provenientes de uma ou mais variedades sociolinguísticas específicas.
	Multilíngue	É composto por textos escritos em diferentes idiomas.
AUTORIA	De aprendiz	Os autores dos textos não são falantes nativos.
	De língua nativa	Os autores dos textos são falantes nativos.
DISPOSIÇÃO INTERNA	Paralelo	Os textos são comparáveis, por exemplo, original e tradução.
	Alinhado	As traduções aparecem abaixo de cada linha original.
FINALIDADE	De estudo	O corpus que se pretende descrever.
	De referência	Usado para fins de contraste com o corpus de estudo.
	De treinamento ou teste	Construído para permitir o desenvolvimento de aplicações e ferramentas de análise.

[8] Do inglês, "sample corpus".

[9] Do inglês, "balanced corpus".

Quadro 1: Tipologia dos corpora. Fonte: Yara (2020 – adaptado de BERBER SARDINHA, 2004, p. 20-22).

Pode-se, então, também classificar o corpus desta pesquisa como: 1) falado – possui transcrições de falas; 2) sincrônico – compreende um período de tempo contínuo e específico; 3) contemporâneo – no tempo atual; 4) de amostragem – composto por porções de textos, planejado para ser uma amostra finita da linguagem como um todo; 5) estático – corpus de amostragem; 6) especializado – os textos são de um tipo (registro) específico; 7) de língua nativa – os autores dos textos são falantes nativos (ou provavelmente possuem o inglês como segunda língua); e 8) estudo – o corpus a ser descrito, estudado e analisado.

É importante ressaltar que esta pesquisa se fundamenta em questões baseadas em sistemas probabilísticos. Uma vez que a linguagem tem um caráter probabilístico, surge "a possibilidade de estabelecer uma relação entre traços [estruturais, lexicais, pragmáticos e discursivos] que são mais comuns e menos comuns em determinado contexto" (BERBER SARDINHA, 2004, p. 23-24). Assim, para se chegar ao corpus desejado, é preciso ter um bom planejamento. Segundo Egbert (2019), Douglas Biber foi o primeiro pesquisador a se aprofundar na questão do desenho (*design*) do corpus, a ponto de influenciar importantes pesquisadores até os dias de hoje:

> O artigo de Biber apresenta uma investigação empírica minuciosa de questões críticas no desenho de corpus, como métodos de amostragem de corpus, tamanho do corpus (em palavras e textos), comprimento do texto, principais passos no desenho de corpus e construção de corpus.
> [...]
> Apesar de existirem muitos corpora, houve pouca discussão na literatura de linguística corpus sobre o processo de desenho e criação de corpus. Assim como a maioria das questões relacionadas a este tema, Biber (1993b) é uma exceção. (EGBERT, 2019, p. 27 e 35).[63]

Biber (1993) definiu o processo de desenho e coleta de corpus como um processo cíclico de quatro passos: 1) Investigação empírica piloto e análise teórica; 2) Desenho do corpus; 3) Compilação de parte do corpus; e 4) Investigação empírica – conforme também podemos observar na figura 6:

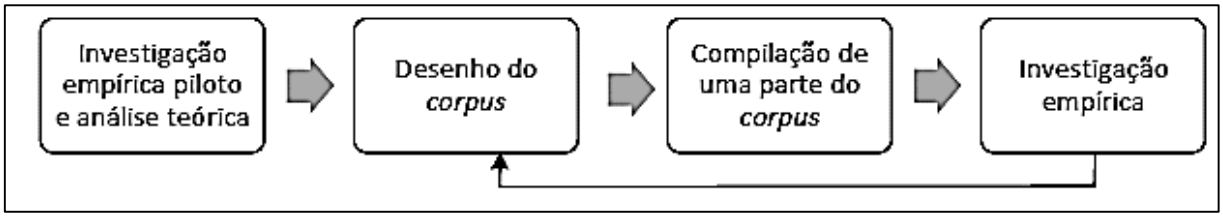


Figura 6: Design de corpus. Fonte: BIBER (1993, p. 256).

No primeiro passo – investigação empírica piloto e análise teórica –, antes mesmo de se pensar no tamanho do corpus a ser analisado, é preciso ter uma definição completa da população linguística ou do domínio-alvo a ser estudado, além da metodologia adotada:

> Geralmente, os pesquisadores se concentram no tamanho da amostra como a consideração mais importante para alcançar a representatividade: quantos textos devem ser incluídos no corpus e quantas palavras por amostra de texto. Livros sobre teoria amostral, no entanto, enfatizam que o tamanho da amostra não é a consideração mais importante na seleção de uma amostra representativa; em vez disso, uma

[63] Original: Biber's paper presents a thorough empirical investigation of critical issues in corpus design, such as corpus sampling methods, corpus size (in words and texts), text length, key steps in corpus design, and corpus construction.
[…]
Despite the many corpora in existence, there has been very little discussion in the corpus linguistics literature about the process of corpus design and creation. As with most issues related to this topic, Biber (1993b) is an exception.

> definição minuciosa da população-alvo e as decisões relativas ao método de amostragem são considerações prévias. (BIBER, 1993, p. 244).[64]

No segundo passo – desenho do corpus –, ao desenharmos um corpus para alcançarmos a representatividade da população ou domínio-alvo, devemos considerar tanto as perspectivas situacionais quanto as linguísticas. Para chegar na definição da população ou do domínio-alvo, Biber (1993, p. 244) elencou dois aspectos importantes: 1) definir quais são os limites dessa população ou domínio-alvo (quais textos são representativos ou não); e 2) definir uma organização hierárquica dessa população ou domínio-alvo (quais são as categorias dos textos, suas definições e distribuições das características linguísticas)[65]:

> A representatividade refere-se à medida em que uma amostra inclui toda a gama de variabilidade de uma população [ou do domínio-alvo]. No desenho do corpus, a variabilidade pode ser considerada a partir de perspectivas situacionais e linguísticas, e ambas são importantes na determinação da representatividade. Assim, um desenho de corpus pode ser avaliado na medida em que ele inclui: (1) uma gama de tipos de texto de uma língua, e (2) uma gama de distribuições linguísticas de uma língua. (BIBER, 1993, p. 244).[66]

No terceiro passo, juntamente com o quarto – compilação de parte do corpus e investigação empírica –, é quando passamos a enxergar a questão cíclica levantada por Biber (1993), pois é a partir da análise empírica do corpus piloto que podemos retrabalhar nos processos anteriores, caso necessário, até chegarmos no corpus definitivo e na sua análise:

> Independentemente do desenho inicial, a compilação de um corpus representativo deve prosseguir de forma cíclica: um corpus piloto deve ser compilado primeiro, representando uma gama relativamente ampla de variação, como também de registros

[64] Original: Typically researchers focus on sample size as the most important consideration in achieving represent: how many texts must be included in the corpus, and how many words per text sample. Books on sampling theory, however, emphasize that sample size is not the most important consideration in selecting a representative sample; rather, a thorough definition of the target population and decisions concerning the method of sampling are prior considerations.

[65] Original: Definition of the target population has at least two aspects: (1) the boundaries of the population—what texts are included and excluded from the population; (2) hierarchical organization within the population—what text categories are included in the population, and what are their definitions.

[66] Original: Representativeness refers to the extent to which a sample includes the full range of variability in a population. In corpus design, variability can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness. Thus a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions in a language.

> e textos. A marcação gramatical deve ser realizada nesses textos, como base para investigações empíricas. Em seguida, análises empíricas devem ser realizadas neste corpus piloto para confirmar ou modificar os vários parâmetros do projeto. Partes desse ciclo poderiam ser realizadas de forma quase contínua, com novos textos sendo analisados à medida que se tornam disponíveis, mas também deve haver etapas distintas de extensa investigação empírica e revisão do desenho do corpus. (BIBER, 1993, p. 256).[67]

Todavia, com os avanços tecnológicos e da internet – em comparação com 1993 até o presente momento –, aumentaram-se as possibilidades de pesquisas na área da Linguística Aplicada. Egbert (2019, p. 28)[68] comenta que, além desses avanços tecnológicos, grandes mudanças também ocorreram no campo da Linguística de Corpus, aumentando em ritmo acelerado o uso de métodos linguísticos de corpus e corpora na pesquisa linguística. Diante disso, Egbert (2019, p. 36) propõe ampliarmos os quatro passos de Biber (1993) para nove[69]:

1. Estabelecer (e projetar) os objetivos e o planejamento da pesquisa.
2. Definir o domínio-alvo (ou população).
3. Desenhar o corpus.
4. Coletar a amostra.

---

[67] Original: Regardless of the initial design, the compilation of a representative corpus should proceed in a cyclical fashion: a pilot corpus should be compiled first, representing a relatively broad range of variation but also representing depth in some registers and texts. Grammatical tagging should be carried out on these texts, as a basis for empirical investigations. Then empirical research should be carried out on this pilot corpus to confirm or modify the various design parameters. Parts of this cycle could be carried out in an almost continuous fashion, with new texts being analysed as they become available, but there should also be discrete stages of extensive empirical investigation and revision of the corpus design.

[68] Original: Much has changed in the twenty-five years since the publication of Biber's paper on representativeness. Advances in computing speed and memory have grown at an exponential rate (Moore 2006). This, combined with the power of the internet, has made it possible to collect and analyze electronic texts on a scale and in ways that were unfathomable in 1993. In addition to these technological advances, major changes have also taken place in the field of corpus linguistics.
[…]
This trend provides evidence that the use of corpora and corpus linguistic methods in linguistics research is increasing at an accelerated rate.

[69] Original: […] I propose a nine-step process for designing and collecting a representative corpus that builds on the cycle Biber proposed. The steps in this process are the following:
1. Establish (and project) research objectives and design
2. Define the target domain (population)
3. Design the corpus
4. Collect the sample
5. Annotate the corpus
6. Evaluate target domain representativeness
7. Evaluate linguistic representativeness
8. Repeat steps 3–5, if necessary
9. Report

5. Fazer a anotação do corpus.
6. Avaliar a representatividade do domínio-alvo (ou população).
7. Avaliar a representatividade linguística.
8. Repetir passos 3-5, se necessário.
9. Criar relatório.

Segundo Toledo (2020, p. 23-24), as principais diferenças entre o processo cíclico de Biber, de 1993, e de Egbert, de 2019, podem ser encontradas no primeiro, terceiro, quinto e nono passos. No primeiro, "além de ressaltar a importância do estabelecimento de objetivos claros de pesquisa para a subsequente compilação do *corpus*," Egbert "preconiza a projeção de objetivos adicionais para o *corpus*, para que ele possa ser utilizado em pesquisas futuras, fazendo assim com que o *corpus* compilado se torne mais atrativo a outros pesquisadores". No terceiro, "além de sugerir que o pesquisador tome as decisões de planejamento", de forma a "garantir que a amostragem compilada seja representativa da população [ou do domínio-alvo] em estudo", Egbert (2019) "sugere que decisões de âmbito mais prático, como prazos, custos e modos de armazenagem dos arquivos também devem ser levadas em consideração". No quinto passo, fazer a anotação do corpus "não figura como uma etapa apartada das demais apresentadas por Biber, mas como parte integrante da etapa *Investigação empírica*". No nono passo, por sua vez, Egbert (2019) fala sobre a "documentação do processo de concepção e compilação do *corpus*", sendo "uma atividade contínua durante toda a pesquisa", em que, "a descrição detalhada da metodologia utilizada na pesquisa baseada em *corpus* não somente servirá para orientar o próprio pesquisador durante todos os passos de sua própria pesquisa, como também servirá de base para futuras pesquisas".

Por fim, segundo Egbert (2019), toda essa responsabilidade no desenho, na qualidade e na representatividade do corpus não precisa recair somente nos ombros do pesquisador, mas pode ser partilhada entre os criadores, pesquisadores e consumidores de tais corpus:

> Para encerrar, farei uma pergunta que faço aos alunos das minhas aulas de linguística de corpus. Quem é responsável pela qualidade e representatividade de um corpus – o criador do corpus, o pesquisador que usa o corpus, ou o consumidor de resultados e materiais baseados no corpus? Podemos argumentar que o criador do corpus é o responsável.
> [...]
> Depois de uma calorosa discussão, meus alunos sempre chegam à mesma conclusão que cheguei: todos os três grupos – criadores de corpus, pesquisadores e consumidores

> – têm a responsabilidade de avaliar o desenho e a representatividade do corpus e tomar suas decisões com base em suas conclusões. (EGBERT, 2019, p. 36).[70]

No caso específico desta pesquisa, podemos dizer que, a princípio, foram analisadas todas as 3.411 transcrições coletadas em conjunto (TED Geral ou o CoTED todo) – não houve a necessidade de se fazer um corpus piloto. Porém, considerando os resultados obtidos nas análises prévias (seção 3), foi decidido dividir e analisar o corpus CoTED em três partes, que correspondem às três categorias aqui atribuídas: TED Tradicional, TEDx e TED-Ed. Desta forma, foi possível decidir qual design de corpus melhor contribuiria para esta pesquisa.

2.7 Linguística de Corpus (LC) – Algumas considerações

Segundo O'keeffe e Mccarthy (2010, p. 7), o uso da Linguística de Corpus (LC) tem se demonstrado bastante extenso como ferramenta de estudos da linguagem humana, do ensino e aprendizagem de idiomas, da análise do discurso, da estilística literária, da linguística forense, da pragmática, da tecnologia da fala, da sociolinguística, da comunicação em saúde, etc. Nesse sentido, os autores explicam que, "a LC é um meio para um fim em vez de um fim em si mesmo", ou seja, a LC é uma ferramenta de pesquisa que "leva as descobertas para além dos domínios do léxico ou da gramática, aplicando suas técnicas a outras questões, algumas mais facilmente respondidas pela análise computacional do que outras"; e tais áreas podem ser "tão diversas quanto a aquisição de uma segunda língua ou os estudos de mídia"[71]. Também, a LC consegue estar presente não somente em questões de pesquisas acadêmicas como também em áreas diversas da comunicação humana, como a empresarial, a comercial, a editorial etc. (BERBER SARDINHA, 2004, p. 6). Desta forma, foi com essa "revolução no pensamento linguístico", apadrinhada pela tecnologia e a Linguística de Corpus, que passamos – com mais firmeza – da idealização para a sistematização da observação da evidência:

---

[70] Original: In closing I will ask a question that I pose to students in my corpus linguistics classes. Who is responsible for the quality and representativeness of a corpus—the corpus creator, the researcher who uses the corpus, or the consumer of results and materials based on the corpus? We might argue that the creator is the responsible party.
[...]
After a lively discussion, my students always come to the same conclusion I have come to: all three groups—corpus creators, researchers, and consumers—have a responsibility to evaluate corpus design and representativeness and make informed decisions based on their conclusions.

[71] Original: In this sense, CL is a means to an end rather than an end in itself. That is, CL leads to insights beyond the realms of lexis or grammar by applying its techniques to other questions, some more easily answered by computational analysis than others. In areas as diverse as second language acquisition and media studies, CL can be applied as a research tool.

> Está em curso uma verdadeira revolução no pensamento lingüístico, com implicações sérias sobre como respondemos a questões fundamentais, tais como o que é língua, como ela é organizada, como deve ser estudada, como deve ser ensinada. A mola propulsora dessa revolução é a tecnologia, mais especificamente o computador. Já foi dito que o computador pessoal, com memória poderosa e capacidade de armazenamento, começa a desempenhar, nas ciências humanas, o papel transformador que o telescópio teve na física e nas ciências exatas. Passamos da idealização para a sistematização da observação da evidência. (BERBER SARDINHA, 2004, p. XVII)

Deste modo, como a LC trabalha dentro de um quadro conceitual formado por uma abordagem empirista e uma visão da linguagem como sistema probabilístico, no qual certos traços linguísticos são mais frequentes que outros, é papel do pesquisador discutir os porquês de a linguagem ser usada de tal forma a exibir tais padrões ou fenômenos (BERBER SARDINHA, 2004).

2.8 Análise Multidimensional (AMD) – Contextualização e embasamento teórico

Chegamos ao aspecto teórico principal desta pesquisa, a Análise Multitraço e Multidimensional de Variação de Registro (*Multifeature Multidimensional Analysis of Register Variation*), ou simplesmente Análise Multidimensional, abreviada como AMD (BIBER, 1988). A AMD é um quadro teórico-metodológico de análise linguística criado por Douglas Biber nos anos de 1980, primeiramente discutida em sua tese de doutorado (BIBER, 1984) e depois publicada no livro *Variation Across Speech and Writing* (BIBER, 1988). Até então, havia poucos estudos da variação de registros linguísticos (variedades de texto definidas situacionalmente, ou seja, definidas pelo contexto em que ocorrem na sociedade) que fossem tão abrangentes, pois um grande número desses estudos trazia um foco gerativo-transformacional de Chomsky em suas análises, que não considerava a linguagem natural (ou verdadeiramente utilizada) nem os modos escrito e falado de forma empírica e de igualitária importância:

> Assim, os dados para análise dentro desse paradigma deliberadamente excluem os erros que ocorrem na "fala real", no dialeto, na variação de registro e em quaisquer características linguísticas que dependam de um discurso ou contexto situacional de interpretação. Pelo contrário, esses dados não são retirados da fala real ou da escrita

> real. Eles estão muito mais próximos da escrita estereotipada do que da fala em si. (BIBER, 1988, p. 7).[72]

Conforme Biber (1988, p. 13) explica, a maioria dos estudos prévios analisava a variação de registro em termos de um único parâmetro subjacente, sugerindo que existia uma distinção situacional básica e única entre os registros. Muitos desses estudos assumiam que a variação de registro poderia ser analisada em termos de distinções simples e dicotômicas, de modo que as variedades fossem consideradas como, por exemplo, formais ou informais, planejadas ou não planejadas, e assim por diante. Além disso, como Biber (1988) também explica, nenhuma dessas abordagens aplicou métodos empíricos para identificar os conjuntos de características linguísticas que coocorrem mas, em vez disso, os pesquisadores propunham conjuntos de características que pareciam funcionar em conjunto com base em suas percepções e intuições, influenciados pela visão gerativo-transformacional de Chomsky[73]. Ademais, outro fator importante que influenciava as pesquisas da época eram as dificuldades metodológicas ou tecnológicas, ou seja, não era possível analisar grandes quantidades de textos, registros e características linguísticas com a tecnologia e ferramentas disponíveis (BIBER, 2019, p. 12)[74].

Em contrapartida, Biber (1988, p. 54-55) relata que, as pesquisas sociolinguísticas realizadas mostravam que a variação da linguagem natural é bastante complexa, dando-nos motivos para presumir a existência de múltiplas dimensões de variação entre os gêneros falado e escrito, por exemplo. Assim, segundo essa premissa, a expectativa de que múltiplas características linguísticas e múltiplas dimensões sejam necessárias para uma descrição adequada da variação linguística entre os registros, apoiados pelo nosso domínio do uso da linguagem em sociedade:

---

[72] Original: Thus the data for analysis within this paradigm deliberately exclude performance errors of 'actual speech', dialect, and register variation, and any linguistic features that depend on a discourse or situational context of interpretation. Although these data are not taken from actual speech or actual writing, they are much closer to stereotypical writing than speech in their form.

[73] Original: […] most previous studies analyzed register variation in terms of a single underlying parameter, suggesting that there was a single basic situational distinction among registers. Second, most previous studies assumed that register variation could be analyzed in terms of simple, dichotomous distinctions, so that varieties are either formal or informal, planned or unplanned, and so on. And finally, none of these early approaches applied empirical methods to identify sets of co-occurring linguistic features. Rather, researchers proposed sets of features that seemed to work together, based on their perceptions and intuitions.

[74] Original: However, despite its fundamental importance, there were few comprehensive linguistic analyses of register variation before the 1980s. This disregard was due mostly to methodological difficulties: until that time, it was simply not feasible to analyze the full range of texts, registers, and linguistic characteristics required for a comprehensive analysis of register variation. However, with the availability of large online text corpora and computational analytical tools, such analyses became possible.

> Os trabalhos de Hymes, Labov, Gumperz – entre outros – descreveram a variação linguística sistemática com base em uma ampla gama de parâmetros sociais e situacionais, incluindo a classe social e o grupo étnico dos participantes, a relação social e situacional entre os participantes, além do cenário e o propósito da comunicação (Brown e Fraser, 1979). O quadro resultante desta pesquisa é de um complexo acoplamento de características e funções linguísticas, com características únicas servindo muitas funções e funções únicas sendo marcadas por muitas características. (BIBER, 1988, p. 54-55).[75]

Vale ressaltar que, o termo antes usado por Biber (1988) para "registro" era "gênero" (*genre*). O termo foi mudado a partir de 1995, o qual também se diferencia do termo "tipo de texto" (*text type*), que designa um conjunto de textos formado exclusivamente com base em critérios linguísticos. Biber (2019, p. 170) entende que, as categorias de registro são determinadas com base em critérios externos, relacionados ao objetivo e ao assunto tratado pelo falante, ou seja, elas são atribuídas com base no uso e não na forma. Assim, os registros caracterizam textos com base em critérios externos, ao passo que, os tipos de texto representam agrupamentos de textos semelhantes em sua forma linguística, independentemente do registro:

> Por exemplo, um artigo acadêmico sobre história asiática representa uma exposição formal e acadêmica segundo o propósito do autor, mas sua forma linguística pode ser narrativa e mais semelhante a alguns tipos de ficção do que a artigos acadêmicos científicos ou de engenharia. O gênero [registro] de tal texto seria exposição acadêmica, mas seu tipo de texto pode ser narrativa acadêmica" (BIBER, 2019, p. 170)[76].

Biber (2019, p. 12) explica que, segundo a premissa da LC, a variação linguística ou de registro(s) é inerente à linguagem humana, pois cada falante faz escolhas sistemáticas tanto na pronúncia, na morfologia, na seleção de palavras e na gramática, e tais escolhas são

[75] Original: Work by Hymes, Labov, Gumperz, and others has described systematic linguistic variation across a wide range of social and situational parameters, including the social class and ethnic group of participants, the social and situational relationship between the participants, the setting, and the purpose of communication (Brown and Fraser 1979). The picture emerging from this research is one of a complex coupling of linguistic features and functions, with single features serving many functions and single functions being marked by many features.

[76] Original: For example, an academic article on Asian history represents formal, academic exposition in terms of the author's purpose, but its linguistic form might be narrative-like and more similar to some types of fiction than to scientific or engineering academic articles. The genre of such a text would be academic exposition, but its text type might be academic narrative.

associadas a diferentes registros, refletindo suas características situacionais[77]. Em outras palavras, os traços linguísticos encontrados nos textos (registros) variam sistematicamente de acordo com os contextos e propósitos comunicativos específicos nos quais são produzidos, ou seja, tal variação não ocorre com a mesma frequência assim como também não é aleatória, pois existe uma correlação com os contextos de uso. Assim sendo, Biber (1988, p. 20) nos diz que, para descobrir os padrões de coocorrência das características linguísticas subjacentes, os quais realmente definem as dimensões linguísticas de uma língua ou de registro(s), é preciso ter uma seleção representativa de textos e características linguísticas para análise como crucial requisito. O autor afirma que, a gama de possíveis padrões de coocorrência deve ser representada nas características escolhidas para análise[78]. À vista disso, conforme Biber (1988; 2019) nos explica, somente com a ajuda da tecnologia – ao fazer o uso de ferramentas computacionais automáticas e semiautomáticas – que é possível colocar em prática a proposição da AMD de descrever de forma comparativa e multidimensional um grande número de textos autênticos – corpus ou corpora –, utilizando uma quantidade maior de parâmetros situacionais e de características linguísticas. Por parâmetros temos quesitos como formalidade, impessoalidade, oralidade, período histórico e estilo. Porém, essas categorias não são limitadas ao aspecto dicotômico de A versus B como formal versus informal ou planejado versus espontâneo; mas de aspectos de variação ou graduação de cada parâmetro, com as polaridades entre o mais formal versus o menos formal, o mais informal versus o menos informal, o mais espontâneo versus o menos espontâneo etc. (BIBER, 1988, p. 20)[79]. As características linguísticas – ou variáveis – consideradas por Biber (1988), por sua vez, são: os adjetivos; os advérbios; os substantivos; os verbos; as conjunções; as preposições; os pronomes; as orações complementares e subordinadas formadas com o pronome relativo *that* e com a partícula de

[77] Original: Register variation is inherent in human language: a single speaker will make systematic choices in pronunciation, morphology, word choice, and grammar associated with different registers, reflecting the situational characteristics of those registers.

[78] Original: To uncover the strong co-occurrence patterns that actually define linguistic dimensions in English, we need to analyze much longer texts, a much larger number of texts taken from many genres [registers], and frequency counts of many linguistic features. Those features that co-occur in different texts across several genres [registers] are the ones that define the basic linguistic dimensions of English. A representative selection of texts and linguistic features for analysis is thus a crucial requisite to this type of analysis; the range of possible variation must be represented in the texts chosen for analysis, and the range of possible co-occurrence patterns must be represented in the features chosen for analysis.

[79] Original: All of these conclusions regarding similarities and differences among texts are inadequate, because the relations among texts cannot be defined unidimensionally Fiction is not simply similar to or different from scientific prose; rather it is more or less similar or different with respect to each dimension. Given that the linguistic variation among texts comprises several dimensions, it is no surprise that the relations among texts must be conceptualized in terms of a multi-dimensional space.

infinitivo *to*; e as orações relativas com pronomes wh.

Isto posto, a abordagem nos estudos linguísticos proposta por Biber (1988) segue o modo inverso das pesquisas que buscavam por meio da análise situacional e funcional identificar as características linguísticas, ou seja, a AMD identifica os conjuntos dessas características linguísticas de forma a utilizá-las na interpretação funcional e situacional que elas exercem nos textos. Biber (1988, p. 13) afirma que, em outras abordagens, as análises começavam com uma distinção situacional ou funcional, e a identificação das características linguísticas associadas a essa distinção vinham como um segundo passo. Como exemplos, ele menciona pesquisadores que davam prioridade às dimensões funcionais, como formal versus informal, restrito versus elaborado ou envolvido versus não envolvido, e somente depois identificavam as características linguísticas associadas à cada dimensão. Como crítica à essa abordagem, Biber (1988) diz que, apesar de os agrupamentos de características serem identificados em termos de função compartilhada, eles não representam necessariamente as dimensões linguísticas, ou seja, esses agrupamentos de características não representam necessariamente as características linguísticas que coocorrem frequentemente nos textos (falado ou escrito). Foi por tal motivo que ele apresentou uma abordagem oposta, utilizando técnicas quantitativas para identificar os grupos de características linguísticas que realmente coocorrem nos textos para, posteriormente, interpretar esses agrupamentos em termos funcionais[80]. Deste modo, para Biber (1988, p. 13), a "dimensão linguística, em vez da dimensão funcional, é dada como prioridade. Esta abordagem baseia-se no pressuposto de que significativos padrões de coocorrência de características linguísticas marcam as dimensões funcionais subjacentes"[81].

Tendo em vista os preceitos logo acima expostos, em seu estudo inicial, Biber (1988) elencou 23 registros diferentes da língua inglesa – conforme podemos ver na tabela 1 – para representar a maior gama de possibilidades situacionais encontradas nos corpora utilizados – conforme podemos ver no quadro 2 –, que compreendem desde biografias até linguagem de

[80] Original: Most analyses begin with a situational or functional distinction and identify linguistic features associated with that distinction as a second step. For example, researchers have given priority to functional dimensions such as formal/informal, restricted/elaborated, or involved/ detached, and subsequently they have identified the linguistic features associated with each dimension. In this approach, the groupings of features are identified in terms of shared function, but they do not necessarily represent linguistic dimensions in the above sense; that is these groupings of features do not necessarily represent those features that co-occur frequently in texts. The opposite approach is used here: quantitative techniques are used to identify the groups of features that actually co-occur in texts, and afterwards these groupings are interpreted in functional terms.

[81] Original: The linguistic dimension rather than functional dimension is given priority. This approach is based on the assumption that strong co-occurrence patterns of linguistic features mark underlying functional dimensions.

rádio e TV. Desse número de registros, Biber (1988) analisou 481 textos, com um total de 960 mil palavras:

#	Registros	Tradução	Número Palavras aproximando (tokens)	Textos
1	Biographies	Biografias	30.000	14
2	Personal letters	Cartas Pessoais	6.000	6
3	Professional letters	Cartas profissionais	10.000	10
4	Face-to-face conversations	Conversas face-a-face	115.000	44
5	Telephone conversations	Conversas ao telefone	32.000	27
6	Popular lore	Cultura popular	30.000	14
7	Official documents	Documentos oficiais	28.000	14
8	Editorial	Editoriais jornalísticos	54.000	27
9	Interview (Public conversations, debates and interviews)	Entrevistas	48.000	22
10	Science-fiction	Ficção científica	12.000	6
11	Adventure fiction	Ficção de aventura	26.000	13
12	Mystery fiction	Ficção de mistério	26.000	13
13	General fiction	Ficção geral	58.000	29
14	Romantic fiction	Ficção romântica	26.000	13
15	Humor	Humor	18.000	9
16	Spontaneous speeches	Palestras espontâneas	26.000	16
17	Planned speeches	Palestras preparadas	31.000	14
18	Hobbies (skills and hobbies)	Passatempos	30.000	14
19	Academic prose	Prosa acadêmica	160.000	80
20	Broadcast	Rádio e TV	38.000	18
21	Religion	Religião	34.000	17
22	Press reportage	Reportagem jornalística	88.000	44
23	Press reviews	Resenhas jornalísticas	34.000	17
Total			908.000	481

Tabela 1: Os 23 registros utilizados por Biber (1988).

Como será melhor explicado na Metodologia deste trabalho, para se realizar a AMD Funcional Aditiva do CoTED, foi feita uma análise contrastiva observando as principais semelhanças e diferenças gramático-funcionais entre os 23 registros escritos e falados da língua

inglesa – considerados por Biber em 1988 – e o corpus das TED Talks, tanto no geral quanto as três categorias antes definidas (TED tradicional, TEDx e TED-Ed).

Lancaster-Oslo/Bergen Corpus ou somente LOB Corpus	Continha por volta de 1 milhão de palavras – representando a linguagem escrita
London-Lund Corpus of Spoken English ou somente London-Lund Corpus	Continha por volta de 500 mil palavras – representando a linguagem falada
Cartas profissionais e pessoais do próprio Biber	Continha 10.000 palavras

Quadro 2: Corpora utilizados por Biber (1988).

Conforme podemos ver no quadro 2, o corpus da língua inglesa criado por Biber em 1988 é formado por outros corpora já existentes. No caso do CoTED – como será visto na Metodologia deste trabalho –, não foram utilizados corpora em sua composição, mas transcrições de textos dos vídeos TED Talks de modo a formar o corpus das TED Talks (seção 3.1.1).

Por sua vez, para identificar e classificar as características linguísticas dos textos, Biber (1988) criou um programa ou etiquetador chamado *Biber Tagger* (seção 3.1.2), que abrange as características morfossintáticas, semânticas e de marcação de posicionamento – programa que foi aprimorado ao longo do tempo aumentando o número de características linguísticas que podem ser analisadas (inicialmente eram 67, atualmente são consideradas 128 características). A seguir, no quadro 3, temos a lista (atual) das variáveis linguísticas do *Biber Tagger*:

Lista de etiquetas contabilizadas pelo programa *Biber Tagger*			
#	**Variável**	**Descrição**	**Exemplo**
1	<abstracn>	substantivo relacionado a coisas ou processos (abstract/process noun)	*thing, development, stress*
2	<act_ipv>	phrasal verb intransitivo relacionado a atividade (activity - intransitive phrasal verb)	*come on, hold on,*
3	<act_tpv>	phrasal verb transitivo relacionado a atividade (activity - transitive phrasal verb)	*bring up, find out*
4	<actv>	verbo relacionado a atividade/ação (activity verb)	*walk, dance*
5	<adj_attr>	adjetivo atributivo (attributive adjective)	*handsome, smiling*
6	<advs>	advérbio (adverb)	*well, too, rather*
7	<agls_psv>	voz passiva sem agente (agentless passive)	*are given, is plotted*
8	<all_advl>	todos os advérbios de posicionamento (sum of stance adverbs)	*of course, probably, really*
9	<all_jth>	todas as categorias de orações complementares com that controladas por adjetivos (sum stance that complement clauses controlled by adjectives)	*that this was a tactical decision*

10	<all_jto>	todas as categorias de orações complementares com to controladas por adjetivos (sum stance to complement clauses controlled by adjectives)	*it is difficult to mantain*
11	<all_nth>	todas as categorias de orações complementares com that controladas por substantivos (sum stance that complement clauses controlled by nouns)	*they believe that the wage*
12	<all_nto>	to usado em oração controlada por substantivos de posicionamento (to complement clause controlled by stance nouns)	*they say that failure to do it is…*
13	<all_th>	soma das orações complementares com that (sum stance that complement clauses)	*he was aware that*
14	<all_to>	todas as orações complementares com to (sum stance to complement clauses)	*my goal is to look to the future*
15	<all_vth>	todas as orações complementares com that controlada por verbos (sum stance that complement clauses controlled by verbs)	*there is a fear that*
16	<all_vto>	todas as orações complementares com to controlada por verbos (sum stance to complement clauses controlled by verbs)	*it is a dangerous thing to do*
17	<alladj>	todas as categorias de adjetivos (all adjectives)	*great, good*
18	<allconj>	todas as categorias de conjunções (all conjunctions)	*and, but, or*
19	<allmodal>	todas as categorias de verbos modais (all modals)	*must, can, could, may*
20	<allpasv>	todos os usos de voz passiva (all passives)	*is done, was made*
21	<allpro>	todas as categorias de pronomes (all pronouns)	*I, you, he, she, it, we, you, they*
22	<allverb>	todas as categorias de verbos, excluindo verbos auxiliares (all verbs)	*go, buy, see, sell,*
23	<allwh>	todas as palavras iniciadas por WH- (all WH- words)	*what, when, when, who*
24	<allwhrel>	todas as orações relativas com pronome WH- (all WH- relative clauses)	*he was asking what happened*
25	<amplifr>	advérbio qualificador/amplificador (amplifier)	*absolutely, entirely*
26	<aspectpv>	phrasal verb acurativo/determinativo (aspectual verb – phrasal verb)	*carry out, look at*
27	<aspectv>	verbo acurativo/determinativo (aspectual verb)	*keep, stop, start, begin*
28	<atadvl>	advérbio atitudinal (attitudinal adverb)	*unfortunately, suprisingly*
29	<att_jth>	that usado em oração complementar controlada por adjetivo atitudinal (that complement clause controlled by atitudinal adjective)	*so obnoxious that*
30	<att_nth>	that usado em oração complementar controlada por substantivo atitudinal (that complement clause controlled by attitudinal noun)	*there were also rumors that*
31	<att_vth>	that usado em oração complementar controlada por verbo atitudinal (that complement clause controlled by attitudinal verb)	*I thought that it was good*
32	<be_state>	verbo to be indicativo de estado (be state)	*am, is are, was, were*
33	<by_pasv>	voz passiva com agente e preposição by (BY-passive)	*he was arrested by the officer*
34	<causev>	verbo causativo (causative verb)	*cause, enable, allow, help*
35	<cognitn>	substantivo cognitivo (cognitive noun)	*believe, find, remember*
36	<colorj>	adjetivo – cor (color adjective)	*blue, white*
37	<commpv>	phrasal verb transitivo relacionado a comunicação (communication – transitive phrasal verb)	*you might find out it works*

38	<commv>	verbo relacionado a comunicação (communication verb)	*say, shout, ask, offer, talk*
39	<concrtn>	substantivo concreto (concrete noun)	*table, wall*
40	<conjncts>	conjunção (conjunction)	*however, therefore, thus*
41	<contrac>	contração (contraction)	*isn't, don't*
42	<copulapv>	phrasal verb de ligação/copula (copular phrasal verb)	*people get pissed off*
43	<downtone>	advérbio suavizador (downtoner)	*nearly, only, merely*
44	<dsre_vto>	to usado em oração complementar controlada por verbos de desejo, intenção e decisão (to complement clauses controlled by verbs of desire, intention, and decision)	*I want to finish it*
45	<efrt_vto>	to usado em oração controlada por verbos de modalidade, causalidade e esforço (to complement clause controlled by verbs of modality, causation and effort)	*to introduce us to the*
46	<evalj>	adjetivo avaliativo (evaluative adjective)	*bad, beautifil, fine, good, poor*
47	<existv>	verbo relacionado a existência ou relacionamento (existence verb)	*appear, indicate, represent*
48	<fact_jth>	that usado em oração complementar controlada por adjetivo factivo (that complement clause controlled by factive adjective)	*it is clear that the presence is*
49	<fact_vth>	that usado em oração complementar controlada por verbo factivo (that complement clause controlled by a factive verb)	*she demonstrated that it*
50	<factadvl>	advérbio factivo (factive adverb)	*the observation that he put*
51	<fct_nth>	that usado em oração complementar controlada por substantivo factivo (that complement clause controlled by a factive noun)	*actually, really*
52	<finlprep>	preposição desacompanhada (stranded preposition)	*about, from, at*
53	<gen_emph>	advérbio ou palavra quantificador(a) enfático(a) (general emphatics)	*just, really, so*
54	<gen_hdg>	advérbio delimitador/atenuador (general hedges)	*almost, maybe*
55	<groupn>	substantivo relacionado a grupo ou instituição (group / institution noun)	*crowd of people, flock of*
56	<have>	verbo have (have)	*have*
57	<humann>	substantivo animado (animate noun)	*the soldier, guy*
58	<inf>	infinitivo (infinitive)	*to go, going*
59	<it>	pronome it (pronoun it)	*it*
60	<jcmp>	that usado em oração complementar controlada por adjetivo (that complement clause controlled by adjective)	*I was confident that it would*
61	<lkly_jth>	that usado em oração complementar controlada por adjetivo de probabilidade (that complement clause controlled by adjective of likelihood)	*it is certain unlikely that*
62	<lkly_nth>	that usado em oração complementar controlada por substantivo de probabilidade (that complement clause controlled by noun of likelihood)	*it is unlikely that the govern*
63	<lkly_vth>	that usado em oração complementar controlada por verbo de probabilidade (that complement clause controlled by verb of likelihood)	*she failed to appear in court*
64	<lklyadvl>	advérbio de probabilidade (likelihood adverb)	*it is so unlikely that*

65	<mentalpv>	phrasal verb transitivo relacionado a atividade mental (mental – transitive phrasal verb)	*haven't you found that out yet*
66	<mentalv>	verbo relacionado a atividade mental (mental verb)	*think, want, love*
67	<mntl_vto>	to usado em oração complementar controlada por verbo de cognição (to complement clause controlled by verb of cognition)	*I would love to kick it*
68	<n_nom>	nominalização no singular (noun nominalization)	*table, vase, mirror*
69	<n>	substantivo (noun)	*house, hand*
70	<nec_mod>	verbo modal de necessidade (necessity modal)	*ought, should, must*
71	<nfct_nth>	that usado em oração complementar controlada por substantivo não factivo (that complement clause controlled by non-factive noun)	*the fact that all doctors are*
72	<nonf_vth>	that usado em oração complementar controlada por verbo não factivo (that complement clause controlled by non-factive verb)	*I didn't realize that he had left*
73	<nonfadvl>	advérbio não factivo (non-factive adverb)	*though*
74	<o_and>	conjunção coordenada – conectivo clausal (coordinating cojunction – clausal connector)	*and, or*
75	<occurpv>	phrasal verb intransitivo relacionado a evento/ocorrência (occurrence – intransitive phrasal verb)	*shut up, go off, stand up*
76	<occurv>	verbo relacionado a evento/ocorrência (occurrence verb)	*change, develop, occur*
77	<p_and>	conjunção coordenada – conectivo frasal (coordinating cojunction - phrasal connector)	*and, if*
78	<pany>	pronome indefinido (indefinite pronoun)	*someone, everything*
79	<pasttnse>	passado (past tense)	*went, looked*
80	<pdem>	pronome demonstrativo (demonstrative pronoun)	*this, that, these, those*
81	<perfects>	aspecto perfeito (perfect aspect)	*have, had + perfect*
82	<pl_adv>	advérbio de lugar (place adverb)	*here, there*
83	<placen>	substantivo – lugar (place noun)	*Egypt, Brazil*
84	<pos_mod>	verbo modal de possibilidade (modal of possibility)	*can, may, might, could*
85	<prcessn>	substantivo relacionado a processos (process noun)	*discharged water*
86	<prd_mod>	verbo modal preditivo (modal of prediction)	*will, would, shall*
87	<pred_adj>	adjetivo predicativo (predicative adjective)	*nice, right, easier,*
88	<prep>	preposição (preposition)	*in, on, from, at*
89	<pres>	verbo no presente (present)	*dances, wants, go*
90	<pro_do>	verbo do como substituto de outro verbo ou sintagma verbal / verbo vicário / pró-verbo (pro-verb do)	*I, my, mine*
91	<pro1>	pronome em 1a pessoa (1st person pronoun)	*you, your, yours*
92	<pro2>	pronome em 2a pessoa (2nd person pronoun)	*he, she, his, her, hers*
93	<pro3>	pronome em 3a pessoa (3rd person pronoun)	*he does*
94	<prob_vto>	to usado em oração complementar controlada por verbos de probabilidade e fato (to complement clause controlled by verbs of probability and simple fact)	*the way to get to our house*
95	<prtcle>	partícula do discurso (discourse particle)	*now*
96	<prv_vb>	verbo de cognição (private verb)	*believe, feel, think*
97	<pub_vb>	verbo dicendi (public verb)	*assert, complain, say*
98	<quann>	substantivo – quantidade (quantity noun)	*each, all, every, pair of*
99	<rel_obj>	oração WH- em posição de objeto (WH- relative clause on object position)	*I don't know what it is*
100	<rel_pipe>	oração WH- com preposição inicial (pied-piping construction)	*in how many different places*

101	<rel_sub>	oração WH- em posição de sujeito (WH- relative clause on subject position)	*that's what I am saying*
102	<relatnj>	adjetivo – relacionamentos (relational adjective)	*additional, average, chief*
103	<sizej>	adjetivo – tamanho (size adjective)	*big, small, tiny*
104	<spch_vto>	to usado em oração complementar controlada por verbos de atos de fala (to complement clause controlled by speech act verbs)	*I didn't say I agree*
105	<spl_aux>	advérbio usado entre verbo auxiliar e verbo principal (split auxiliary)	*she looked exactly like*
106	<sua_vb>	verbo de persuasão (suasive verb)	*ask, command, insist*
107	<sub_cnd>	conjunção subordinativa condicional (conditional subordination)	*if, unless*
108	<sub_cos>	subordinação causativa (causative subordination)	*because*
109	<sub_othr>	outros advérbios/conjunções usadas em orações subordinadas (other adverbial subordinators)	*as, except, until*
110	<tccncrtn>	substantivo relacionado a assuntos técnicos ou concretos (technical / concrete nouns)	*rock, chickens*
111	<that_del>	omissão de that em oração subordinada (that deletion)	*I thought it was a good film*
112	<that_rel>	that em orações subordinadas com pronome relativo (that relative clauses)	*the way in which this happens*
113	<timej>	adjetivo – tempo (time adjective)	*late, new, recent, young*
114	<tm_adv>	advérbio de tempo (time adverb)	*Wednesday*
115	<topicj>	adjetivo relacionado a tópicos (topical attributive adjective)	*chemical, political, sexual*
116	<typetokn>	relação entre item e ocorrência (type token ratio)	*type-token ratio*
117	<vcmp>	that usado em oração complementar controlada por verbo (that complement clause controlled by verb)	*they warned me that it's bad*
118	<vprogrsv>	presente contínuo (present progressive)	*to be + verbo com -ing*
119	<wh_cl>	oração com pronome WH- (WH- clauses)	*tokens*
120	<wh_ques>	pronome WH- usado em perguntas (WH- questions)	*the rings that she wore*
121	<whiz_vbn>	modificador pós-nominal da voz passiva (passive postnominal modifier)	*who, whose, what, which*
122	<wordcnt>	quantidade de palavras (word count)	*by the way in which*
123	<wrdlngth>	tamanho de palavra (word length)	-
124	<x1_jto>	to usado em oração complementar controlada por adjetivos de certeza (to complement clause controlled by adjectives of certainty)	*I am certain to regret it*
125	<x2_jto>	to usado em oração complementar controlada por adjetivos de habilidade/desejo (to complement clause controlled by adjectives of ability/willingness)	*I am axious to go, willing to*
126	<x3_jto>	to usado em oração complementar controlada por adjetivos de afeição pessoal (to complement clause controlled by adjectives of personal affect)	*I am sorry to hear that*
127	<x4_jto>	to usado em oração complementar controlada por adjetivos de facilidade/dificuldade (to complement clause controlled by adjectives of ease/difficulty)	*they are easy to steal*
128	<x5_jto>	to usado em oração complementar controlada por adjetivos de avaliatividade (to complement clause controlled by evaluative adjectives)	*this one is nice to smell*

Quadro 3: Lista de etiquetas contabilizadas pelo programa Biber Tagger (baseado em Biber, 1988) – Fonte: Resende (2019).

Biber (1988, p. 76) também explica que, após a definição e atribuição das características linguísticas, é preciso calcular sua frequência em cada texto, que incluem[82]: 1) a frequência média; 2) as frequências máximas e mínimas, ou seja, as ocorrências máximas e mínimas em qualquer texto; 3) a amplitude, ou seja, a diferença entre os valores máximos e os mínimos; e 4) o desvio padrão, ou seja, a medida do grau de dispersão.

Para a contagem das frequências das palavras ou características linguísticas do corpus etiquetado, Biber (1988) desenvolveu o programa *Biber Tag Count* (seção 3.1.2), que automaticamente padroniza as frequências das etiquetas por 1.000 palavras, permitindo assim a comparação entre textos ou registros de um corpus ou corpora. Como próximo passo, são utilizados programas de computador na busca e identificação das coocorrências (variação) dessas características linguísticas (padrões de coocorrências gramaticais) nos textos de corpora (registro(s)), mensurando e analisando as diversas dimensões encontradas com tais resultados. Conforme a definição feita por Biber (1988), entende-se por dimensão como um conjunto de traços comunicativos latentes compartilhados pelos textos de um corpus, ou seja, uma dimensão linguística é determinada com base em um padrão consistente de coocorrência entre as características linguísticas. Em outras palavras, uma dimensão linguística é identificada quando um grupo de características linguísticas sistematicamente coocorrem em textos; e cada dimensão compreende a um grupo independente de características linguísticas coocorrendo, e cada padrão de coocorrência pode ser interpretado em termos funcionais. "O resultado é uma avaliação empírica de quantas dimensões independentes existem; uma avaliação de quais funções são independentes e quais funções são associadas a uma mesma dimensão; além de uma avaliação da importância relativa às diferentes dimensões" (BIBER, 1988, p. 13 e 14)[83]. Para se chegar aos padrões de coocorrências gramaticais e verificar as dimensões de variação, que nos permite visualizar as características linguísticas partilhadas entre os textos, é preciso

[82] Original: […] descriptive statistics for the frequencies of the linguistic features in the entire corpus of texts. Included are: (l) the mean frequency (2) the maximum and minimum frequencies, that is, the maximum and minimum occurrences in any text, (3) the 'range, that is, the difference between the maximum and the minimum values, and (4) the 'standard deviation', a measure of the spread of the distribution — 68% of the texts in the corpus have frequency values within the spread of plus or minus one standard deviation from the mean score.

[83] Original: […] a linguistic dimension is determined on basis of a consistent co-occurrence pattern among features. That is, when a group of features consistently Co-occur in texts, those feature a linguistic dimension.
[…]
By defining 'dimension' from a strictly linguistic perspective, it is possible to identify the set of dimensions required to account for the linguistic variation within a set of texts. Each dimension comprises an independent group of co-occurring linguistic features, and each co-occurrence pattern can be interpreted in functional terms. The result is an empirical assessment of how many independent dimensions there are; an assessment of which functions are independent and which are associated with the same dimension; and an assessment of the relative importance of different dimensions.

adotar a análise fatorial (seção 2.6), que possibilita resumir as interrelações entre um grande grupo de variáveis de forma concisa e construir as dimensões subjacentes, pois:

> [...] é possível decifrar uma dimensão unificada subjacente a cada conjunto de características linguísticas que coocorrem. Nesse sentido, uso a análise fatorial – comumente utilizada em outras ciências sociais e comportamentais –, buscando resumir as interrelações entre um grande grupo de variáveis de forma concisa, e construir as dimensões subjacentes (ou construtos) – que são conceitualmente muito mais claras do que as muitas medidas linguísticas consideradas individualmente. (BIBER, 1988, p. 64).[84]

Desta forma, a análise das dimensões permite visualizar as características linguísticas partilhadas por uma porção significativa de dados, ou seja, os textos podem se situar ao longo de uma escala que vai de mais a menos em relação a cada traço comunicativo (BERBER SARDINHA, 2004). Tais "dimensões são caracterizadas por uma polaridade", isto é, os "registros com escores positivos representam melhor um dos polos, enquanto os registros com os maiores escores negativos ilustram o polo oposto" (BERBER SARDINHA, 2004).

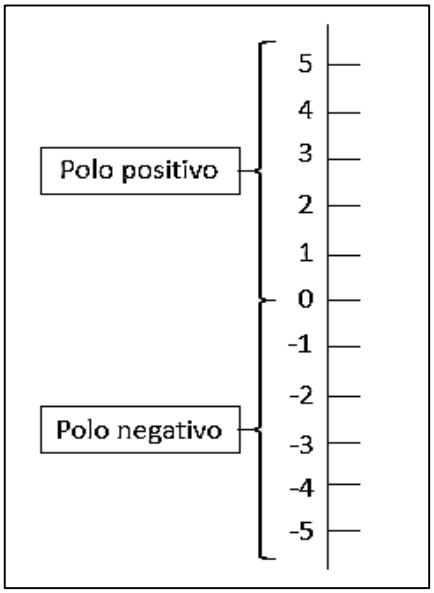


Figura 7: escala de variação da dimensão baseada em Biber (1988).

[84] Original: Working from this assumption, it is possible to decipher a unified dimension underlying each set of co-occurring linguistic features. In this sense, I am using factor analysis as it is commonly used in other social and behavioral sciences: to summarize the interrelationships among a large group of variables in a concise fashion; to build underlying dimensions (or constructs) that are conceptually clearer than the many linguistic measures considered individually.

A figura 7 ilustra um exemplo no qual podemos posicionar os textos ou registros ao longo de uma escala de variação. Caso estivéssemos considerando os registros palestras espontâneas e palestras preparadas, cada um deles poderia estar em uma posição diferente ou semelhante na escala de acordo com as características linguísticas que os mais representam. Considerando que as palestras espontâneas estivessem no polo positivo e as palestras preparadas no polo negativo, isso seria o indício de que o grupo de características linguísticas que estão mais presentes nas palestras espontâneas estariam menos presentes nas palestras preparadas, e vice-versa. Tal escala de variação foi utilizada nesta pesquisa para alocar as TED Talks ao longo da escala de registros feita por Biber (1988) – como parte da AMD Funcional Aditiva do CoTED.

2.9 Análise Multidimensional (AMD) – Análise fatorial

A análise fatorial é uma ferramenta estatística de abordagem multitraço ou multidimensional usada para o agrupamento de características linguísticas com base em sua coocorrência, ou seja, agrupamento desses dados em fatores. De uma grande quantidade de variáveis (características linguísticas, nesse caso), a análise fatorial consegue separar em fatores aquelas variáveis que coocorrem:

> Em uma análise fatorial, um grande número de variáveis originais, neste caso as frequências de características linguísticas, é reduzido a um pequeno conjunto de variáveis derivadas, os "fatores". Cada fator representa alguma área nos dados originais, que foram resumidos ou generalizados. Em outras palavras, cada fator representa uma área de alta variância compartilhada nos dados, ou seja, é um agrupamento de características linguísticas que coocorrem com alta frequência. Os fatores são combinações lineares das variáveis originais, derivadas de uma matriz de correlação de todas as variáveis. (BIBER, 1988, p. 79).[85]

Mas a análise fatorial por si não faz todo o trabalho. Ela é uma ferramenta de análise que não funcionaria bem sem uma boa fundamentação teórica de pesquisa, adequação dos dados coletados que serão analisados e a inclusão de múltiplas características linguísticas. Assim,

[85] Original: Factor analysis is the primary statistical tool of the multi-feature/multi-dimensional approach to textual variation. In a factor analysis, a large number of original variables, in this case the frequencies of linguistic features, are reduced to a small set of derived variables, the 'factors'. Each factor represents some area in the original data that can be summarized or generalized. That is, each factor represents an area of high shared variance in the data, a grouping of linguistic features that co-occur with a high frequency. The factors are linear combinations of the original variables, derived from a correlation matrix of all variables.

como Biber (1988, p. 65) explica, embora a análise fatorial permita a identificação quantitativa de dimensões subjacentes dentro de um conjunto de textos, ela não poderá ser empregada de forma profícua sem que a pesquisa tenha sido previamente delineada com embasamento teórico; em outros termos, antes de realizar a análise fatorial, devem ser determinados a gama de situações e os propósitos comunicativos disponíveis em uma língua. Além disso, ele argumenta que, devem ser coletados textos que representem essa gama de variação, ou seja, as características linguísticas – que são indicadores potencialmente importantes dentro de um domínio – devem ser identificadas com antecedência e medidas em cada um dos textos. Também, ele nos diz que, a preparação inadequada ou a distorção desses pré-requisitos teóricos pode invalidar os resultados de uma análise fatorial, pois, mesmo sendo uma ferramenta analítica primordial, ela depende de uma base teórica advinda de uma base de dados adequada de textos e da inclusão de múltiplas características linguísticas[86].

Segundo Biber (1988), na análise fatorial, a base de dados deve incluir cinco vezes mais textos do que características linguísticas a serem analisadas, isso porque, para se representar uma gama de possibilidades situacionais e de processamento do objeto estudado, é preciso ter um grande número de textos.[87] Contudo, também é preciso incluir a maior gama possível de características linguísticas que possam ser importantes na análise fatorial (BIBER, 1988, p. 71-72).[88] Conforme previamente mencionado, em seu estudo inicial, Biber (1988, p. 71-72) elencou 67 características linguísticas (atualmente são 128) que foram consideradas relevantes para se determinar as dimensões funcionais subjacentes da língua inglesa.

Também, Biber (1988, p. 75) fala que, ao se utilizar a estatística na análise dos textos de corpora, é importante também implementar a normalização das frequências das características linguísticas encontradas. Isso significa que, uma frequência bruta (*raw*

[86] Original: Although factor analysis enables quantitative identification of underlying dimensions within a set of texts, it cannot be employed usefully apart from a theoretically-motivated research design. That is, before performing a factor analysis, the range of communicative situations and purposes available in a language must be determined, and texts representing that range of variation must be collected. In the same way, linguistic features that are potentially important indicators within the domain must be identified in advance and measured in each of the texts. Inadequate preparation or skewing in these theoretical prerequisites can invalidate the results of a factor analysis (Gorsuch 1983:336ff). That is, factor analysis provides the primary analytical tool, but is dependent on the theoretical foundation provided by an adequate data base of texts and inclusion of multiple linguistic features.

[87] Original: In factor analysis, the data base should include five times as many texts as linguistic features to be analyzed (Gorsuch 1983:332). In addition, simply representing the range of situational and processing possibilities in English requires a large number of texts.

[88] Original: Prior to any comparison of texts, principled decision must be made concerning the linguistic features to be used. For the purposes of this study, previous research was surveyed to identify potentially important linguistic features — those that have been associated with particular communicative functions and therefore might be used to differing extents in different types of texts. No a priori commitment is made concerning the importance of an individual linguistic feature or the validity of a previous functional interpretation during the selection of features. Rather, the goal is to include the widest possible range of potentially important linguistic features.

frequency) não pode ser utilizada na comparação entre os textos caso eles sejam de tamanhos diferentes. Isso porque influenciaria no cálculo de frequência das características. Por isso que é importante realizar a normalização por 1.000 (por exemplo), ou seja, pega-se o total de frequência de uma característica linguística, divide-se pelo total de palavras no texto analisado e multiplica o resultado por 1.000[89]. É por esse motivo que o próprio *Biber Tag Count* (utilizado na presente pesquisa) foi elaborado para automaticamente padronizar as frequências das etiquetas por 1.000 palavras; podendo-se, desta forma, calcular o escore de fator de cada texto, cujo valor mínimo estabelecido por Biber (1988, p. 87) é de .30.

A princípio, é feita a extração fatorial inicial sem rotação para que o número máximo de fatores possa ser extraído. Porém, Biber (1988, p. 82) ressalta que, como o objetivo da análise fatorial é reduzir o número de variáveis observáveis para um número relativamente pequeno de construtos subjacentes, a análise fatorial continua extraindo fatores até que toda a variância compartilhada entre as variáveis tenha sido contabilizada; mas apenas os primeiros fatores são propensos a explicar a quantidade não trivial de variância compartilhada e, portanto, merecem uma maior consideração. E como não existe um método matematicamente exato para determinar o número de fatores a serem extraídos, Biber (1988) optou por uma diretriz consideravelmente simples ao examinar um conjunto de valores Eigen (*Eigenvalues*)[90], que são índices diretos da quantidade de variação encontrada em cada fator (tabela 2). Com esses valores, é possível observar quais fatores apresentam os maiores *eigenvalues*, responsáveis por uma maior variância explicada:

Fator	Valores Eigen	% de variação compartilhada
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%

[89] Original: The frequency counts of all linguistic features are normalized to a text length of 1,000 words (except for type/token ratio and word length — see discussion in Appendix Il). This normalization is crucial for any comparison of frequency counts across texts, because text length can vary widely. A comparison of non-normalized counts will give an inaccurate assessment of the frequency distribution in texts.

[90] Eigenvalue: valores Eigen, a quantidade de variância compartilhada por um fator específico após a primeira redução de dimensão, ou seja, a extração preliminar, não rotacionada. (Fonte: RESENDE, 2019).

10	1,27	1,9%
11	1,23	1,9%

Tabela 2: Primeiros 11 Eigenvalues da língua inglesa (BIBER, 1988, p. 83).

Na tabela 2 – logo acima –, é possível ver, por exemplo, que na dimensão 1 da língua inglesa existe 26,8 % de variação compartilhada entre os textos analisados por Biber (1988). Diante de tal variação encontrada, é feito o gráfico de sedimentação (*scree plot*)[91] – figura 8 –, o qual mostra um ponto de ruptura (comumente chamado de "cotovelo") indicando quais fatores contribuem mais ou menos para a análise geral[92]:

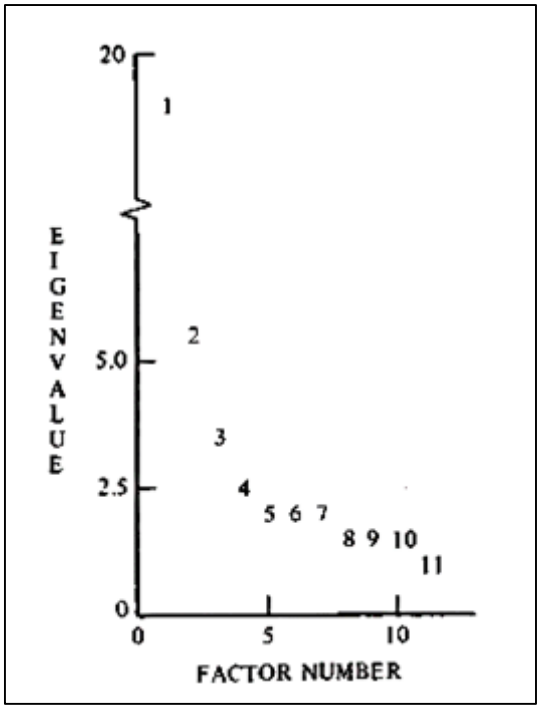


Figura 8: *Scree plot* dos valores *Eigen* da língua inglesa (BIBER, 1988, p. 83).

[91] Gráfico de sedimentação: também chamado de scree plot, é o gráfico gerado a partir das comunalidades, que é quanto cada variável se relaciona com as outras. Ele possibilita determinar o número de fatores por meio da representação gráfica dos valores Eigen. Isto é, ele é um gráfico dos autovalores versus o número de fatores por ordem de extração. (Fonte: RESENDE, 2019).

[92] Original: [...] the purpose of factor analysis is to reduce the number of Observed variables to a relatively small number of underlying Constructs. A factor analysis will continue extracting factors until all of the shared variance among the variables has been accounted for; but only the first few factors are likely to account for a nontrivial amount of shared variance and therefore be worth further consideration. There is no mathematically exact method for determining the number of factors to be extracted. There are, however. several guidelines for this decision. One of the simplest is to examine a plot of the eigenvalues, which are direct indices of the amount of variance accounted for by each factor. Such a plot is called a scree plot, and Will normally show a characteristic break indicating the point at which additional factors contribute little to the overall analysis.

Conforme podemos ver na figura 8, existe uma curva acentuada que engloba os fatores de 4 a 7, a qual levou Biber a inicialmente considerar 7 fatores em sua pesquisa feita em 1988 (BIBER, p. 89-90).

Como próximo passo, chega-se, então, à fase final da análise fatorial, que é a análise rotacionada. Assim, dos sete fatores – ou sete agrupamentos de características linguísticas –, seis foram de fato aproveitados, porém, após alguns estudos posteriores e revisões, acabaram por restar cinco deles (BIBER, 2009). Logo a seguir, são apresentados os cinco agrupamentos de características linguísticas (fatores) que compõem as cinco dimensões da língua inglesa, definidas por Biber (1988; 2009) – com tais resultados, obtemos os escores médios dos registros nas cinco dimensões (lembrando que, o valor mínimo estabelecido por Biber (1988, p. 87) é de .30) – tabelas 3-7:

Estrutura do Fator 1 (BIBER, 1988; 2009)			
Produção marcada por envolvimento versus informacional			
Polo positivo		**Polo negativo**	
verbo privado	0,96	substantivo	-0,47
apagamento de ‘that’	0,91	tamanho de palavra	-0,54
contração	0,90	preposição	-0,54
verbo no tempo presente	0,86	razão forma-ocorrência	-0,58
pronome de segunda pessoa	0,86	adjetivo em posição atributiva	-0,80
verbo ‘do’	0,82	(advérbio de lugar	-0,32)
negação analítica	0,78	(voz passiva sem agente	-0,38)
pronome demonstrativo	0,76	(oração adjetiva reduzida de particípio	-0,39)
enfatizador	0,74	(oração adjetiva reduzida de gerúndio	-0,42)
pronome de primeira pessoa	0,74		
pronome ‘it’	0,71		
‘be’ como verbo principal	0,71		
subordinação causativa	0,66		
partícula discursiva	0,66		
pronome indefinido	0,62		
atenuador	0,58		
advérbio / qualificador - amplificador	0,56		
pronome relativo	0,55		
pergunta ‘wh’	0,52		
verbo modal de possibilidade	0,50		
coordenação não-frasal	0,48		
oração ‘wh’	0,47		
preposição final	0,43		
(advérbio	0,42)		
(subordinação condicional	0,32)		

Tabela 3: Estrutura do Fator 1 da língua inglesa (BIBER, 1988; 2009).

Estrutura do Fator 2 (BIBER, 1988; 2009)			
Discurso narrativo versus não narrativo			
Polo positivo		**Polo negativo**	
verbo no tempo passado	0,90	(verbo no tempo presente	-0,47)
pronome de terceira pessoa	0,73	(adjetivo em posição atributiva	-0,41)
verbo no aspecto perfeito	0,48	(oração adjetiva reduzida de particípio	-0,34)
verbo público	0,43	(tamanho de palavra	-0,31)
negação sintética	0,40		
oração reduzida de gerúndio	0,39		

Tabela 4: Estrutura do Fator 2 da língua inglesa (BIBER, 1988; 2009).

Estrutura do Fator 3 (BIBER, 1988; 2009)			
Referência dependente de situação versus elaborada			
Polo positivo		**Polo negativo**	
oração wh em posição de objeto	0,63	advérbio de tempo	-0,60
oração wh com preposição inicial	0,61	advérbio de lugar	-0,49
oração wh em posição de sujeito	0,45	advérbios	-0,46
coordenação frasal	0,36		
nominalização	0,36		

Tabela 5: Estrutura do Fator 3 da língua inglesa (BIBER, 1988; 2009).

Estrutura do Fator 4 (BIBER 1988; 2009)	
Argumentação explícita	
Polo positivo	
verbo no infinitivo	0,76
verbo modal de antecipação	0,54
verbo de persuasão	0,49
subordinação condicional	0,47
verbo modal de necessidade	0,46
advérbio encaixado no auxiliar	0,44
(verbo modal de possibilidade	0,37)

Tabela 6: Estrutura do Fator 4 da língua inglesa (BIBER, 1988; 2009).

Estrutura do Fator 5 (BIBER, 1988; 2009)			
Estilo abstrato versus não abstrato			
Polo positivo		**Polo negativo**	
conjuntivos	0,48	(razão forma-ocorrência	-0,31) voz
passiva sem agente	0,43		
orações adjetivas reduzidas de particípio	0,42		
voz passiva com preposição ‘by’	0,41		
modificador pós-nominal	0,40		

outros advérbios subordinativos	0,39
(adjetivo em posição predicativa)	0,31

Tabela 7: Estrutura do Fator 5 da língua inglesa (BIBER, 1988; 2009).

Como será visto na Metodologia deste trabalho, tais fatores foram utilizados na AMD Aditiva Funcional do CoTED, ou seja, os textos do corpus foram mensurados segundo tais fatores, que caracterizam as dimensões da língua inglesa encontradas por Biber (1988).

A seguir, na figura 9, temos o exemplo do mapeamento (alocação e comparação) dos registros analisados por Biber que se encontram na dimensão 1 da língua inglesa, ou seja, foram elencados e ordenados os registros de acordo com seus escores na dimensão em que se encontram. A ordenação dos registros ocorrem de acordo com seus escores médios em cada dimensão (tabela 8), com isso, conseguimos traçar paralelos entre eles, dentro das polaridades positiva e negativa. Os registros com os maiores escores positivos representam o polo positivo e os registros com os maiores escores negativos representam o polo negativo – a ordenação dos demais registros, que ocorrem de acordo com seus escores médios em cada dimensão, são apresentados no anexo 1 desta pesquisa:

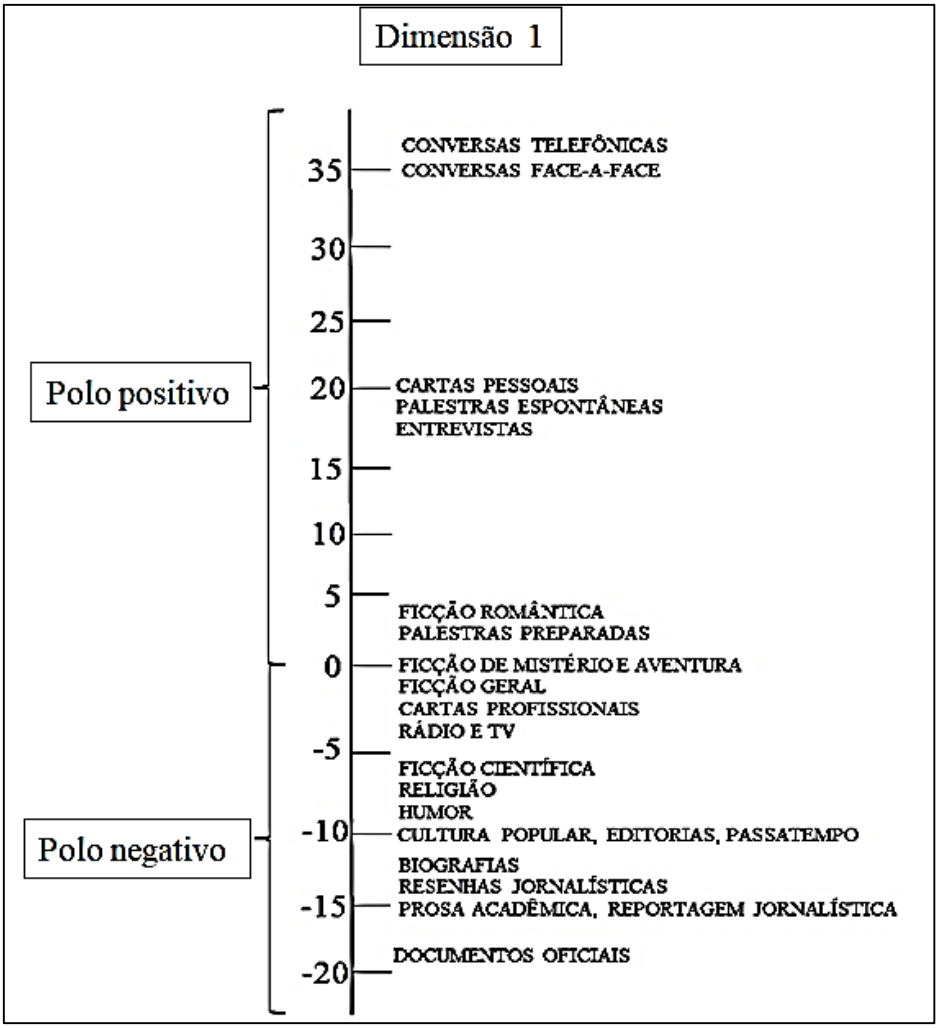


Figura 9: Ordenação dos registros de acordo com seus escores médios na dimensão 1 (BIBER, 1988).

Conforme podemos ver na figura 9, no polo positivo, temos um compartilhamento – em nível de escala para mais ou para menos – de características linguísticas entre os registros que vão desde conversa telefônica até palestras preparadas. Alguns registros estão em um nível mais "equilibrado", que são as ficções de mistério e de aventura. No polo negativo, por sua vez, temos desde ficção geral até documentos oficiais, compartilhando características linguísticas também no nível de escala para mais ou para menos.

Escores médios da dimensão 1 da língua inglesa (BIBER, 1988)		
Variáveis	**N**	**Média**
dim1	Conversas telefônicas	37,2
dim1	Conversas face-a-face	35,3
dim1	Cartas pessoais	19,5
dim1	Palestras espontâneas	18,2
dim1	Entrevistas	17,1

dim1	Ficção romântica	4,3
dim1	Palestras preparadas	2,2
dim1	Ficção de mistério	-0,2
dim1	Ficção de aventura	-0,0
dim1	Ficção geral	-0,8
dim1	Cartas profissionais	-3,9
dim1	Rádio e TV	-4,3
dim1	Ficção científica	-6,1
dim1	Religião	-7,0
dim1	Humor	-7,8
dim1	Cultura popular	-9,3
dim1	Editoriais	-10,0
dim1	Passatempo	-10,1
dim1	Biografias	-12,4
dim1	Resenhas jornalísticas	-13,9
dim1	Prosa acadêmica	-14,9
dim1	Reportagem jornalística	-15,1
dim1	Documentos oficiais	-18,1

Tabela 8: Escores médios da dimensão 1 da língua inglesa (BIBER, 1988).

Na tabela 8, é possível ver quais foram as médias consideradas para cada registro da língua inglesa, possibilitando o mapeamento da dimensão 1 da língua inglesa, conforme figura 9. Tais procedimentos estatísticos, então, fazem parte tanto da AMD Aditiva quanto da AMD Completa do CoTED.

2.10 Análise Multidimensional (AMD) – Cálculo estatístico univariado (ANOVA)

Após a identificação dos fatores, começa a próxima fase da Análise Multidimensional (AMD), que corresponde à interpretação de cada fator para que possamos rotular as dimensões encontradas (BIBER, 1988). Porém, antes de prosseguir com essa análise qualitativa da pesquisa, ainda existe uma última etapa nos procedimentos estatísticos chamada ANOVA Fatorial (cálculo estatístico univariado). Segundo Cantos-Gomez (2019, p. 120), as ANOVAs são usadas para determinar se há diferenças significativas entre as médias de dois ou mais grupos independentes[93]. Com os resultados desses procedimentos, podemos encontrar o percentual de variação dos textos de um corpus explicado por suas variáveis dependentes – cujo

[93] Original: ANOVAs are used to determine whether there are any significant differences between the means of two or more independent groups. (CANTOS-GOMEZ, 2019, p. 120).

valor é medido pela pesquisa – e/ou independentes – cujos valores são anteriores à pesquisa. Também fazemos o uso do chamado Modelo Linear Geral (GLM – *General Linear Model*), um substituto da ANOVA fatorial – essa abordagem é mais geral e suporta o uso de variáveis dependentes categóricas[94].

Com os resultados do primeiro procedimento (ANOVA Fatorial), considerando as variáveis dependentes de um corpus – os fatores do CoTED, nesta pesquisa –, é possível observar o percentual de variação da linguagem ao longo de cada uma das cinco dimensões da língua inglesa encontradas por Biber (1988). Com os resultados do segundo procedimento (GLM), considerando as variáveis independentes de um corpus – no caso do CoTED, as variáveis independentes são "apresentador", que corresponde aos 2.845 apresentadores/ autores dos vídeos/textos das TED Talks analisadas; e "evento", que corresponde aos 412 títulos de eventos TED encontrados (seção 3.1.1) –, é possível verificar se a variação presente é estatisticamente significante, e se os fatores exercem influência em alguma variável independente, permitindo verificar se os agrupamentos das variáveis nos fatores são significativos ou simplesmente causais, ou seja, resultantes da variabilidade natural da amostra. (BERBER SARDINHA; VEIRANO PINTO, 2019, p. 6).

Os itens analisados nas ANOVAs são a razão F, o coeficiente de determinação (R2) – R quadrado – e o valor de p. Segundo Berber Sardinha e Veirano Pinto (2019, p. 6), a razão de F "indica se a variação nos dados é estatisticamente significativa em todos os componentes do corpus", ou seja, a razão F indica a diferença entre os conjuntos em análise medindo a quantidade de variação existente em cada grupo, isso, por meio do cálculo dos escores médios dos textos e dos escores médios de cada dimensão; e quanto maior o valor de F, mais significantes serão os resultados. Quanto ao valor de p, ele é uma estimativa probabilística para verificar se o valor de um teste estatístico ocorre aleatoriamente. Neste caso, para que os resultados sejam considerados significativos, o valor de p deve ser menor que 0,05 (5%), o que significa que a probabilidade de a variação ocorrer por acaso é de uma a cada 20 vezes, ou seja, ela se torna improvável se ocorrer menos de 5% das vezes – para que o valor de p seja inferior a 0,05, o valor de F deve estar acima de 3,35, pois p é o nível de significância de F. O R2, por sua vez, mede a porcentagem de variação capturada em cada dimensão para cada variável, dependente ou independente, analisada. Logo a seguir, temos um exemplo tomado na dimensão 2 da língua inglesa (BIBER, 1988) – tabela 9:

[94] Fonte: http://www.mat.ufrgs.br/~viali/estatistica/mat2282/material/laminaspi/Anova_OWay.pdf

Anova da Dimensão 2 de Biber					
Fator	**Variável**	**F**	**p**	**R2**	**%**
2	Dimensão 2	32,30	<.0001	0,608	60,8

Tabela 9: Anova da Dimensão 2 de Biber (1988).

Na pesquisa de Biber (1988), o resultado da Anova para a dimensão 2 apresentou valor de F é significativo (32,30), o valor de p está abaixo de 0,05 e o valor de R2 (0,608) significa que mais de 60% da variação é explicada pelos registros analisados por Biber. Isso significa que o escore de fator da dimensão 2 é estatstisticamente importante na discriminação entre os registros.

2.11 Análise Multidimensional (AMD) – Interpretação dos fatores

Após a identificação dos fatores, começa a última fase da Análise Multidimensional (AMD), que corresponde à interpretação de cada fator – de modo funcional e polarizado – para que possamos rotular as dimensões encontradas (BIBER, 1988). É por isso que, segundo Biber (1988, p. 28), a AMD é considerada como uma análise de caráter tanto macro quanto micro, ou seja, a análise macroscópica identifica as dimensões de variação entre os textos e especifica todas as relações entre os registros quanto às dimensões; e, por sua vez, a análise microscópica descreve as funções das características linguísticas em relação às situações de fala de cada texto, ou seja, as características linguísticas indicam os componentes específicos de situação, além de suas funções como marcadores de relações dentro de um texto[95]. Em outras palavras, as análises macroscópicas "são efetuadas quando da computação dos fatores", isto é, "as várias análises de cada texto são agrupadas de modo que se possa perceber a das mesmas em nível macro"; quanto às análises microscópicas, "se dão quando da interpretação dos fatores de modo funcional. Nesse nível, são levados em conta cada texto e [ou] cada registro individualmente" (BERBER SARDINHA, 2004, p. 306).

Na interpretação dos fatores – para que possamos nomear as dimensões encontradas – devemos nos basear nos resultados da análise fatorial, que identifica as características linguísticas que coocorrem com frequência nos textos do corpus. Conforme Biber nos explica (1988, p. 91 e 92), na interpretação de um fator, busca-se uma dimensão funcional subjacente

[95] Original: Macroscopic analysis identify the dimensions of variation among texts and specify the overall relations among genres with respect to those dimensions. Microscopic analysis describe the functions of linguistic features in relation to the speech situations of individual texts. Linguistic features mark particular components of the situation, in addition to their functions as markers of relations within a text.

para explicar o padrão de coocorrência entre as características identificadas pelo fator. Em outras palavras, ele nos diz que, um conjunto de características coocorrem frequentemente em textos porque elas estão exercendo alguma função em comum nesses textos, e é nesse ponto que as microanálises das características linguísticas se tornam crucialmente importantes. Assim, as análises funcionais de características individuais em textos permitem identificar a função subjacente compartilhada por um grupo de características em uma análise de fatores. No entanto, embora os padrões de coocorrência sejam quantitativamente oriundos da análise fatorial, a interpretação da dimensão subjacente a um fator é provisória e requer confirmação, semelhante a qualquer outra análise interpretativa. Deste modo, as características linguísticas agrupadas em cada fator podem ser interpretadas como uma dimensão textual, por meio de uma avaliação das funções comunicativas amplamente compartilhadas entre as características, sendo que, a relação complementar entre as cargas positivas e negativas também deve ser considerada na interpretação[96].

Desta forma, com a interpretação dos fatores (ver tabelas 3 a 7), Biber (1988) identificou as dimensões de variação de registros (tanto falados quanto escritos) da língua inglesa, separando seis delas na pesquisa de 1988, se tornando cinco em 2009 – dimensões consideradas até hoje:

1) Dimensão 1: Nomeada como "Produção marcada por envolvimento versus informacional" (*Involved versus Informational Production*) (BIBER, 1988, p. 104-108). Nesta dimensão, existe uma clara divisão entre a fala (interação) e a escrita (informação), sendo ela considerada como uma dimensão universal, pois é comum encontrar sua ocorrência nas AMDs (seção 4.2.2). O polo positivo – que carrega as variáveis que possuem correlações positivas – traz variáveis gramaticais que indicam interação, como verbos mentais (*think, feel, believe* etc.), verbos no presente, enfatizadores, contrações e pronomes de primeira e de segunda pessoa. Os registros que melhor representam o modo de produção com interação são as conversas, tanto ao telefone

[96] Original: In the interpretation of a factor, an underlying functional dimension is sought to explain the co-occurrence pattern among features identified by the factor. That is, it is claimed that a cluster of features co-occur frequently in texts because they are serving some common function in those texts. At this point, micro-analyses of linguistic features become crucially important. Functional analyses of individual features in texts enable identification of the shared function underlying a group of features in a factor analysis. It must be emphasized, however, that while the co-occurrence patterns are derived quantitatively through factor analysis, interpretation of the dimension underlying a factor is tentative and requires confirmation, similar to any other interpretive analysis. […]
The linguistic features grouped on each factor can be interpreted as a textual dimension through an assessment of the communicative functions most widely shared by the features. The complementary relationship between positive and negative loadings must also be considered in the interpretation.

quanto face a face. O polo negativo – que carrega as variáveis de correlação negativa – possui variáveis gramaticais que sinalizam densidade de informação, assim como substantivos, preposições, tamanho de palavras e adjetivos atributivos. Os registros que melhor representam a produção informacional são os documentos oficiais, a reportagem jornalística e a prosa acadêmica (BERBER SARDINHA, 2004, p. 312).

2) Dimensão 2: Nomeada como "Discurso narrativo versus não narrativo" (*Narrative Discourse*) (BIBER, 1988, p. 108 e 109). Nesta dimensão, existe uma divisão clara entre o discurso narrativo e o não narrativo. O polo positivo – que carrega as variáveis que possuem correlações positivas – traz variáveis gramaticais que indicam narração, como verbos no passado, pronomes de terceira pessoa, verbos dicendi (*reply, say, ask*, etc.) e negação sintética (contração do *not*, por exemplo). Os registros que melhor demonstram uma preocupação com a narração são os registros de ficção. O polo negativo – que carrega as variáveis de correlação negativa – possui variáveis gramaticais que sinalizam discurso não narrativo, assim como verbos no presente e adjetivos em posição atributiva. Os registros que melhor exprimem uma orientação não-narrativa são os registros de rádio e TV, passatempos e documentos oficiais (BERBER SARDINHA, 2004, p. 312).

3) Dimensão 3: Nomeada como "Referência dependente de situação versus elaborada" (*Situation-Dependent versus Elaborated Reference*) (BIBER, 1988, p. 110). Nesta dimensão, existe uma divisão clara entre referência explícita nos textos e referência dependente da situação, externa aos textos. O polo positivo – que carrega as variáveis que possuem correlações positivas – traz variáveis gramaticais que indicam referência explícita, como orações relativas, coordenação frasal e nominalizações. Os registros que apresentam referência explícita em maior grau são os documentos oficiais, cartas profissionais, resenhas jornalísticas e prosa acadêmica. O polo negativo – que carrega as variáveis de correlação negativa – possui variáveis gramaticais que sinalizam referência dependente da situação, assim como advérbios de tempo e lugar, que são geralmente utilizados para fazer referência a algo exterior aos textos. Os registros de rádio e TV, conversas telefônicas, face a face e ficção romântica exprimem referência dependente de situação (BERBER SARDINHA, 2004, p. 312).

4) Dimensão 4: Nomeada como "Argumentação explícita" (*Overt Expression of Persuasion*) (BIBER, 1988, p. 111). Nesta dimensão, temos a persuasão explícita marcada nos textos. O polo positivo – que carrega as variáveis que possuem correlações positivas – traz variáveis

gramaticais que indicam persuasão, como verbos suasivos (de persuasão: *command*, *stipulate*, *order* etc.), verbos modais e verbos no infinitivo. Os registros de caráter mais persuasivo são as cartas profissionais, os editoriais jornalísticos e a ficção romântica. O polo negativo – que carrega as variáveis de correlação negativa – não foi carregado. Mas é possível definir os registros nos quais a persuasão é menos explícita, que são os de rádio e TV, resenhas jornalísticas e ficção de aventura (BERBER SARDINHA, 2004, p. 312-313).

5) Dimensão 5: Nomeada como "Estilo abstrato versus não abstrato" (*Impersonal Style*) (BIBER, 1988, p. 111-113). Nesta dimensão, os textos veiculam informação mais abstrata versus menos abstrata. O polo positivo – que carrega as variáveis que possuem correlações positivas – traz variáveis gramaticais que indicam informação mais abstrata, como conjunções, voz passiva (sem sujeito; com *by*), orações reduzidas de particípio e adjetivos em posição predicativa. Os registros que veiculam informação mais abstrata são os acadêmicos, os documentos oficiais e os religiosos. O polo negativo – que carrega as variáveis de correlação negativa – carregou somente uma variável. Mas é possível definir os registros nos quais as informações são menos abstratas, que são as conversas telefônicas, face a face e ficção romântica. (BERBER SARDINHA, 2004, p. 313).

2.12 Análise Multidimensional (AMD) – Algumas considerações

Segundo Berber Sardinha (2004, p. 305), podemos resumir a Análise Multidimensional (AMD) em três etapas básicas: 1) a primeira seria de caráter preliminar, compreendendo a revisão da literatura em busca de traços linguísticos relevantes, a coleta do corpus e a codificação dos textos de acordo com o elenco de características linguísticas selecionadas; 2) a segunda fase seria a análise fatorial, durante a qual é feito um agrupamento das características linguísticas em fatores, e a interpretação funcional desses fatores, a fim de descobrir um traço comunicativo dominante subjacente ao fator, dando origem às dimensões; e 3) a terceira etapa seria o cálculo de escores de cada texto em relação a cada fator, e a interpretação das dimensões à luz dos textos que as compõem.

Ademais, outra característica importante da AMD é que essa abordagem metodológica também possui um caráter cumulativo, pois permite a comparação com outras análises feitas posteriormente, podendo ser elas de larga escala ou mesmo individuais. Também, ela consegue acomodar diversos traços linguísticos, não somente gramaticais e lexicais, como traços discursivos, por exemplo. Tais adaptações trazem a possibilidade de se encontrar as dimensões

de variação de vários registros na língua inglesa e até em outros idiomas. Segundo Biber (1988, p. 200), embora seu estudo tenha começado como uma investigação da fala e da escrita da língua inglesa, a análise final apresenta uma descrição geral das relações entre os textos na língua inglesa e, portanto, pode ser utilizada como base para a investigação de várias outras questões relacionadas. Conforme ele nos explica, uma vez que os textos utilizados em seu estudo abrangem muitos dos possíveis tipos de discurso da língua inglesa, e as características linguísticas utilizadas abrangem muitas das funções comunicativas marcadas por características superficiais do idioma, as dimensões resultantes não são estritamente parâmetros de variação entre a fala e a escrita, mas são parâmetros fundamentais da variação linguística entre os textos na língua inglesa. Deste modo, é por isso que as dimensões podem ser utilizadas para especificar as relações entre diferentes registros, como por exemplo, textos de diferentes períodos históricos, textos de diferentes dialetos sociais ou textos de aprendizes de diferentes habilidades. Portanto, é por isso que Biber (1988) afirma que a abordagem geral da Análise Multitraço e Multidimensional de Variação de Registro (MF/MD) da variação textual pode ser usada para investigar uma série de outras questões do discurso, podendo até ser usada para especificar as relações entre textos em outras línguas e fornecer uma base para comparações interlinguísticas entre os tipos de texto[97].

Logo em seguida, será apresentada a metodologia aplicada para chegar nos resultados preliminares da pesquisa. Logo após, tais resultados preliminares serão analisados para se chegar aos resultados finais desta pesquisa.

3. Metodologia de pesquisa

O presente capítulo está dividido em duas seções. Na primeira seção encontramos os procedimentos de compilação, desenho, coleta e etiquetagem do corpus CoTED. Na segunda, temos o primeiro passo metodológico de análise da AMD, a Análise Multidimensional

[97] Original: Although this study began as an investigation of speech and writing, the final analysis presents an overall description of the relations among texts in English, and it can therefore be used as a basis for the investigation of several related issues. That is, since the texts used in this study cover many of the possible discourse types in English, and the linguistic features used here cover many of the communicative functions marked by surface features in English, the resulting dimensions are not strictly parameters of variation between speech and writing; rather they are fundamental parameters of linguistic variation among English texts. As such, the dimensions can be used to specify the relations among many different types of texts, for example, texts from different historical periods, texts from different social dialects, or texts from student writers of differing abilities. Similarly, the general MF/MD approach to textual variation, which I apply here to the relations among spoken and written texts in English, can be used to investigate a number of other discourse issues. In particular, this approach can be used to specify the relations among texts in other languages and provide a basis for cross-linguistic comparisons of text types.

Funcional Aditiva, a qual busca saber como o corpus CoTED (até o momento, considerado como 1 registro com 3 sub-registros: TED tradicional, TEDx e TED-Ed) se encaixa nas 5 dimensões de variação da língua inglesa (BIBER, 1988; 2009). Em seguida, temos o segundo passo metodológico de análise da AMD, a Análise Multidimensional Funcional Completa, a qual busca extrair as dimensões de variação funcionais do CoTED. Também, é discutido como se dá o percentual de variação da linguagem das TED Talks ao longo de cada uma das cinco dimensões da língua inglesa encontradas por Biber (1988), por meio da Análise de Variância (ANOVA); e como se dá a variação multidimensional funcional em termos das variáveis independentes "apresentador" e "evento", por meio do Modelo Linear Geral (GLM).

3.1 Corpus TED Talks (CoTED)

A presente pesquisa teve o intuito de compilar um corpus que atendesse aos procedimentos teórico-metodológicos na construção de corpora propostos por Biber (1988), e complementados por Egbert (2019), compreendendo os textos com as transcrições dos vídeos das TED Talks em inglês – divididos, a princípio, como um tipo específico de categoria ou registro – TED Talks Geral – em contraposição às demais três categorias ou sub-registros – TED tradicional, TEDx e TED-Ed. Desta forma, buscou-se seguir os nove passos propostos por Egbert (2019), baseados nas diretrizes do processo cíclico de quatro passos de Biber (1988), na construção, desenho e coleta de corpus, que são:

1. Estabelecer (e projetar) os objetivos e o planejamento da pesquisa.
2. Definir o domínio-alvo (ou população).
3. Desenhar o corpus.
4. Coletar a amostra.
5. Fazer a anotação do corpus.
6. Avaliar a representatividade do domínio-alvo (ou população).
7. Avaliar a representatividade linguística.
8. Repetir passos 3-5, se necessário.
9. Criar relatório.

Os passos 1 ao 3 já foram previamente abordados neste trabalho (seção 2.3). Quanto ao passo 1, o objetivo principal desta pesquisa é responder as seguintes perguntas sobre a

linguagem verbal das TED Talks:

1) Como o corpus das TED Talks (CoTED) se encaixa nas dimensões de variação da língua inglesa encontradas por Biber (1988)?

2) Quais são as dimensões de variação do corpus das TED Talks (CoTED) sob a perspectiva da AMD Funcional Completa?

3) Como se dá a variação multidimensional funcional em termos das variáveis independentes "apresentador" e "evento" do corpus das TED Talks (CoTED)?

Para responder à primeira pergunta, foi feita a AMD Funcional Aditiva e, para responder as demais perguntas, foi feita a AMD Funcional Completa. Referente ao passo 2, ao definir o domínio-alvo desta pesquisa, foram consideradas as transcrições dos vídeos em inglês das TED Talks em geral, de 1984 até o final de 2019, e divididos em três categorias – TED tradicionais, TEDx e TED-Ed. Com o passo 3, classificamos o corpus desta pesquisa – de 3.411 transcrições de vídeos TED – como oral retratado em formato de texto escrito, falado, sincrônico, contemporâneo, de amostragem, estático, especializado, de língua nativa (no caso, inglês) e de estudo. Desta forma, focaremos nos passos 4 ao 9, a partir daqui.

Quanto à coleta do corpus TED Talks (CoTED), inicialmente, foram manualmente coletados os endereços no formato .html dos vídeos TED Talks em inglês, disponíveis no site oficial das TED Talks[98] (de 1984 até setembro de 2019), gerando um total de 3.979 arquivos. Os textos das transcrições dos vídeos foram extraídos via linguagem de scripts[99], sendo salvos no formato .txt. Outros scripts foram utilizados para organizar e "limpar" tais textos, retirando-se códigos como seus endereços eletrônicos e etc. Também foi feita a conferência e limpeza manual dos arquivos – alguns estavam com nenhuma marcação de tempo, sem informações sobre as visualizações, sem transcrições e até em outros idiomas (como espanhol, japonês, francês e português), mesmo estando na categoria de vídeos em inglês, segundo o site da TED. Todos os textos apresentavam formatação e espaçamentos que poderiam dificultar sua etiquetagem, por isso, foi feita uma revisão manual via *Notepad++*. Também foi feita a substituição dos símbolos musicais "♫" e "♪" por "(*Singing*)" e a exclusão de vídeos exclusivamente de performances artísticas, musicais, teatrais etc. Desta forma, dos 3.979

[98] https://www.ted.com/talks

[99] Linguagem de Script: Uma linguagem de programação executada dentro de um programa; Usada para automatizar comandos que seriam feitos por uma pessoa; São "interpretadas", ou seja, um interpretador traduz o código para linguagem de máquina; (Fonte: https://docente.ifrn.edu.br/pedrobaesse/disciplinas/programacao-web/material-de-aula/aula-02-linguagem-de-script/at_download/file). Obs: Os scripts foram elaborados pelo professor Tony Berber Sardinha.

arquivos, restaram 3.411. As transcrições foram separadas de acordo com cada vídeo considerado, e os metadados foram salvos em uma tabela no formato .xls (serão mencionados em mais detalhes logo abaixo). A nomeação dos textos foi padronizada seguindo a seguinte ordem: 1) T inicial para todos; 2) numeração de quatro dígitos de acordo com o número de identificação do arquivo original (.html); 3) abreviação do nome do autor do texto/vídeo; 4) abreviação do tipo de TED. Exemplos:

- Para TED tradicional: T0007_AAB_TED19 – esse é um vídeo TED tradicional, cujo arquivo de transcrição coletado está na 7° posição na tabela .xls, seu autor tem as iniciais A, A, e B em seu nome, e o vídeo pertence ao evento chamado TED2019.
- Para TEDx: T0044_DP_TEDXPSU – esse é um vídeo TEDx, cujo arquivo de transcrição coletado está na 44° posição na tabela .xls, seu autor tem as iniciais D e P em seu nome, e o vídeo pertence ao evento chamado TEDxPSU.
- Para TED-Ed: T0001_AG_TEDED – esse é um vídeo TED-Ed, cujo arquivo de transcrição coletado está na 1° posição na tabela .xls, seu autor tem as iniciais A e G, e o vídeo pertence à categoria dos vídeos educacionais chamado TED-Ed.

A princípio, foram anotados todos os textos das TED Talks – atribuição das características linguísticas – e foi feita a análise fatorial de todos eles em conjunto. Contudo, com os resultados obtidos na primeira análise fatorial, também foi feita a análise fatorial dos três grupos aqui elencados – TED tradicional, TEDx e TED-Ed – de forma a comparar os resultados. Os resultados e sua análise são apresentados na seção 4 desta pesquisa.

3.1.1 Metadados

Conforme ilustrado no quadro 4 logo abaixo, o CoTED é formado de 3.411 transcrições de vídeos TED em inglês – relembrando que a linguagem verbal das TED Talks pode ser considerada como oral retratada em formato de texto escrito –, referentes aos anos de 1984 até o final de 2019. Segundo a divisão feita aqui neste trabalho – com relação ao tipo de vídeo TED –, temos o TED Tradicional – evento realizado pela própria TED, cujo "carro-chefe" são as palestras; o TEDx – evento realizado por entidades autônomas em todo o mundo, cujo "carro-chefe" são as palestras; o TED-Ed – vídeos educacionais animados feitos, segundo o estilo TED, por palestrantes, professores, designers, pesquisadores, jornalistas etc.; e o TED Geral (o CoTED em si) – que corresponde à soma do TED Tradicional, TEDx e TED-Ed.

Do total de 3.411 TED Talks, 2.400 são classificados como TED tradicionais, 603 como TEDx e 408 como TED-Ed. Desse total de vídeos, ou melhor, de textos, temos o total de 6.370.138 palavras. O valor mínimo de palavras encontrado é de 378 (*Photos from a storm chaser* – TED2013 "TED tradicional"), o valor médio de palavras é de 1.868 (*Sculpted space, within and without* – TEDGlobal 2012 "TED tradicional") e o valor máximo de palavras é de 9.186 (*Nationalism vs. Globalism* – TED Dialogues (2017) "TED tradicional"). O tempo mínimo em minutos é por volta de dois minutos (*How the button changed fashion* – Small Thing Big Idea (2018) "TED tradicional"), o tempo médio é por volta de 12 minutos (*Why we should build wooden skyscrapers* – TED2013 "TED tradicional") e o tempo máximo chegou a mais de uma hora (*Nationalism vs. Globalism* – TED Dialogues (2017) "TED tradicional"). Quanto ao total de visualizações, desde 2006 até o final de 2019, temos 7.309.610.334. Desse total, temos o valor mínimo de visualizações, sendo 10.667 (*Community health heroes* – Torchbearers (2018) "TED tradicional"). A média de visualizações é de 2.143.283 (*The Happy Planet Index* – TEDGlobal (2010) "TED tradicional"). E o número máximo de visualizações é de 62.542.472 (*Do schools kill creativity?* – TED2006 "TED tradicional"). Quanto aos números referentes ao gênero biológico do apresentador/criador do texto da transcrição, tivemos 1.261 vídeos TED Talks apresentados somente por mulheres (quase 37% do total), 2.133 somente por homens (mais de 62% do total) e somente 17 por ambos os gêneros (menos de 0,5% do total).

Acredita-se que os dados aqui apresentados sejam não somente formas de validar o corpus da presente pesquisa como representativo do domínio-alvo – a linguagem verbal das TED Talks – como também formas de ajudar a analisar os resultados obtidos. Desta forma, também, pode-se afirmar que o presente corpus segue os parâmetros de representatividade da variedade linguística aqui analisada – a linguagem verbal das TED Talks –, pois é uma amostragem significativa da variedade textual pesquisada, contendo características linguísticas articuladas ao contexto de uso, e atendendo às funções comunicativas específicas do contexto estudado (BIBER, 1993; EGBERT, 2019).

Corpus TED Talks (CoTED)	
Total de vídeos	3.411
Total de palavras	6.370.138
N° mínimo de palavras	378
N° médio de palavras	1.868
N° máximo de palavras	9.186
Tempo mínimo	140s (+ 2min.)
Tempo médio	742s (+ 12min.)
Tempo máximo	3.608s (+ 60min.)
Total de visualizações	7.309.610.334
Mínimo de visualizações	10.667
Média de visualizações	2.143.283
Máximo de visualizações	62.542.472
Gênero biológico (feminino)	1.261
Gênero biológico (masculino)	2.133
Gênero biológico (ambos)	17
TED tradicional	2.400
TEDx	603
TED-Ed	408
Apresentador/criador	2.845

Quadro 4: Corpus das TED Talks (metadados).

Outros valores de metadados encontrados são referentes ao número de vídeos por ano, a quantidade de *tags* (anteriormente mencionadas como tópicos – seção 2.1) utilizadas na pesquisa do site das TED Talks, a variedade de nomes dos eventos TED e a quantidade de vídeos TED por apresentador/criador (além de verificar a presença de algumas celebridades das mais variadas áreas). No quadro 5, é interessante ver a progressão dos números de vídeos TED Talks que foram sendo postados on-line ao longo dos anos – fato que deve continuar até o presente momento:

anos	total de vídeos TED
2019	197
2018	336
2017	370
2016	313
2015	286
2014	282
2013	323
2012	271

2011	237
2010	219
2009	201
2008	74
2007	101
2006	41
2005	61
2004	30
2003	31
2002	25
2001	4
1998	6
1994	1
1990	1
1984	1

Quadro 5: Vídeos TED Talks ao longo dos anos.

No quadro 6, temos os 50 primeiros tópicos (*tags*) – dentre os 440 encontrados no site oficial[100] – que podem ser atribuídos aos vídeos das TED Talks. Por conta da grande variedade de tópicos que podem ser atribuídos a cada vídeo, sem ter uma definição de qual seria o principal, foram atribuídas as três classificações aqui nesta pesquisa – TED tradicional, TEDx e TED-Ed.

#	**Ordem alfabética (até posição 50 de um total de 440)**	**quantidade de tags (tópicos)**
1	science	886
2	technology	846
3	global issues	484
4	society	481
5	culture	476
6	social change	437
7	design	423
8	innovation	331
9	humanity	332
10	health	327
11	future	291
12	history	283
13	animation	277

[100] https://www.ted.com/talks

14	biology	253
15	entertainment	228
16	communication	239
17	community	241
18	ted-ed	239
19	health care	228
20	tedx	215
21	medicine	218
22	business	213
23	education	208
24	creativity	205
25	personal growth	208
26	nature	205
27	environment	199
28	invention	192
29	collaboration	186
30	art	182
31	economics	186
32	women	185
33	identity	185
34	psychology	174
35	politics	168
36	brain	151
37	life	146
38	medical research	145
39	engineering	137
40	storytelling	135
41	inequality	135
42	activism	131
43	war	131
44	computers	131
45	human body	131
46	sustainability	125
47	government	121
48	public health	121
49	children	116
50	disease	119

Quadro 6: Os 50 primeiros tópicos (*tags*) das TED Talks.

No quadro 7, temos os primeiros 50 títulos de eventos TED – dentre os 412 encontrados no site. Por conta da quantidade e variedade de títulos, foi decidido que as três classificações aqui atribuídas – TED tradicional, TEDx e TED-Ed – ainda seriam as melhores

opções.

#	**Títulos dos eventos TED Talks (50 de 412)**
1	Arbejdsglaede Live
2	Business Innovation Factory
3	Chautauqua Institution
4	DIY Neuroscience
5	DLD 2007
6	EG 2007
7	EG 2008
8	Full Spectrum Auditions
9	Global Witness
10	INK Conference
11	LIFT 2007
12	Mission Blue II
13	Mission Blue Voyage
14	Serious Play 2008
15	Skoll World Forum 2007
16	Small Thing Big Idea
17	Taste3 2008
18	TED Dialogues
19	TED Fellows 2015
20	TED Fellows Retreat 2013
21	TED Fellows Retreat 2015
22	TED in the Field
23	TED Prize Wish
24	TED Residency
25	TED Salon
26	TED Salon Brightline Initiative
27	TED Salon Doha Debates
28	TED Salon Optum
29	TED Salon Samsung
30	TED Salon The Macallan
31	TED Salon U.S. Air Force
32	TED Salon Verizon
33	TED Salon Zebra Technologies
34	TED Senior Fellows at TEDGlobal 2010
35	TED Studio
36	TED Talks Education
37	TED Talks India
38	TED Talks Live
39	TED@Bangalore
40	TED@BCG Berlin

41	TED@BCG London
42	TED@BCG Milan
43	TED@BCG Paris
44	TED@BCG San Francisco
45	TED@BCG Singapore
46	TED@BCG Toronto
47	TED@Cannes
48	TED@IBM
49	TED@Intel
50	TED@Johannesburg

Quadro 7: 50 títulos de eventos TED Talks.

No quadro 8, temos os 50 primeiros nomes dos apresentadores das TED Talks, segundo a quantidade de apresentações – apresentadores e não somente palestrantes, pois também são consideradas as TED-Eds.

#	**Nomes**	**n° palestras - decrescente (até posição 50 de 2.845)**
1	Alex Gendler	22
2	Iseult Gillespie	17
3	Daniel Finkel	10
4	Emma Bryce	9
5	Greg Gage	9
6	Hans Rosling	9
7	Juan Enriquez	8
8	Dean Kamen	6
9	Elizabeth Cox	6
10	Jonathan Haidt	6
11	Joy Lin	6
12	Chris Anderson	5
13	Clay Shirky	5
14	Colm Kelleher	5
15	Dan Ariely	5
16	Jacqueline Novogratz	5
17	Julian Treasure	5
18	Marco Tempest	5
19	Mia Nacamulli	5
20	Nicholas Negroponte	5
21	Rives	5

22	AJ Jacobs	4
23	Al Gore	4
24	Barry Schwartz	4
25	Bill Gates	4
26	Christina Greer	4
27	Dan Dennett	4
28	Dan Finkel	4
29	Danny Hillis	4
30	David Pogue	4
31	Eleanor Nelsen	4
32	Eric Liu	4
33	John McWhorter	4
34	Jonathan Drori	4
35	Kaitlyn Sadtler	4
36	Kevin Kelly	4
37	Lucianne Walkowicz	4
38	Margaret Heffernan	4
39	Michael Green	4
40	Paola Antonelli	4
41	Pico Iyer	4
42	Robert Full	4
43	Rose Eveleth	4
44	Sajan Saini	4
45	Sarah Parcak	4
46	Stefan Sagmeister	4
47	Steven Johnson	4
48	Steven Pinker	4
49	Tom Wujec	4
50	Adam Savage	3

Quadro 8: 50 apresentadores TED Talks.

É interessante perceber que existe a possibilidade de uma pessoa poder ter mais de uma apresentação postada no site. No quadro 9, temos nomes de algumas pessoas famosas, que se fazem presente nas TED Talks (uma das marcas dos eventos TED), trazendo (provavelmente) uma visão de credibilidade.

Nomes	**Número de vídeos**
Chris Anderson	5
Al Gore	4
Bill Gates	4

Bono	2
Elon Musk	2
Bill Clinton	1
Sua Santidade o Papa Francisco	1
Sua Santidade o Karmapa	1
Greta Thunberg	1
John Legend	1
Monica Lewinsky	1
Sting	1

Quadro 9: Personalidades nas TED Talks.

3.1.2 Etiquetagem

Conforme anteriormente explicado, Biber (1988) criou um programa ou etiquetador, o *Biber Tagger* (figura 10) – (seção 2.5) –, que abrange as características morfossintáticas, semânticas e de marcação de posicionamento – programa que foi aprimorado ao longo do tempo, aumentando o número de características linguísticas que podem ser analisadas (RESENDE, 2019). O etiquetador "serve para inserir automaticamente, no corpus, códigos que indicam a classe gramatical de cada palavra [com precisão média de 95%]" (BERBER SARDINHA, 2004, p. 113). Para a presente pesquisa, foram computadas 128 características gramaticais, semânticas e referentes aos marcadores de posicionamento; compreendendo nove categorias morfológicas: adjetivos, advérbios, substantivos, verbos, conjunções, preposições, pronomes, orações complementares e subordinadas.

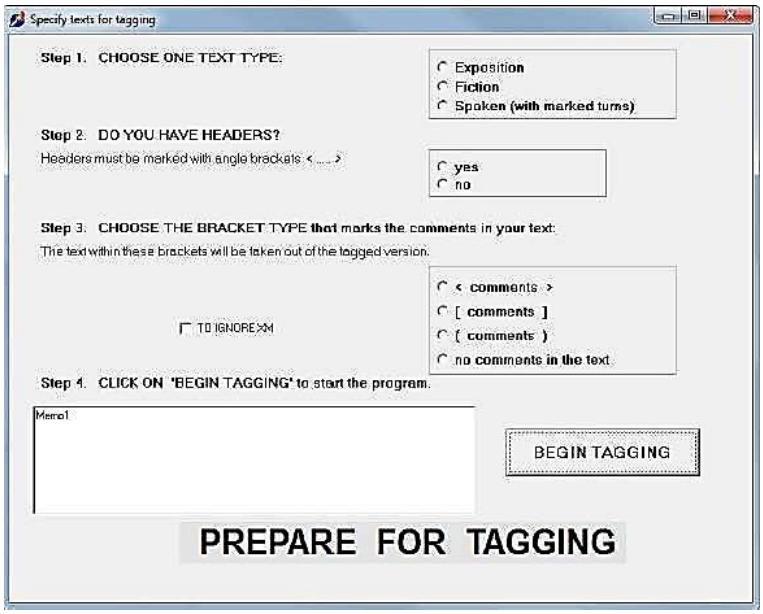


Figura 10: Interface do *Biber Tagger*.

Na figura 11 temos o exemplo de um trecho de texto do corpus CoTED etiquetado em arquivo no formato .txt – *Inside the bizarre world of internet trolls and propagandists* (TED2019 – TED tradicional). Nesse exemplo, lemos o texto na vertical, em que, cada palavra é posicionada em uma linha, seguida pelo símbolo ^, pela etiqueta (característica linguística), pelo sinal = e pelo termo etiquetado. No caso das marcas de pontuação, elas são codificadas como EXTRAWORD.

```
I ^pp1a+pp1+++=I
spent ^vbd+++xvbnx+=spent
the ^ati++++=the
past ^jjb++++=past
three ^cd++++=three
years ^nns++++=years
talking ^vwbg+++xvbg+=talking
to ^in++++=to
some ^dti++++=some
of ^in++++=of
the ^ati++++=the
worst ^jjt+atrb+++=worst
people ^nns++++=people
on ^in++++=on
the ^ati++++=the
internet ^nn+++??+=internet.
. ^zz++++=EXTRAWORD
```

Figura 11: Trecho de texto etiquetado pelo *Biber Tagger*.

Ao analisarmos as características linguísticas atribuídas nesse trecho acima (figura 11), ou seja, na frase "*I spent the past three years talking to some of the worst people on the internet*", obtivemos as seguintes etiquetas:

- pp1a+pp1+++ (pronome de primeira pessoa em posição de sujeito)
- vbd+++xvbnx+ (verbo no passado)
- ati++++ (artigo definido singular)
- jjb++++ (adjetivo)
- cd++++ (número cardinal)
- nns++++ (substantivo no plural)

- vwbg+++xvbg+ (verbo no presente contínuo modificador pós-nominal)
- in++++ (preposição)
- dti++++ (determinante no singular ou plural)
- in++++ (preposição)
- ati++++ (artigo definido singular)
- jjt+atrb+++ (adjetivo atributivo)
- nns++++ (substantivo comum no plural)
- in++++ (preposição)
- ati++++ (preposição)
- nn+++??+ (substantivo comum no singular)
- zz++++ (pontuação)

A tabela completa com as descrições das etiquetas do *Biber Tagger* encontra-se no anexo 2 deste trabalho. As contagens das características linguísticas foram automaticamente normalizadas pelo *Biber Tag Count* (com base em mil palavras) – (seção 2.5) – a fim de permitir a comparação direta na mesma base de contagens entre textos de tamanhos diferentes:

								t0001_ag_teded.txt				34.8	4.9	677
7.4	0.0	0.0	29.5	0.0	0.0	0.0	3.0	4.4	16.2	0.0	0.0	0.0	0.0	0.0
0.0	0.0	1.5	25.1	0.0	0.0	317.6	115.2	90.1	62.0	16.2	8.9	1.5	1.5	1.5
0.0	0.0	0.0	8.9	4.4	25.1	5.9	5.9	0.0	0.0	0.0	3.0	3.0	14.8	3.0
3.0	4.4	3.0	1.5	4.4	7.4	29.5	20.7	0.0	3.0	115.2	20.7	0.0	113.7	5.9
1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.5	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.4	0.0
3.0	0.0	0.0	3.0	1.5	0.0	1.5	4.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	3.0	0.0	0.0	0.0	0.0	0.0	16.2	11.8	3.0	16.2	31.0	4.4	16.2	1.5
0.0	1.5	4.4	0.0	0.0	5.9	10.3	22.2	7.4	16.2	0.0	8.9	5.9	1.5	0.0
	-18.87		-0.99		-2.68		-4.02		3.68	1.5	0.0	1.5	0.0	0.0

Tabela 10: Exemplo de resultados com o *Biber Tag Count*.

A tabela 10 traz um exemplo de frequência normalizada do arquivo 0001 do corpus CoTED (T0001_AG_TEDED). Como próximo passo, por meio do programa *SAS OnDemand for Academics*[101] (figura 12), as características linguísticas identificadas foram computadas e salvas em arquivos do tipo planilha no formato .xls (figura 13):

[101] https://welcome.oda.sas.com/login

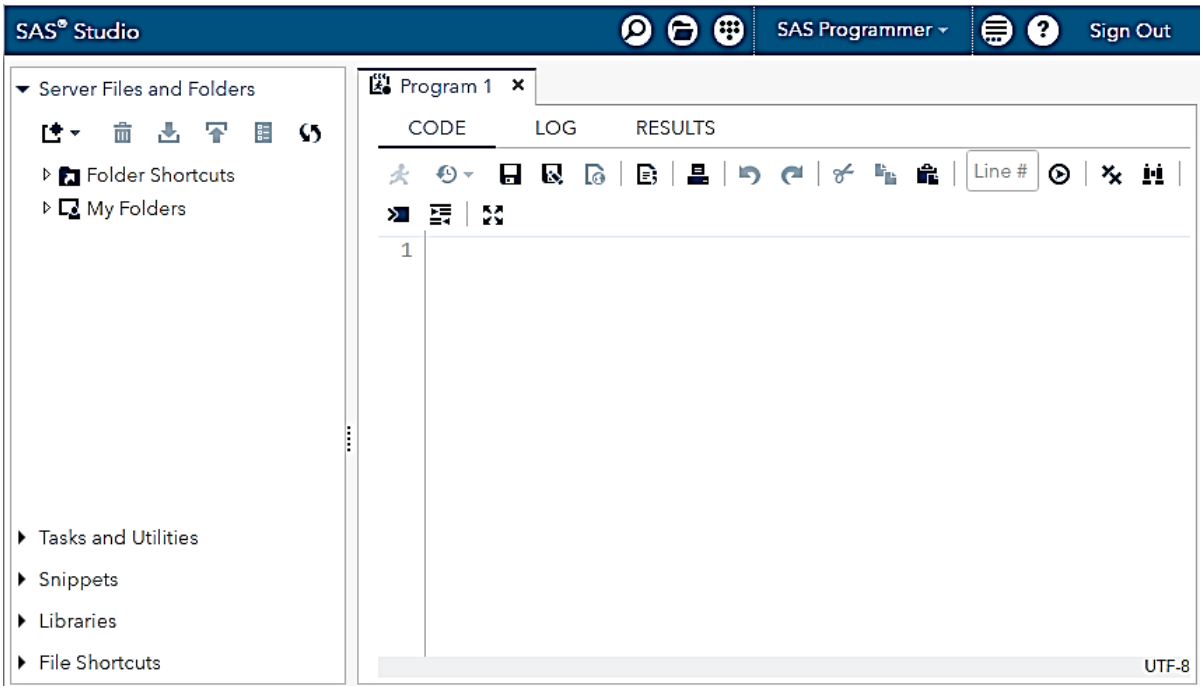
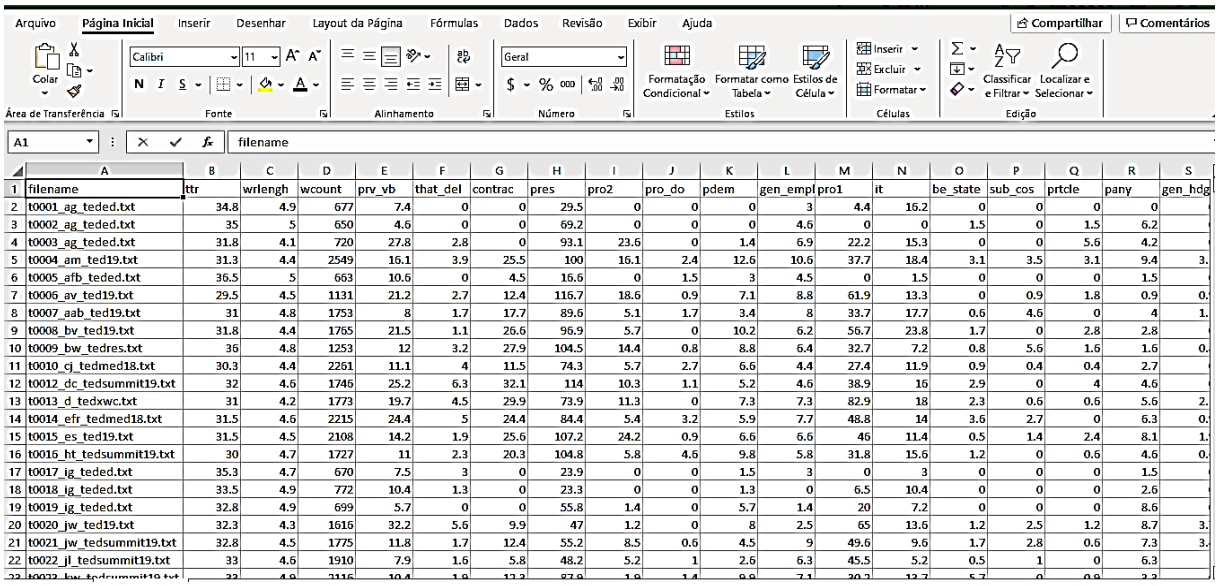


Figura 12: Interface do programa *SAS OnDemand for Academics.*



	filename	ttr	wrlengh	wcount	prv_vb	that_del	contrac	pres	pro2	pro_do	pdem	gen_empl	pro1	it	be_state	sub_cos	prtcle	pany	gen_hdg
2	t0001_ag_teded.txt	34.8	4.9	677	7.4	0	0	29.5	0	0	0	3	4.4	16.2	0	0	0	0	
3	t0002_ag_teded.txt	35	5	650	4.6	0	0	69.2	0	0	0	4.6	0	0	1.5	0	1.5	6.2	
4	t0003_ag_teded.txt	31.8	4.1	720	27.8	2.8	0	93.1	23.6	0	1.4	6.9	22.2	15.3	0	0	5.6	4.2	
5	t0004_am_ted19.txt	31.3	4.4	2549	16.1	3.9	25.5	100	16.1	2.4	12.6	10.6	37.7	18.4	3.1	3.5	3.1	9.4	3.
6	t0005_afb_teded.txt	36.5	5	663	10.6	0	4.5	16.6	0	1.5	3	4.5	0	1.5	0	0	0	1.5	
7	t0006_av_ted19.txt	29.5	4.5	1131	21.2	2.7	12.4	116.7	18.6	0.9	7.1	8.8	61.9	13.3	0	0.9	1.8	0.9	0.
8	t0007_aab_ted19.txt	31	4.8	1753	8	1.7	17.7	89.6	5.1	1.7	3.4	8	33.7	17.7	0.6	4.6	0	4	1.
9	t0008_bv_ted19.txt	31.8	4.4	1765	21.5	1.1	26.6	96.9	5.7	0	10.2	6.2	56.7	23.8	1.7	0	2.8	2.8	
10	t0009_bw_tedres.txt	36	4.8	1253	12	3.2	27.9	104.5	14.4	0.8	8.8	6.4	32.7	7.2	0.8	5.6	1.6	1.6	0.
11	t0010_cj_tedmed18.txt	30.3	4.4	2261	11.1	4	11.5	74.3	5.7	2.7	6.6	4.4	27.4	11.9	0.9	0.4	0.4	2.7	
12	t0012_dc_tedsummit19.txt	32	4.6	1746	25.2	6.3	32.1	114	10.3	1.1	5.2	4.6	38.9	16	2.9	0	4	4.6	
13	t0013_d_tedxwc.txt	31	4.2	1773	19.7	4.5	29.9	73.9	11.3	0	7.3	7.3	82.9	18	2.3	0.6	0.6	5.6	2.
14	t0014_efr_tedmed18.txt	31.5	4.6	2215	24.4	5	24.4	84.4	5.4	3.2	5.9	7.7	48.8	14	3.6	2.7	0	6.3	0.
15	t0015_es_ted19.txt	31.5	4.5	2108	14.2	1.9	25.6	107.2	24.2	0.9	6.6	6.6	46	11.4	0.5	1.4	2.4	8.1	1.
16	t0016_ht_tedsummit19.txt	30	4.7	1727	11	2.3	20.3	104.8	5.8	4.6	9.8	5.8	31.8	15.6	1.2	0	0.6	4.6	0.
17	t0017_ig_teded.txt	35.3	4.7	670	7.5	3	0	23.9	0	0	1.5	3	0	3	0	0	0	1.5	
18	t0018_ig_teded.txt	33.5	4.9	772	10.4	1.3	0	23.3	0	0	1.3	0	6.5	10.4	0	0	0	2.6	
19	t0019_ig_teded.txt	32.8	4.9	699	5.7	0	0	55.8	1.4	0	5.7	1.4	20	7.2	0	0	0	8.6	
20	t0020_jw_ted19.txt	32.3	4.3	1616	32.2	5.6	9.9	47	1.2	0	8	2.5	65	13.6	1.2	2.5	1.2	8.7	3.
21	t0021_jw_tedsummit19.txt	32.8	4.5	1775	11.8	1.7	12.4	55.2	8.5	0.6	4.5	9	49.6	9.6	1.7	2.8	0.6	7.3	3.
22	t0022_jl_tedsummit19.txt	33	4.6	1910	7.9	1.6	5.8	48.2	5.2	1	2.6	6.3	45.5	5.2	0.5	1	0	6.3	

Figura 13: Planilha com as frequências normalizadas do CoTED.

3.1.3 Análise Multidimensional Funcional Aditiva do Corpus TED Talks (CoTED)

Esta seção começa a tratar diretamente sobre a primeira pergunta da presente pesquisa: Como o corpus das TED Talks (CoTED) se encaixa nas dimensões de variação da língua inglesa encontradas por Biber (1988)? A Análise Multidimensional Funcional Aditiva (AMD Aditiva) possibilita a comparação entre o CoTED, ou seja, a linguagem verbal das TED Talks – definido como o registro considerado nesta pesquisa –, com os registros da língua inglesa. Como

principal requisito nessa comparação, é preciso que os textos do CoTED tenham sido etiquetados com as mesmas variáveis do estudo de Biber (1988). Portanto, considerando os passos anteriores – até a contagem das características linguísticas pelo *Biber Tag Count* – foram seguidos os seguintes passos, por meio do programa *SAS OnDemand for Academics*:

1) Cálculo dos escores de fator da linguagem verbal das TED Talks com base nos desvios médios e padrão do estudo de Biber (1988).
2) Cômputo dos escores mínimos e máximos, média, desvio padrão e abrangência (tabelas 11 e 12).
3) Cômputo dos escores de fator de cada texto em cada dimensão (figura 14). Um *escore* de fator é um valor calculado com base na presença das características linguísticas que carregaram em cada fator. Assim, cada texto recebe um *escore* que determina o nível da presença de cada dimensão.
4) Percentual de variação da linguagem das TED Talks ao longo de cada uma das cinco dimensões da língua inglesa encontradas por Biber (1988), por meio da Análise de Variância (ANOVA).

As contagens das características linguísticas, já normalizadas, foram inseridas no programa *SAS OnDemand for Academics* – obtendo os escores mínimos e máximos, bem como o cálculo do escore médio, do desvio padrão e da abrangência das dimensões de variação do CoTED no geral (tabela 11) e suas três subdivisões (tabela 12) – resultados que serão explorados na seção 4 desta pesquisa:

Escores médios do CoTED						
Variáveis	**N**	**Total**	**Média**	**Mínimo**	**Máximo**	**Desvio-padrão**
dim1	3411	48163.66	14.1200997	-24.1600000	59.4200000	11.5131177
dim2	3411	-6585.88	-1.9307769	-5.3800000	7.2900000	1.5522293
dim3	3411	-1104.52	-0.3238112	-11.9300000	12.1200000	3.0749068
dim4	3411	-2505.58	-0.7345588	-6.6100000	11.4800000	2.2262327
dim5	3411	7164.55	2.1004251	-3.6300000	15.2800000	2.2092495

Tabela 11: Escores médios do CoTED.

Na tabela 11, na primeira coluna, encontramos as 5 dimensões da língua inglesa encontradas por Biber (1988) – variáveis dependentes; na segunda coluna, encontramos o total de textos das TED Talks analisados; e, nas demais colunas temos, na seguinte ordem, os escores

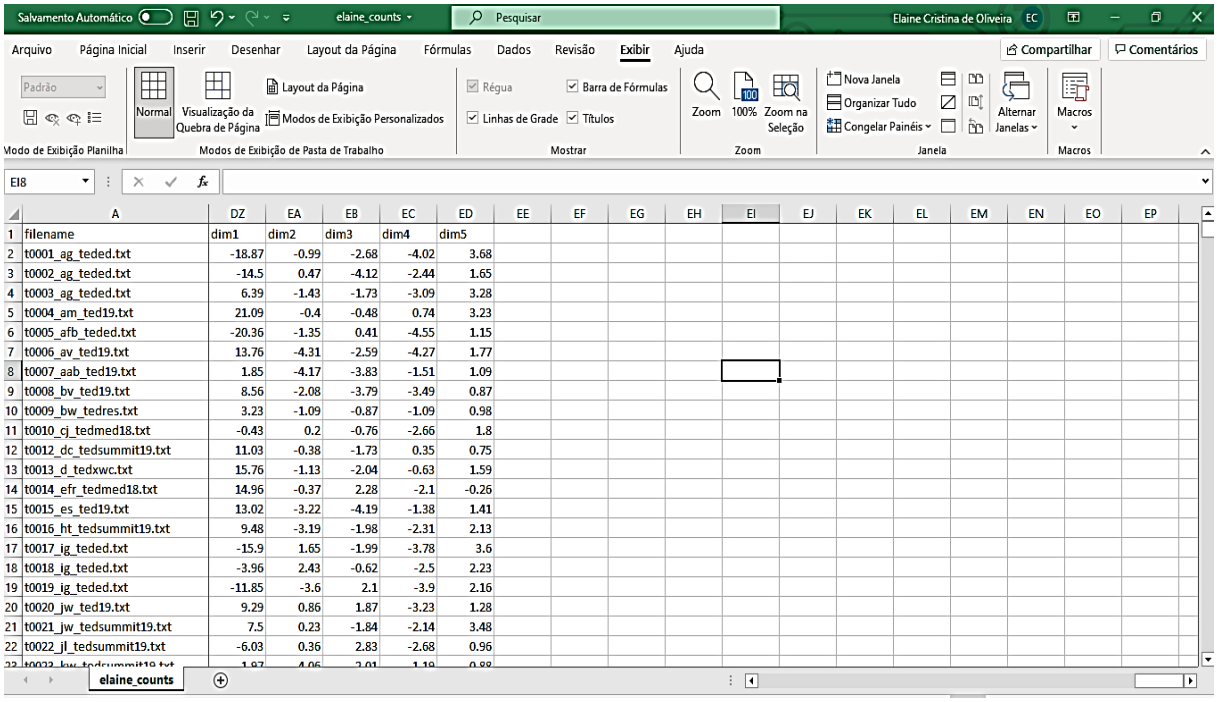
totais, médios, máximos, além do desvio-padrão. Por conta do desvio-padrão alto na dimensão 1, valor de 11.5131177, também foi feita a análise fatorial dos três grupos de TED Talks aqui elencados (TED tradicional, TEDx e TED-Ed), de forma a tentar descobrir se essa subdivisão das TED Talks carregam alguma influência na variação linguística, ou na sua linguagem verbal.

Escores médios - TED trad/TEDx/TED-Ed								
ted_type	**N Obs**	**Variáveis**	**N**	**Total**	**Média**	**Mínimo**	**Máximo**	**Desvio-padrão**
TED trad	2400	dim1	2400	37437.31	15.5988792	-24.1600000	52.3600000	10.1742315
		dim2	2400	-4523.30	-1.8847083	-5.3800000	7.2900000	1.4721854
		dim3	2400	-1178.75	-0.4911458	-11.9300000	11.8800000	2.9945606
		dim4	2400	-1615.97	-0.6733208	-6.6100000	10.5100000	1.9974807
		dim5	2400	4817.66	2.0073583	-3.6300000	12.5000000	2.1543752
TEDx	603	dim1	603	10643.22	17.6504478	-7.3800000	46.2800000	9.2603001
		dim2	603	-1041.16	-1.7266335	-5.1400000	3.6200000	1.5194295
		dim3	603	-229.9100000	-0.3812769	-11.6000000	12.1200000	2.8606174
		dim4	603	-197.2100000	-0.3270481	-6.100000	8.0100000	1.9561183
		dim5	603	1194.32	1.9806302	-3.2900000	11.4800000	1.8186425
TED-Ed	408	dim1	408	83.1300000	0.2037500	-23.5900000	59.4200000	12.1686680
		dim2	408	-1021.42	-2.5034804	-5.3700000	5.9900000	1.8958276
		dim3	408	304.1400000	0.7454412	-11.6300000	10.7700000	3.5973799
		dim4	408	-692.4000000	-1.6970588	-6.6100000	11.4800000	3.3391060
		dim5	408	1152.57	2.8249265	-3.1600000	15.2800000	2.8363310

Tabela 12: Escores médios - TED trad/TEDx/TED-Ed.

Na tabela 12, na primeira coluna, encontramos os 3 grupos de TED Talks aqui definidos – TED tradicional, TEDx e TED-Ed – considerados como possíveis sub-registros nesta pesquisa; na segunda coluna, encontramos o total de textos de cada grupo analisado; e, nas demais colunas temos, na seguinte ordem, os escores totais, médios, máximos, além do desvio-padrão.

Depois, cada texto do CoTED recebeu um escore de fator dentro das 5 dimensões da língua inglesa encontradas por Biber (1988) – figura 14. Esse passo nos ajuda a encontrar os textos mais representativos de cada dimensão, neste caso, por meio do valor da média para os polos positivos e negativos.



	A	DZ	EA	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK	EL	EM	EN	EO	EP
1	filename	dim1	dim2	dim3	dim4	dim5												
2	t0001_ag_teded.txt	-18.87	-0.99	-2.68	-4.02	3.68												
3	t0002_ag_teded.txt	-14.5	0.47	-4.12	-2.44	1.65												
4	t0003_ag_teded.txt	6.39	-1.43	-1.73	-3.09	3.28												
5	t0004_am_ted19.txt	21.09	-0.4	-0.48	0.74	3.23												
6	t0005_afb_teded.txt	-20.36	-1.35	0.41	-4.55	1.15												
7	t0006_av_ted19.txt	13.76	-4.31	-2.59	-4.27	1.77												
8	t0007_aab_ted19.txt	1.85	-4.17	-3.83	-1.51	1.09												
9	t0008_bv_ted19.txt	8.56	-2.08	-3.79	-3.49	0.87												
10	t0009_bw_tedres.txt	3.23	-1.09	-0.87	-1.09	0.98												
11	t0010_cj_tedmed18.txt	-0.43	0.2	-0.76	-2.66	1.8												
12	t0012_dc_tedsummit19.txt	11.03	-0.38	-1.73	0.35	0.75												
13	t0013_d_tedxwc.txt	15.76	-1.13	-2.04	-0.63	1.59												
14	t0014_efr_tedmed18.txt	14.96	-0.37	2.28	-2.1	-0.26												
15	t0015_es_ted19.txt	13.02	-3.22	-4.19	-1.38	1.41												
16	t0016_ht_tedsummit19.txt	9.48	-3.19	-1.98	-2.31	2.13												
17	t0017_ig_teded.txt	-15.9	1.65	-1.99	-3.78	3.6												
18	t0018_ig_teded.txt	-3.96	2.43	-0.62	-2.5	2.23												
19	t0019_ig_teded.txt	-11.85	-3.6	2.1	-3.9	2.16												
20	t0020_jw_ted19.txt	9.29	0.86	1.87	-3.23	1.28												
21	t0021_jw_tedsummit19.txt	7.5	0.23	-1.84	-2.14	3.48												
22	t0022_jl_tedsummit19.txt	-6.03	0.36	2.83	-2.68	0.96												

Figura 14: Escores de fator de cada texto em cada dimensão da língua inglesa (BIBER, 1988).

3.1.4 Análise Multidimensional Funcional Completa do Corpus TED Talks (CoTED)

Esta seção começa a tratar diretamente sobre a segunda pergunta da presente pesquisa: Quais são as dimensões de variação do corpus das TED Talks (CoTED) sob a perspectiva da AMD Funcional Completa? Considerando os passos anteriores – até a contagem das características linguísticas pelo *Biber Tag Count* – foram seguidos os seguintes passos:

1) Análise estatística das planilhas normalizadas referentes aos textos etiquetados – feita por meio do programa *SAS OnDemand for Academics*.
2) Condução da Análise Fatorial inicial das contagens das características linguísticas, para identificar os fatores – feita por meio do programa *SAS OnDemand for Academics*. Por meio da solução não rotacionada do CoTED, podemos encontrar a variância explicada por cada fator. Com o gráfico *scree plot* – figura 15 – é possível ver o ponto de ruptura (comumente chamado de "cotovelo") indicando quais fatores contribuem mais ou menos para a análise geral.
3) Condução da Análise Fatorial rotacionada, a fim de mostrar os fatores consolidados – feita por meio do programa *SAS OnDemand for Academics*.
4) Cômputo dos *escores* de fator de cada texto em cada dimensão – passo realizado também por meio do programa *SAS OnDemand for Academics*. Um *escore* de fator é um valor calculado com base na presença das características linguísticas que carregaram em cada fator. Assim, cada

texto recebe um *escore* que determina cada dimensão. O conjunto dos *escores* representa o perfil multidimensional de cada texto, denotando sua caracterização no espaço multidimensional.

5) Interpretação dos fatores – os fatores foram interpretados em termos de propósitos comunicativos. Como resultado, encontra-se as dimensões funcionais presentes no corpus. Essa etapa da análise é de cunho qualitativo, baseada na leitura e interpretação de textos do corpus que a análise estatística identificou como marcados pelas dimensões, ou seja, que possuem mais carga de cada dimensão.

6) Após a análise qualitativa, as dimensões foram nomeadas, o que eleva os fatores ao estatuto de dimensões de variação.

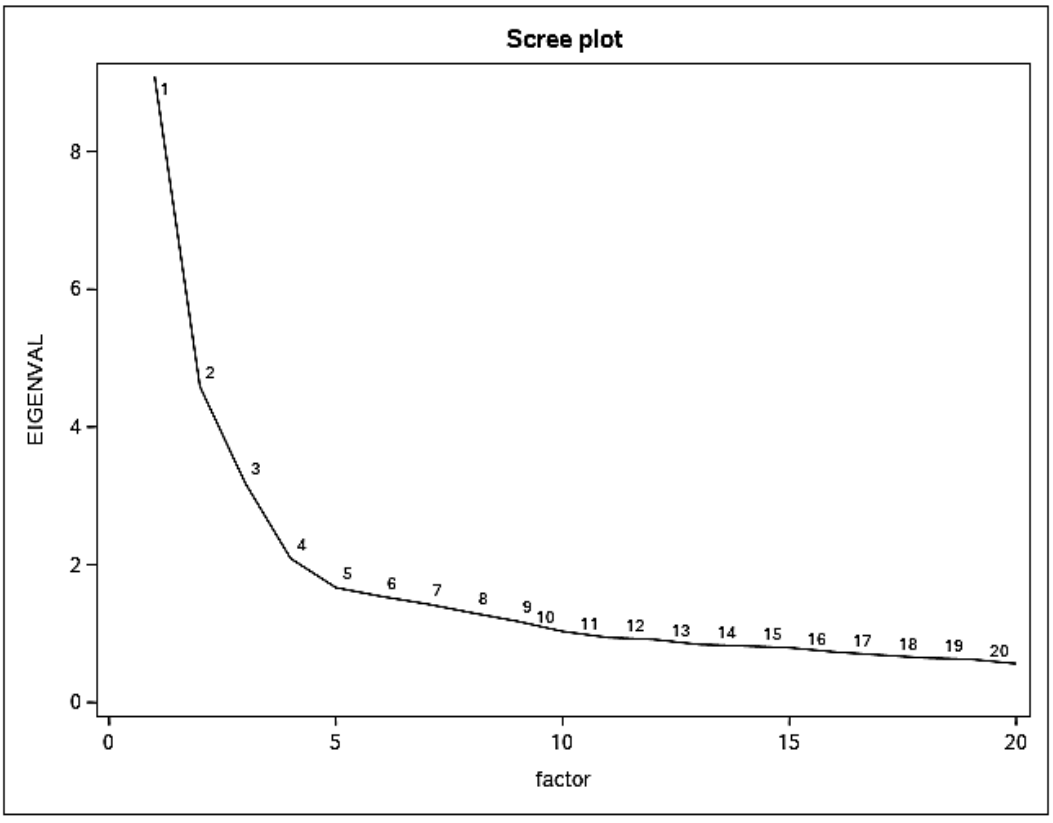


Figura 15: gráfico *scree plot* do CoTED.

Na figura 15, já é possível presumir o número provável de fatores que contribuem mais ou menos para a análise geral das dimensões de variação da linguagem verbal das TED Talks, pois existe uma curva ou um "cotovelo" entre os fatores 3, 4 e 5. Também é possível verificar os valores da solução não rotacionada para decidir quantas dimensões de variação melhor explicam a linguagem verbal das TED Talks – tabela 13:

Variância explicada por cada fator: solução não rotacionada do CoTED							
Fator 1	**Fator 2**	**Fator 3**	**Fator 4**	**Fator 5**	**Fator 6**	**Fator 7**	**etc.**
9.0954709	4.5889020	3.1849807	2.0909929	1.6667570	1.5392419	1.4299470	etc.

Tabela 13: Variância explicada por cada fator (solução não rotacionada do CoTED).

Na tabela 13, os valores de cada fator encontrado tendem a diminuir a cada dimensão. Por isso, é importante escolher as dimensões que melhor explicam o objeto estudado, no caso, a linguagem verbal das TED Talks. Todos esses resultados serão apresentados na seção 4 deste trabalho, que corresponde à sua análise.

3.1.5 Análise de Variância (ANOVA) e Modelo Linear Geral (GLM) do corpus TED Talks (CoTED)

A Análise de Variância (cálculo estatístico univariado – ANOVA) da Análise Multidimensional Funcional Aditiva do CoTED mostra o percentual de variação dos textos das TED Talks explicado pelas variáveis das 5 dimensões da língua inglesa (BIBER, 1988). Os resultados da ANOVA da Análise Multidimensional Funcional Aditiva de CoTED – tabela 14 – serão discutidos na seção 4 deste trabalho.

Cálculo estatístico univariado (ANOVA) da Análise Multidimensional Funcional Aditiva do CoTED					
Fator	**Variável**	**F**	**p**	**R2**	**%**
1	Dimensão 1	434.15	<.0001	0.203050	20,3
2	Dimensão 2	34.72	<.0001	0.019966	2,0
3	Dimensão 3	28.79	<.0001	0.016614	1,7
4	Dimensão 4	50.57	<.0001	0.028823	2,9
5	Dimensão 5	25.31	<.0001	0.014636	1,5

Tabela 14: Cálculo estatístico univariado (ANOVA) da Análise Multidimensional Funcional Aditiva do CoTED.

Os resultados do Modelo Linear Geral (*General Linear Model* – GLM) para as duas variáveis independentes selecionadas, "apresentador" e "evento", são utilizados para responder à terceira pergunta desta pesquisa: como se dá a variação multidimensional funcional em termos das variáveis independentes "apresentador" e "evento" do corpus das TED Talks (CoTED)? Os resultados – tabela 15 – serão discutidos na seção 4 deste trabalho.

Modelo Linear Geral (GLM) das variáveis independentes do CoTED: “apresentador” e “evento”					
Fator	**Variável**	**F**	**p**	**R2**	**%**
1	apresentador	1.49	<.0001	0.342302	34,2
	evento	4.26	<.0001	0.365712	36,6
2	apresentador	1.37	<.0001	0.323247	32,3
	evento	4.60	<.0001	0.383898	38,4
3	apresentador	1.26	<.0001	0.303327	30,3
	evento	1.38	<.0001	0.157239	15,7
4	apresentador	1.33	<.0001	0.315607	31,6
	evento	1.49	<.0001	0.167496	16,7

Tabela 15: Modelo Linear Geral (GLM) das variáveis independentes do CoTED: “apresentador” e “evento”.

É importante relembrar que, os itens analisados nas ANOVAs (incluindo o GLM) são a razão F, o coeficiente de determinação (R2) – o R quadrado – e o valor de p. Segundo Berber Sardinha e Veirano Pinto (2019, p. 6), a razão de F “indica se a variação nos dados é estatisticamente significativa em todos os componentes do corpus”, ou seja, a razão F indica a diferença entre os conjuntos em análise medindo a quantidade de variação existente em cada grupo, isso, por meio do cálculo dos escores médios dos textos e dos escores médios de cada dimensão; e quanto maior o valor de F, mais significantes serão os resultados. Quanto ao valor de p, ele é uma estimativa probabilística para verificar se o valor de um teste estatístico ocorre aleatoriamente. Neste caso, para que os resultados sejam considerados significativos, o valor de p deve ser menor que 0,05 (5%), o que significa que a probabilidade de a variação ocorrer por acaso é de uma a cada 20 vezes, ou seja, ela se torna improvável se ocorrer menos de 5% das vezes – para que o valor de p seja inferior a 0,05, o valor de F deve estar acima de 3,35, pois p é o nível de significância de F. O R2, por sua vez, mede a porcentagem de variação capturada em cada dimensão para cada variável, dependente ou independente, analisada.

4. Resultados – Apresentação e discussão

Nesta seção serão apresentados e discutidos os resultados das análises descritas na Metodologia de Pesquisa deste trabalho. A seção está dividida em três partes, respondendo às três perguntas de pesquisa: 1) Como o corpus das TED Talks (CoTED) se encaixa nas dimensões de variação da língua inglesa encontradas por Biber (1988)? Quais são as dimensões de variação do corpus das TED Talks (CoTED) sob a perspectiva da AMD Funcional

Completa? Como se dá a variação multidimensional funcional em termos das variáveis independentes "apresentador" e "evento" do corpus das TED Talks (CoTED)?

A primeira parte é referente aos resultados da Análise Multidimensional Funcional Aditiva do Corpus TED Talks (CoTED) – além da ANOVA, que mostra o percentual de variação dos textos do CoTED explicado pelas variáveis das 5 dimensões da língua inglesa (BIBER, 1988); a segunda traz os resultados da Análise Multidimensional Funcional Completa do Corpus TED Talks (CoTED); e a terceira traz os resultados da variação multidimensional funcional do Corpus TED Talks (CoTED) em termos das variáveis independentes "apresentador" e "evento", isso por meio do Modelo Linear Geral (*General Linear Model* – GLM).

4.1 Resultados da Análise Multidimensional Funcional Aditiva do Corpus TED Talks (CoTED)

Após o desenho, a compilação, a etiquetagem – por meio do *Biber Tagger* – e a contagem das 128 variáveis linguísticas do CoTED – por meio do *Biber Tag Count* –, foram gerados os escores médios de fator padronizado pela análise fatorial – por meio do *SAS OnDemand for Academics* – dos textos do CoTED, no geral, e de suas três categorias aqui estabelecidas, TED tradicional, TEDx e TED-Ed. Com os resultados dos escores médios, foi possível o mapeamento (alocação e comparação) e a análise do CoTED e de suas categorias aqui definidas ao longo das 5 dimensões de variação dos registros falados e escritos da língua inglesa considerados por Biber (1988):

1) Dimensão 1 – Produção marcada por envolvimento versus informacional;
2) Dimensão 2 – Discurso narrativo versus não narrativo;
3) Dimensão 3 – Referência dependente de situação versus elaborada;
4) Dimensão 4 – Argumentação explícita;
5) Dimensão 5 – Estilo abstrato versus não abstrato.

Desta forma, logo abaixo na seção 4.1.1, serão apresentados os valores de escores médios encontrados, o desvio-padrão e uma breve comparação entre eles. Logo depois, nas seções 4.1.2 até 4.1.6, serão apresentados o mapeamento do CoTED e de suas três categorias – TED tradicional, TEDx e TED-Ed – nas 5 dimensões acima citadas. Em seguida, por meio da

observação do comportamento dos textos alocados ao longo das dimensões e de uma análise qualitativa dos resultados, serão apresentadas as diferenças e as semelhanças do CoTED – com suas três categorias ao fundo – em comparação aos registros da língua inglesa, de forma a identificar a coocorrência dos grupos das características linguísticas mais salientes e o caráter comunicativo funcional que possuem.

4.1.1 Escore médio e desvio-padrão

Na tabela 16, na primeira coluna, encontramos as 5 dimensões da língua inglesa encontradas por Biber (1988) – variáveis dependentes; na segunda coluna, encontramos o total de textos das TED Talks analisados; e, nas demais colunas temos os escores médios e o desvio-padrão em cada dimensão.

Escores médios e desvio-padrão do CoTED			
Variáveis	**N**	**Média**	**Desvio-padrão**
dim1	3411	14.1200997	11.5131177
dim2	3411	-1.9307769	1.5522293
dim3	3411	-0.3238112	3.0749068
dim4	3411	-0.7345588	2.2262327
dim5	3411	2.1004251	2.2092495

Tabela 16: Escores médios e desvio-padrão do CoTED.

Conforme pode ser observado na tabela 16, os valores médios obtidos pela análise fatorial são: para a dimensão 1 = 14.1200997; para a dimensão 2 = -1.9307769; para a dimensão 3 = -0.3238112; para a dimensão 4 = -0.7345588; e, para a dimensão 5 = 2.1004251. Os valores dos escores médios são utilizados para mapear (alocar e comparar) a linguagem verbal das TED Talks nas escalas das dimensões da língua inglesa encontradas por Biber (1988). Quanto ao desvio-padrão: para a dimensão 1 = 11.5131177; para a dimensão 2 = 1.5522293; para a dimensão 3 = 3.0749068; para a dimensão 4 = 2.2262327; e, para a dimensão 5 = 2.2092495. Esses valores mostram o quanto de variação ou dispersão existe em relação à média. Logo abaixo, na figura 16, temos uma representação gráfica da distribuição do CoTED (TED Geral) nas Dimensões de 1 a 5 da língua inglesa, ou seja, temos o gráfico de distribuição, também chamado de “gráfico de caixa e bigode”:

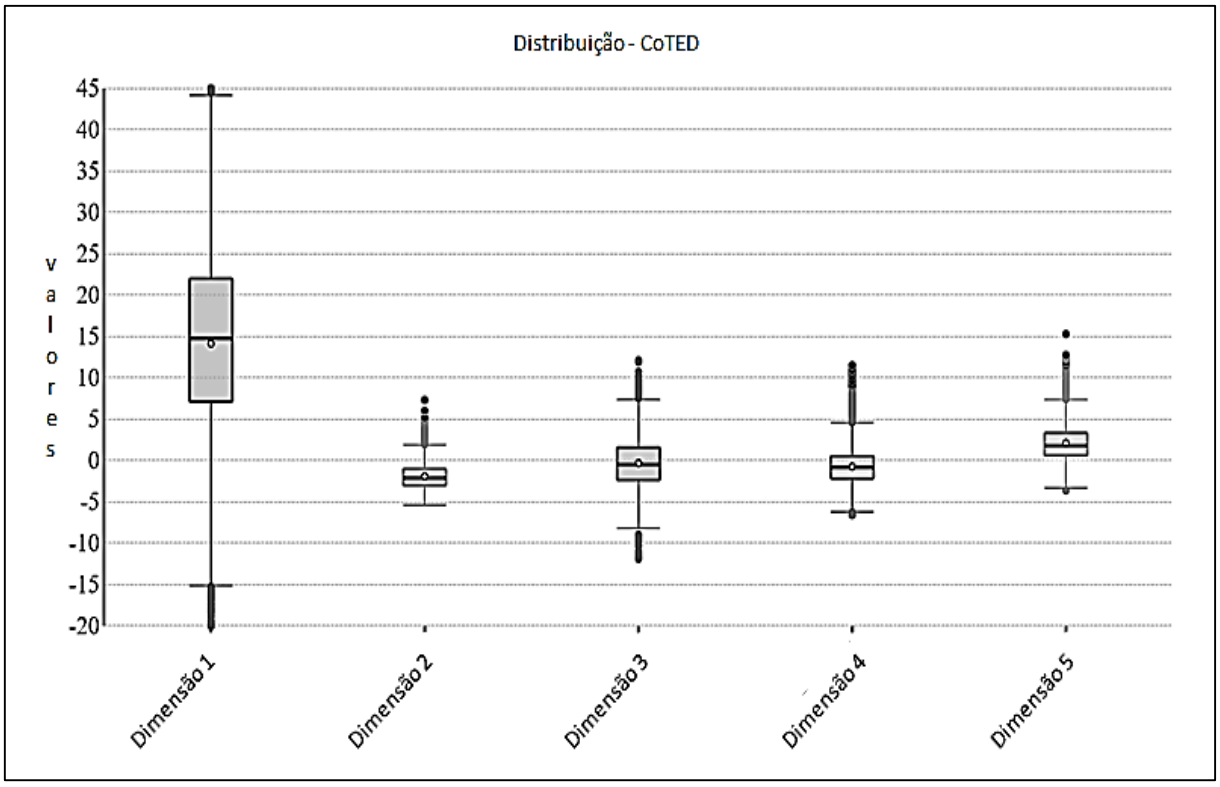


Figura 16: Distribuição do CoTED (TED Geral) nas Dimensões 1-5 (BIBER, 1988) – Ferramenta utilizada: https://goodcalculators.com/box-plot-maker/

Por conta de se ter apresentado um desvio-padrão alto em relação à dimensão 1 da língua inglesa (BIBER, 1988), valor 11.5131177, foi decidido comparar os resultados gerais com os resultados das divisões feitas com relação aos tipos de TED – TED tradicional, TEDx e TED-Ed. Assim, ao considerarmos a dimensão 1, foi possível observar que existe uma considerável variação entre a linguagem da TED-Ed em comparação com a TED tradicional e a TEDx (representada nos mapeamentos na seção 4.1.2). Contudo, isso não ocorre entre as demais dimensões – a tabela 17 mostra os escores médios da TED tradicional, da TEDx e da TED-Ed:

Escores médios e desvio-padrão do TED trad/TEDx/TED-Ed				
ted_type	**N Obs**	**Variáveis**	**Média**	**Desvio-padrão**
TED trad	2400	dim1	15.5988792	10.1742315
		dim2	-1.8847083	1.4721854
		dim3	-0.4911458	2.9945606
		dim4	-0.6733208	1.9974807
		dim5	2.0073583	2.1543752
TEDx	603	dim1	17.6504478	9.2603001
		dim2	-1.7266335	1.5194295
		dim3	-0.3812769	2.8606174

		dim4	-0.3270481	1.9561183
		dim5	1.9806302	1.8186425
TED-Ed	408	dim1	0.2037500	12.1686680
		dim2	-2.5034804	1.8958276
		dim3	0.7454412	3.5973799
		dim4	-1.6970588	3.3391060
		dim5	2.8249265	2.8363310

Tabela 17: Escores médios e desvio-padrão do TED trad/TEDx/TED-Ed.

Contudo, conforme anteriormente mencionado, é possível perceber que as três categorias também apresentam um desvio padrão alto na dimensão 1 – TED tradicional: 10.1742315; TEDx: 9.2603001; e TED-Ed: 12.1686680. Isso significa que existe uma grande variação internamente de cada um deles e que a divisão feita entre os grupos TED Geral, TED tradicional, TEDx e TED-Ed pode não ter tanta relevância na dimensão 1. Certamente devem existir fatores externos à variação da linguagem verbal usada que a influenciam – por isso serão considerados, mais a seguir, os resultados da AMD Funcional Completa para encontrar as dimensões de variação da linguagem verbal das TED Talks e a influência das variáveis independentes "apresentador" e "evento".

Logo abaixo, veremos o mapeamento (alocação e comparação) e a análise do CoTED e de suas categorias aqui definidas – TED Geral, TED tradicional, TEDx e TED-Ed – ao longo das 5 dimensões de variação dos registros falados e escritos da língua inglesa considerados por Biber (1988).

4.1.2 Corpus TED Talks (CoTED) na Dimensão 1 – Produção marcada por envolvimento versus informacional

Conforme previamente explicado neste trabalho, as dimensões da língua inglesa são caracterizadas pela divisão ou oposição entre os polos positivo e negativo. No polo positivo da dimensão 1, temos uma produção textual de caráter de envolvimento ou interacional, como encontrado na linguagem falada – assim como nas conversas (simuladas ou não). Dos 23 registros da língua inglesa considerados por Biber (1988), as conversas telefônicas e as conversas face a face são as que melhor representam o polo positivo dessa dimensão. Quanto ao polo negativo, temos uma produção textual de caráter mais informacional, ou seja, temos mais características da linguagem escrita. Dos 23 registros considerados por Biber (1988), os documentos oficiais e as reportagens jornalísticas são os que melhor representam o polo

negativo da dimensão 1. Logo abaixo temos o mapeamento do corpus TED Talks geral na Dimensão 1 de Biber (1988) – figura 17:

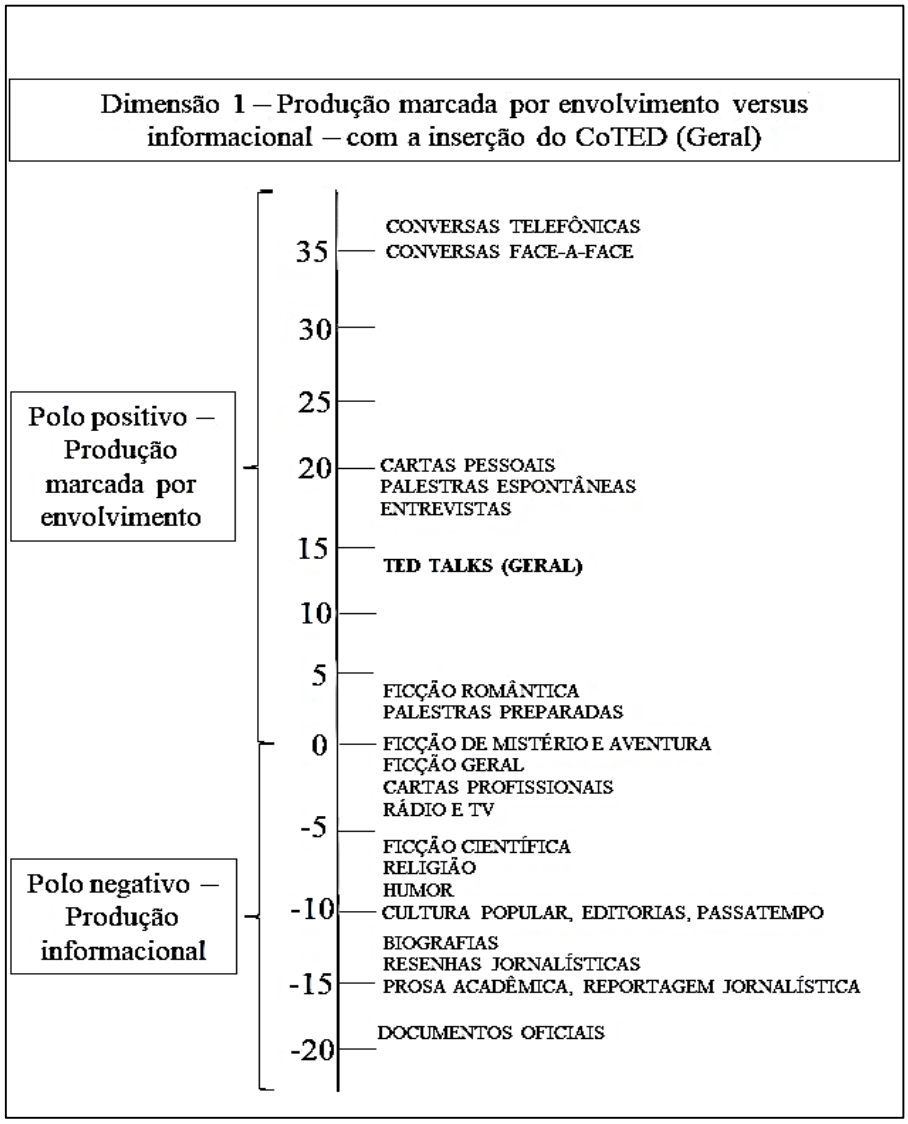


Figura 17: Dimensão 1 – Produção marcada por envolvimento versus informacional – com a inserção do CoTED (TED Geral).

Considerando que o valor de escore médio da linguagem verbal das TED Talks geral é de 14.1200997 (estando no polo positivo) na dimensão 1 da língua inglesa, podemos dizer que, no geral, as TED Talks tem características linguísticas relativamente distintas, se aproximando mais de entrevistas, palestras espontâneas e cartas pessoais. Veremos tais características mais a seguir. Logo abaixo, temos o mapeamento da TED tradicional, da TEDx e da TED-Ed na dimensão 1 – figura 18:

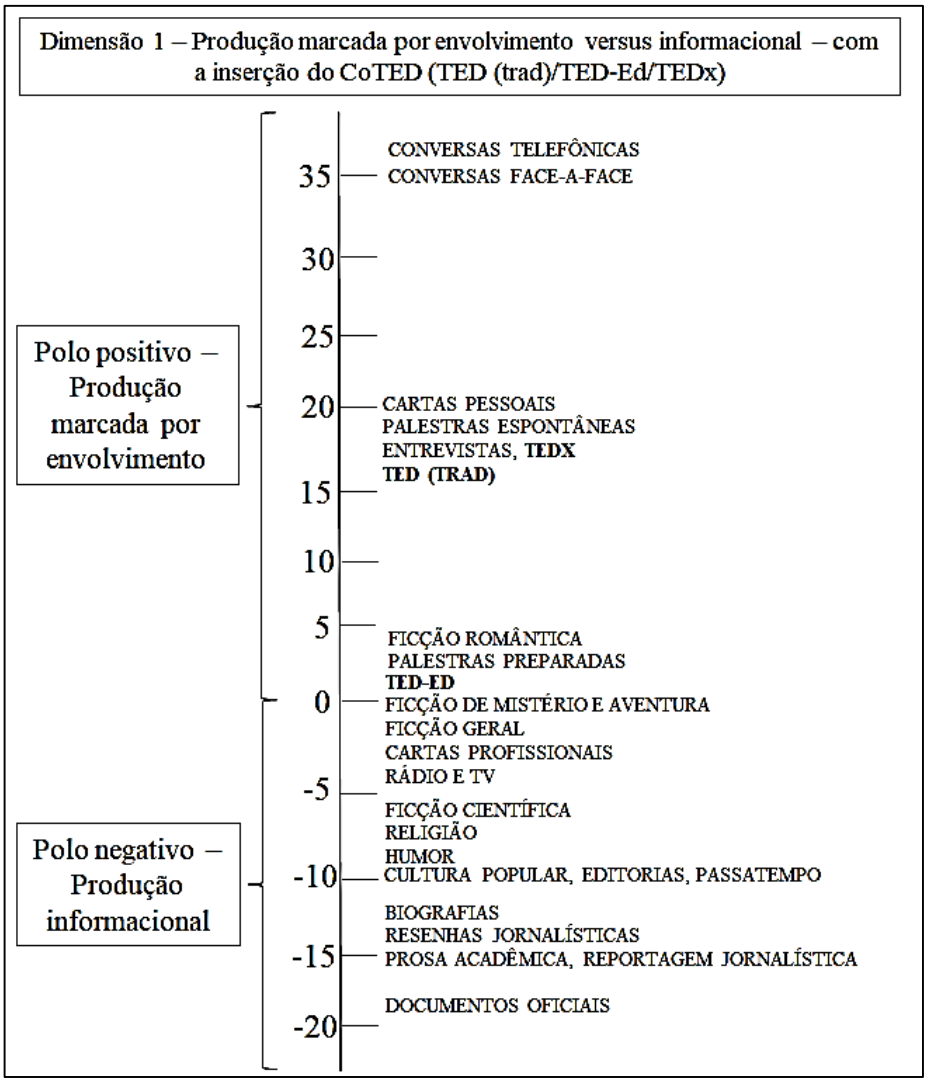


Figura 18: Dimensão 1 – Produção marcada por envolvimento versus informacional – com a inserção do CoTED (TED tradicional/TEDx/TED-Ed).

Considerando o mapeamento das três categorias, TED tradicional (média 15.5988792, no polo positivo), TEDx (média 17.6504478, no polo positivo) e TED-Ed (média 0.2037500, no polo positivo), vê-se uma clara distinção entre TED-Ed com relação ao TED tradicional e ao TEDx na dimensão 1. A princípio, isso apontaria que, por mais que os vídeos TED-Ed sejam indicados como vídeos TED e que sigam o formato ou estilo TED, existe sim uma diferença nas características linguísticas por eles apresentados. Tais características se aproximam das palestras preparadas e das ficções. Isso nos levaria a concluir que, com a média de 0.2037500, existe muito pouco envolvimento ou situação interacional na sua linguagem. Veremos tais características mais a seguir. Porém, conforme anteriormente visto, a variação das TED-Ed na dimensão 1 também é bastante considerável, ou seja, seu desvio-padrão é de 12.1686680, podendo variar mais de 12 pontos acima ou abaixo do seu valor médio de 0.2037500; e com

seu escore máximo podendo chegar à 59.4200000 e o mínimo chegar à -23.5900000 (ver tabela 12). Agora, para poder identificar os grupos das características linguísticas mais salientes e coocorrentes no CoTED, tanto nos polos positivo e negativo da dimensão 1 de Biber (1988), temos que levar em consideração a estrutura do fator 1, conforme representado na tabela 18:

Estrutura do Fator 1 (BIBER, 1988)			
Produção marcada por envolvimento versus informacional			
Polo positivo		**Polo negativo**	
verbo privado	0,96	substantivo	-0,47
apagamento de 'that'	0,91	tamanho de palavra	-0,54
contração	0,90	preposição	-0,54
verbo no tempo presente	0,86	razão forma-ocorrência	-0,58
pronome de segunda pessoa	0,86	adjetivo em posição atributiva	-0,80
verbo 'do'	0,82	(advérbio de lugar	-0,32)
negação analítica	0,78	(voz passiva sem agente	-0,38)
pronome demonstrativo	0,76	(oração adjetiva reduzida de particípio	-0,39)
enfatizador	0,74	(oração adjetiva reduzida de gerúndio	-0,42)
pronome de primeira pessoa	0,74		
pronome 'it'	0,71		
'be' como verbo principal	0,71		
subordinação causativa	0,66		
partícula discursiva	0,66		
pronome indefinido	0,62		
advérbio delimitador/atenuador	0,58		
advérbio / qualificador - amplificador	0,56		
pronome relativo	0,55		
pergunta 'wh'	0,52		
verbo modal de possibilidade	0,50		
coordenação não-frasal	0,48		
oração 'wh'	0,47		
preposição final	0,43		
(advérbio	0,42)		
(subordinação condicional	0,32)		

Tabela 18: Estrutura do Fator 1 da língua inglesa (BIBER, 1988).

Ao considerar o escore médio do CoTED na dimensão 1 **(produção marcada por envolvimento/interativo versus informacional)**, foi separado o exemplo abaixo de um trecho retirado do texto *How AI can save our humanity* (TED tradicional, T0820_KFL_TED18, 2018, média 14.12), considerado como um dos mais representativos no polo positivo – tendo em vista que o escore médio é positivo:

I [pronome de primeira pessoa] ***'m*** [contração] *going to talk about how AI and*

mankind ***can*** [verbo modal de possibilidade] *coexist, but first,* ***we*** [pronome de primeira pessoa] ***have*** [verbo no tempo presente] *to rethink about our human values.* ***So*** [partícula discursiva] *let me first make a confession about my errors in my values.* ***It*** [pronome 'it'] ***was*** ['be' como verbo principal] *11 o'clock, December 16, 1991.* ***I*** [pronome de primeira pessoa] ***was*** ['be' como verbo principal] *about to become a father for the first time. My wife, Shen-Ling, lay in the hospital bed going through a* ***very*** [advérbio / qualificador - amplificador] *difficult 12-hour labor.* ***I*** [pronome de primeira pessoa] *sat by her bedside but looked* ***anxiously*** [advérbio] *at my watch, and* ***I*** [pronome de primeira pessoa] *knew* ***something*** [pronome indefinido] ***that*** [pronome relativo] *she* ***didn't*** [contração]. ***I*** [pronome de primeira pessoa] *knew* ***that*** [pronome relativo] ***if in one hour*** [subordinação condicional], *our child* ***didn't*** [contração] *come,* ***I*** [pronome de primeira pessoa] *was going to leave her there and go back to work and make a presentation about AI to my boss, Apple's CEO.*

[...]

So [partícula discursiva] ***how do we differentiate ourselves as humans in the age of AI?*** [pergunta 'wh'] *We talked about the axis of creativity, and certainly* ***that*** [pronome demonstrativo] *is one possibility, and now we introduce a new axis that we can call compassion, love, or empathy. Those are things that AI cannot* [negação analítica] *do. So as AI takes away the routine jobs, I like to* ***think*** [verbo privado] *we* ***can*** [verbo modal de possibilidade], *we* ***should*** [verbo modal de possibilidade] *and we must create jobs of compassion.*

Podemos ver nesse trecho que, a pessoa que fala usa os pronomes de primeira pessoa "eu" (*I*) e "nós" (*we*), partículas discursivas e contrações, buscando uma aproximação entre o falante e o ouvinte, retratando um estilo informal de conversa. No caso do "nós", por exemplo, os ouvintes são chamados a participar do raciocínio do falante. Outros recursos presentes, como fazer perguntas, usar advérbio/qualificador-amplificador, enfatizar com o uso de advérbios e fazer suposições, também podem ser consideradas como formas de se trazer a atenção do ouvinte. Esse texto está categorizado como palestra, porém, vê-se que não é um estilo tradicional de palestra, pois apresenta um caráter de maior envolvimento e interação.

Conforme já tratado anteriormente, como tivemos um valor alto de desvio-padrão na dimensão 1 para o CoTED (TED Geral), 11.5131177, foram consideradas as três categorias

nesta parte da análise: TED tradicional (média 15.5988792, no polo positivo), TEDx (média 17.6504478, no polo positivo) e TED-Ed (média 0.2037500, no polo positivo). O TED-Ed foi selecionado por estar numa disposição consideravelmente diferente das demais categorias na escala. Assim, o exemplo abaixo vem do texto *The psychology of narcissism* (T1953_WKC_TEDED, 2016, média 0.18), o qual representa o valor de escore médio da TED-Ed:

> *Way before the first selfie, the ancient Greeks and Romans had a myth about someone a little* ***too*** [advérbio / qualificador - amplificador] *obsessed with his own image. In one telling, Narcissus* ***was*** [be como verbo principal] *a handsome guy wandering the world in search of someone to love. After rejecting a nymph named Echo, he caught a glimpse of his own reflection in a river, and fell in love with it. Unable to tear himself away, Narcissus drowned. A flower marked the spot of where he died, and* ***we*** [pronome de primeira pessoa] ***call*** [verbo no tempo presente] *that flower the Narcissus. The myth* ***captures*** [verbo no tempo presente] *the basic idea of narcissism, elevated and sometimes detrimental self-involvement. But* ***it's*** [contração] *not just a personality type that* ***shows*** [verbo no tempo presente] *up in advice columns.* ***It's*** [contração] *actually a set of traits classified and studied by psychologists. The psychological definition of narcissism* ***is*** [‘be’ como verbo principal] *an inflated, grandiose self-image. To varying degrees, narcissists think* ***they're*** [contração] *better looking, smarter, and more important than other people, and that they* ***deserve*** [verbo no tempo presente] *special treatment. Psychologists* ***recognize*** [verbo no tempo presente] *two forms of narcissism as a personality trait: grandiose and vulnerable narcissism.* ***There's*** [contração] *also narcissistic personality disorder, a more extreme form, which* ***we*** [pronome de primeira pessoa] ***'ll*** [contração] *return to shortly. Grandiose narcissism* ***is*** [‘be’ como verbo principal] *the most familiar kind, characterized by extroversion, dominance, and attention seeking. Grandiose narcissists pursue attention and power, sometimes as politicians, celebrities, or cultural leaders.*
>
> *[...]*
>
> ***So*** [partícula discursiva] ***can*** [verbo modal de possibilidade] *narcissists improve on those negative traits? Yes! Anything that promotes honest reflection on their own*

behavior and caring for others, like psychotherapy or practicing compassion towards others, **<u>can</u>** [verbo modal de possibilidade] *be helpful.*

Percebemos nesse trecho que não temos a forte presença do pronome de primeira pessoa "eu" (*I*) como encontrado no exemplo anterior. Outras características presentes no exemplo anterior também aparecem no trecho do texto TED-Ed, mas são utilizadas de formas distintas. Por exemplo, temos o pronome de primeira pessoa "nós" (*we*), que parece trazer uma certa pessoalidade ao texto, contudo, percebe-se que é de fato utilizado de forma geral e indireta ao(s) ouvinte(s). Também temos verbos no tempo presente e o verbo '*be*' como principal, porém, as frases têm um caráter muito mais informativo do que interativo. Temos também a presença de verbos no passado, que traz o foco em se contar algo, como um enredo de uma história. Porém, isso não faz parte das características linguísticas da dimensão 1. Mas, ainda assim, aparecem algumas características que se assemelham, como o uso de contração, partícula discursiva, verbo modal de possibilidade, advérbio/qualificador-amplificador e pergunta.

Com isso, podemos verificar que a TED-Ed pode sim ter traços e características de envolvimento ou interação, mas não tão presentes quanto na TED tradicional e na TEDx, como também pode apresentar traços e características informacionais, características mais esperadas para um vídeo educacional. Isso explicaria o seu valor alto no desvio-padrão na dimensão 1 do CoTED. Contudo, devemos manter em mente que o desvio-padrão da própria TED-Ed na dimensão 1 também é alto, podendo variar de mais informacional/interativo até menos informacional/interativo.

4.1.2.1 ANOVA do CoTED – Dimensão 1 (AMD Aditiva)

Considerando os resultados previamente apresentados na seção 3.1.5 das ANOVAs da Análise Funcional Aditiva do CoTED, referente à dimensão 1, temos as seguintes considerações: os resultados indicam uma variância significativa, com F = 434,15, p = <.0001 e R2 = 0.203050, indicando que 20,3% é o percentual de predição de que a variação da linguagem verbal das TED Talks é explicada pela coocorrência das características linguísticas da dimensão 1 de Biber (1988) – tabela 19:

ANOVA do CoTED – Dimensão 1					
Fator	**Variável**	**F**	**p**	**R2**	**%**
1	Dimensão 1	434.15	<.0001	0.203050	20,3

Tabela 19: ANOVA do CoTED – Dimensão 1.

O valor de R2 pode não ser muito alto, mas é, sem dúvidas, importante para enxergar de modo mais amplo como a linguagem verbal das TED Talks é moldada. Contudo, percebe-se que ainda existe uma segunda parte a ser explicada, que corresponde aos demais 80% da variação. É por isso que a diferença encontrada pela média (ver tabela 20) entre TED-Ed em relação ao TED tradicional e TEDx parece ser óbvia à primeira vista, sendo a TED-Ed menos características na linguagem falada e a TED tradicional e a TEDx mais características na linguagem falada. Mas logo percebemos que essa divisão não é 100% definitiva, pois também encontramos um desvio-padrão alto em todas as três categorias com relação à dimensão 1 (tabela 20). Isso significa que, quando temos um desvio-padrão alto, existem muitas variáveis intervenientes. Como será visto logo mais a seguir, sabe-se que as variáveis situacionais (independentes) – como "apresentador" e "evento", por serem randômicas e não fixas – exercem uma influência importante na linguagem verbal das TED Talks. Deste modo, podemos até considerar a TED-Ed como mais informacional do que a TED tradicional e a TEDx, mas nem sempre.

Escores médios e desvio-padrão – dimensão 1 (TED trad/TEDx/TED-Ed)				
ted_type	**N Obs**	**Variáveis**	**Média**	**Desvio-padrão**
TED trad	2400	dim1	15.5988792	10.1742315
TEDx	603	dim1	17.6504478	9.2603001
TED-Ed	408	dim1	0.2037500	12.1686680

Tabela 20: Escores médios e desvio-padrão – dimensão 1 (TED trad/TEDx/TED-Ed).

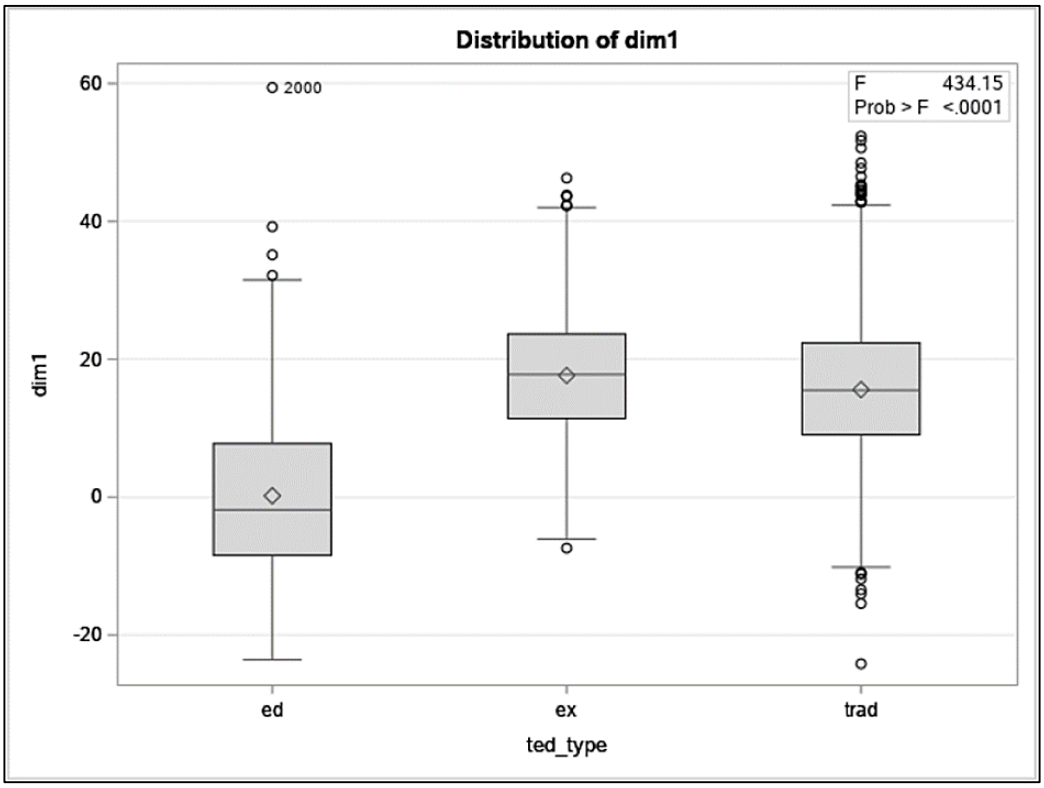


Figura 19: Distribuição – TED tradicional, TEDx e TED-Ed – Dimensão 1 (BIBER, 1988) – Fonte: SAS OnDemand for Academics.

Na figura 19, temos o gráfico de distribuição – chamado de "gráfico de caixa e bigode" – da TED tradicional, TEDx e TED-Ed na dimensão 1, feito pelo *SAS OnDemand for Academics*. Nele, podemos visualmente perceber o distanciamento e a aproximação encontrados entre TED-Ed, Ted tradicional e TEDx na dimensão 1.

4.1.3 Corpus TED Talks (CoTED) na Dimensão 2 – Discurso narrativo versus não narrativo

A dimensão 2 da língua inglesa também é caracterizada pela divisão ou oposição entre os polos positivo e negativo. No polo positivo da dimensão 2, temos uma produção textual com foco narrativo, cujo propósito comunicativo é de relatar um evento. Dos 23 registros da língua inglesa considerados por Biber (1988), os registros de ficção são os que melhor representam o polo positivo dessa dimensão. Quanto ao polo negativo, encontramos uma produção textual de discurso não narrativo. Dos 23 registros considerados por Biber, os registros de rádio e TV, passatempos e documentos oficiais são os que melhor representam o polo negativo da dimensão

2. Logo abaixo temos o mapeamento do corpus TED Talks geral na Dimensão 2 de Biber (1988) – figura 20:

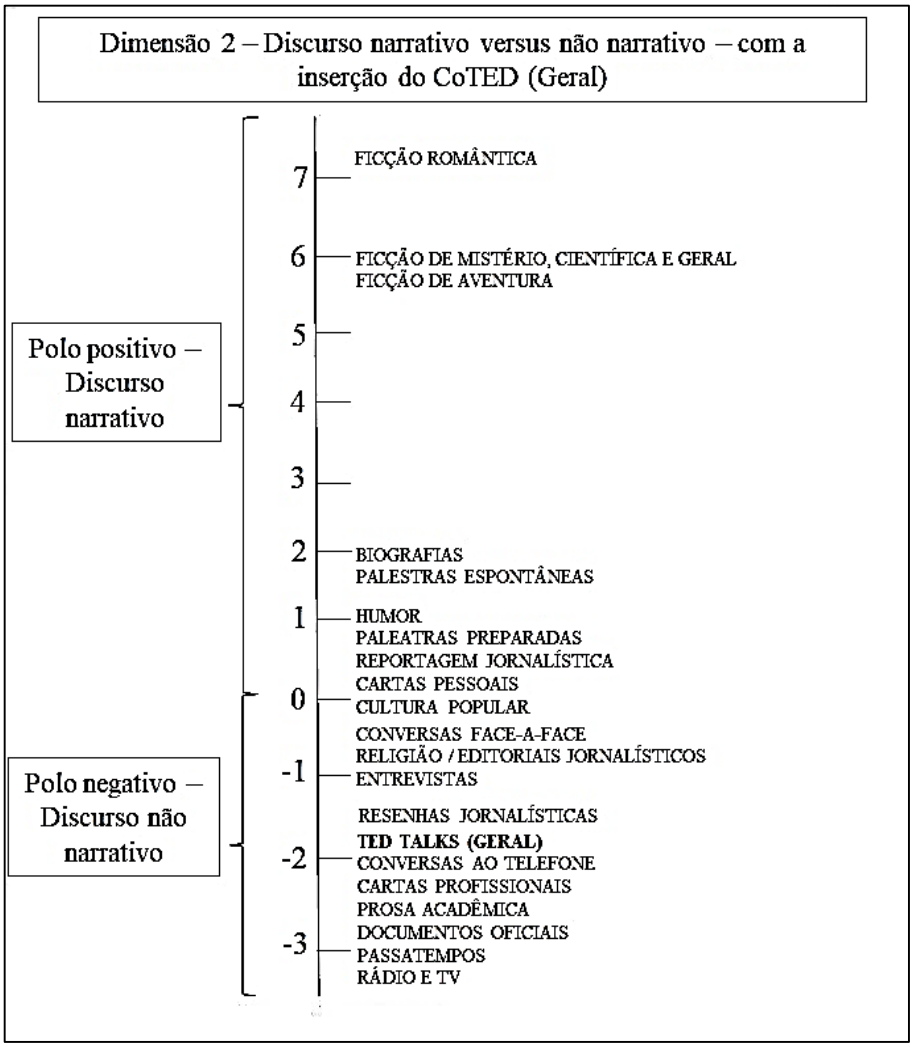


Figura 20: Dimensão 2 – Discurso narrativo versus não narrativo – com a inserção do CoTED (TED Geral).

Segundo o valor de escore médio na dimensão 2, conforme representado no quadro acima, a linguagem verbal das TED Talks geral se localiza na média -1.9307769 da tabela, estando no polo negativo. Essa posição pode nos indicar que as TED Talks têm características semelhantes às resenhas jornalísticas e conversas ao telefone. Veremos tais características mais a seguir. Logo abaixo – apesar de não termos um valor de desvio-padrão alto na dimensão 2 –, temos o mapeamento da TED tradicional, da TEDx e da TED-Ed na Dimensão 2 de Biber (1988), por via de comparação com a TED Geral (ver figuras 20 e 21):

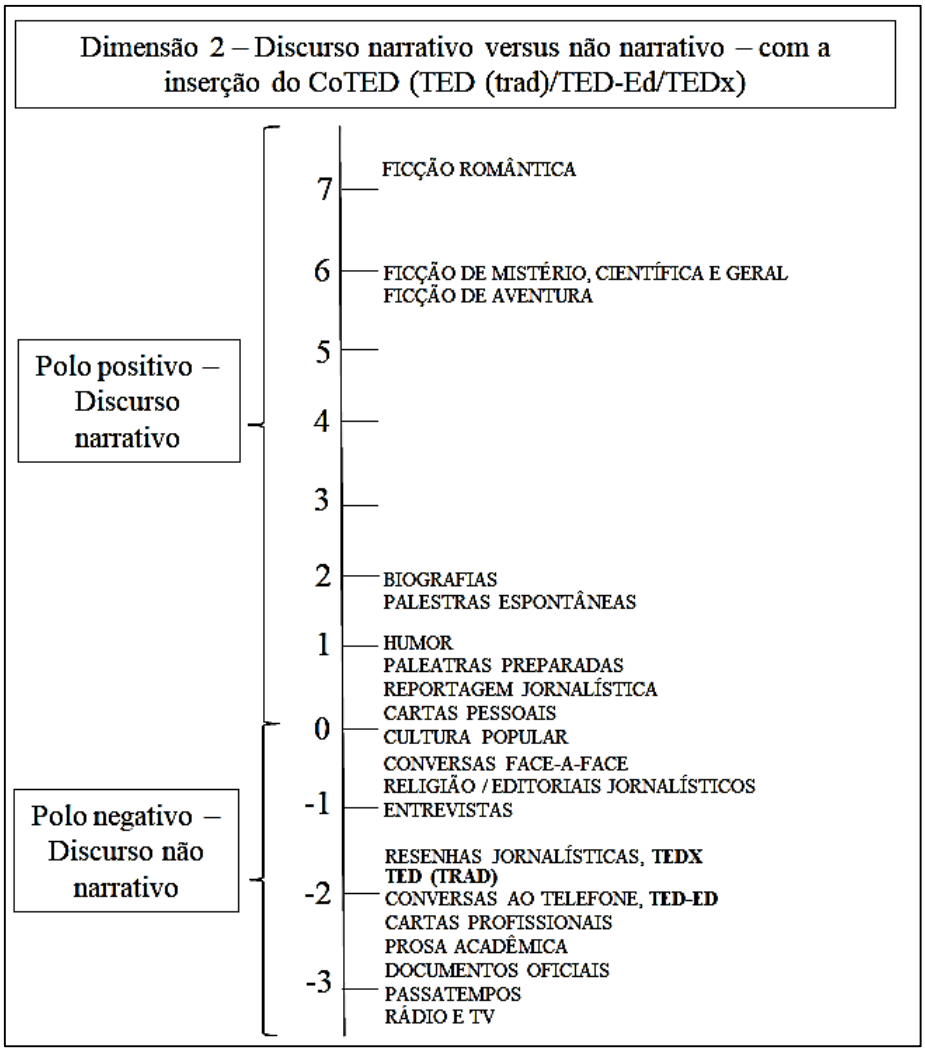


Figura 21: Dimensão 2 – Discurso narrativo versus não narrativo – com a inserção do CoTED (TED (trad)/TED-Ed/TEDx).

Considerando o mapeamento das três categorias TED tradicional (média -1.8847083, no polo negativo), TEDx (média -1.7266335, no polo negativo) e TED-Ed (média -2.5034804, no polo negativo), percebe-se que tanto a TED Talks Geral quanto as suas três categorias se encontram muito próximas umas das outras. Isso indica que eles se assemelham aos registros das resenhas jornalísticas, conversas ao telefone e cartas profissionais, todos os quais, apresentam uma produção textual com foco no discurso não narrativo.

Para poder identificar os grupos das características linguísticas mais salientes e coocorrentes do CoTED tanto nos polos positivo e negativo, desta vez, da dimensão 2 de Biber (1988), temos que levar em consideração a estrutura do fator 2, conforme representado na tabela 21:

Estrutura do Fator 2 (BIBER, 1988)			
Discurso narrativo versus não narrativo			
Polo positivo		**Polo negativo**	
verbo no tempo passado	0,90	(verbo no tempo presente	-0,47
pronome de terceira pessoa	0,73	(adjetivo em posição atributiva	-0,41
verbo no aspecto perfeito	0,48	(oração adjetiva reduzida de particípio	-0,34
verbo público	0,43	(tamanho de palavra	-0,31
negação sintética	0,40		
oração reduzida de gerúndio	0,39		

Tabela 21: Estrutura do Fator 2 da língua inglesa (BIBER, 1988).

Ao considerar o escore médio do CoTED na dimensão 2 **(discurso narrativo versus não narrativo)**, foi separado o exemplo abaixo de um trecho retirado do texto *Are we in control of our own decisions?* (TED tradicional, T0077_DA_EG08, 2008, média -1.9), considerado como um dos mais representativos no polo negativo – tendo em vista que o escore médio é negativo:

> *But there **is*** [verbo no tempo presente] *a **silver*** [adjetivo em posição atributiva] *lining. The **silver*** [adjetivo em posição atributiva] *lining **is*** [verbo no tempo presente], *I **think*** [verbo no tempo presente], *kind of the reason that **behavioral*** [adjetivo em posição atributiva] *economics **is*** [verbo no tempo presente] *interesting and exciting. Are we Superman, or are we Homer Simpson? When it **comes*** [verbo no tempo presente] *to building the **physical*** [adjetivo em posição atributiva] *world, we kind of **understand*** [verbo no tempo presente] *our limitations. We **build*** [verbo no tempo presente] *steps. And we **build*** [verbo no tempo presente] *these things that not everybody can use, obviously. (Laughter) We **understand*** [verbo no tempo presente] *our limitations, and we **build*** [verbo no tempo presente] *around them. But for some reason, when it **comes*** [verbo no tempo presente] *to the **mental*** [adjetivo em posição atributiva] *world, when we **design*** [verbo no tempo presente] *things like healthcare and retirement and stock markets, we somehow **forget*** [verbo no tempo presente] *the idea that we are limited. I **think*** [verbo no tempo presente] *that if we understood our **cognitive*** [adjetivo em posição atributiva] *limitations in the same way we **understand*** [verbo no tempo presente] *our **physical*** [adjetivo em posição atributiva] *limitations, even though they **don't stare*** [verbo no tempo presente] *us in the face the same way, we could design a **better*** [adjetivo em posição atributiva] *world, and that, I **think*** [verbo no tempo presente], ***is*** [verbo no tempo presente] *the*

hope of this thing. Thank you very much. (Applause)

Podemos ver nesse trecho uma significativa frequência dos verbos no tempo presente, com afirmações, suposições e ações no presente, em contraste com o uso do verbo no passado utilizado para se narrar um evento. Outra característica importante é o uso extensivo do adjetivo em posição atributiva. Como o próprio nome já diz, existe uma atribuição dada a algo, no caso aos substantivos. Essa atribuição é dada de forma direta sem a "ajuda" de outra palavra, como um verbo ou uma preposição, podendo ser um recurso utilizado pelo falante para ilustrar sua opinião quanto a algo. Além disso, essa característica não é descrita como pertencente ao discurso narrativo. Uma característica interessante que também podemos considerar é o tamanho das palavras que, geralmente, são maiores em discursos não narrativos do que em discursos narrativos, conforme podemos ver na dimensão 1 da língua inglesa (BIBER, 1988). Deste modo, podemos chegar à conclusão de que o texto acima possui um discurso com foco essencialmente não narrativo, característica que pode ser atribuída às TED Talks no geral. Como não tivemos um valor muito alto de desvio-padrão, não foi separado um exemplo exclusivo do TED-Ed (ou do TEDx), conforme foi feito para a dimensão 1.

4.1.3.1 ANOVA do CoTED – Dimensão 2 (AMD Aditiva)

Considerando os resultados previamente apresentados na seção 3.1.5 das ANOVAs da Análise Funcional Aditiva do CoTED (ver tabela 14), referente à dimensão 2, temos as seguintes considerações: os resultados da ANOVA indicam uma variância não tão significativa em comparação com a dimensão 1, com F = 34.72, p = <.0001 e R2 = 0.019966, indicando que menos de 2,0% é o percentual de predição de que a variação da linguagem verbal das TED Talks é explicada pela coocorrência das características linguísticas da dimensão 2 de Biber (1988). Ainda assim, podemos classificar a linguagem verbal das TED Talks no geral como menos narrativa.

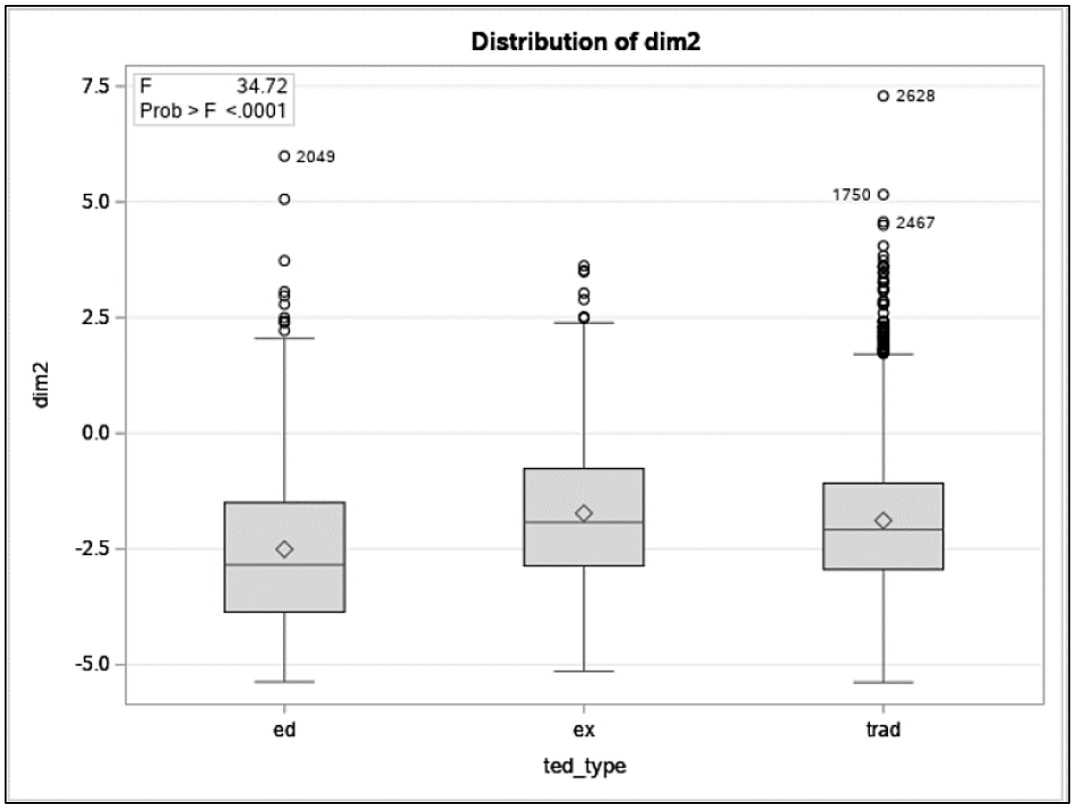


Figura 22: Distribuição – TED tradicional, TEDx e TED-Ed – Dimensão 2 (BIBER, 1988) – Fonte: SAS OnDemand for Academics.

Na figura 22, temos o gráfico de distribuição – chamado de "gráfico de caixa e bigode" – da TED tradicional, TEDx e TED-Ed na dimensão 2, feito pelo *SAS OnDemand for Academics*. Nele, podemos visualmente perceber o quão próximas são a Ted tradicional, a TEDx e a TED-Ed na dimensão 2.

4.1.4 Corpus TED Talks (CoTED) na Dimensão 3 – Referência dependente de situação versus elaborada

A dimensão 3 da língua inglesa também é caracterizada pela divisão ou oposição entre os polos positivo e negativo. No polo positivo da dimensão 3, temos uma produção textual com foco na referência dependente de situação, ou seja, apresenta informações e referências altamente explícitas dentro do próprio texto. Dos 23 registros da língua inglesa considerados por Biber (1988), os registros de documentos oficiais, cartas profissionais, resenhas jornalísticas e prosa acadêmica são os que melhor representam o polo positivo dessa dimensão. Quanto ao polo negativo, encontramos uma produção textual de situação elaborada, ou seja, as referências são dependente de contexto, sendo elas externas ao próprio texto. Dos 23 registros considerados

por Biber, os registros de rádio e TV, conversas telefônicas, conversas face a face e ficção romântica são os que melhor representam o polo negativo da dimensão 3. Logo abaixo temos o mapeamento do corpus TED Talks Geral na Dimensão 3 de Biber (1988) – figura 23:

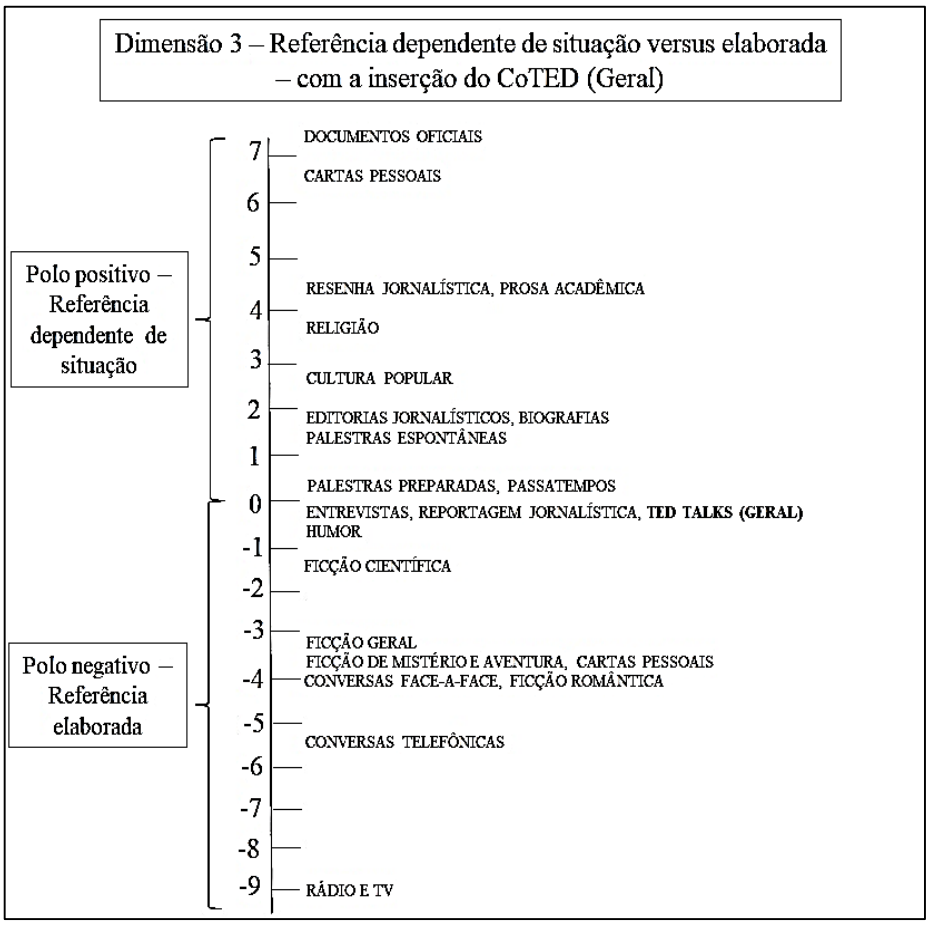


Figura 23: Dimensão 3 – Referência dependente de situação versus elaborada – com a inserção do CoTED (Geral).

Segundo o valor de escore médio na dimensão 3, conforme representado na figura acima, a linguagem verbal das TED Talks geral se localiza na média -0.3238112 da tabela, estando no polo negativo. Essa posição pode nos indicar que a TED Talks em geral tem características semelhantes às entrevistas e reportagens jornalísticas. Veremos tais características mais a seguir. Logo abaixo, temos o mapeamento da TED tradicional, da TEDx e da TED-Ed na Dimensão 3 de Biber (1988) – figura 24:

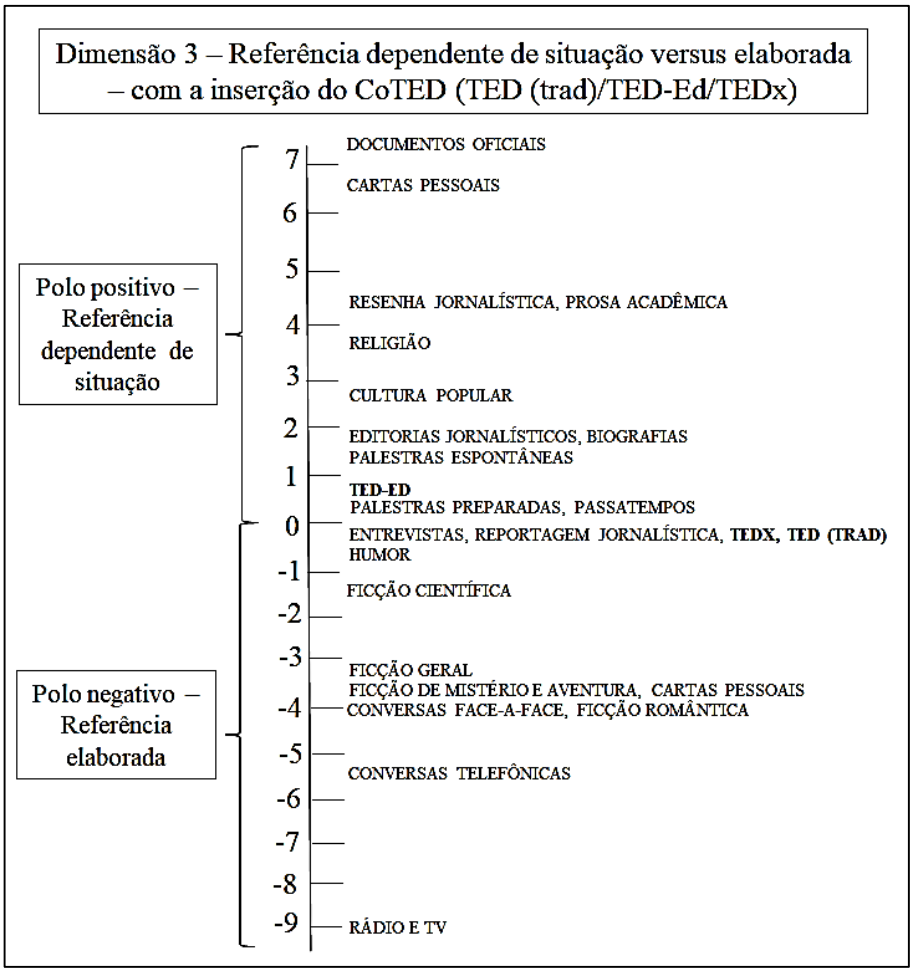


Figura 24: Dimensão 3 – Referência dependente de situação versus elaborada – com a inserção do CoTED (TED (trad)/TED-Ed/TEDx).

Considerando o mapeamento das três categorias TED tradicional (média -0.4911458, no polo negativo), TEDx (média -0.3812769, no polo negativo) e TED-Ed (média 0.7454412, no polo positivo), percebe-se que tanto a TED tradicional quanto a TEDx estão praticamente no mesmo mapeamento da TED geral, ou seja, elas possuem características semelhantes às entrevistas e reportagens jornalísticas. Somente a TED-Ed sofreu uma pequena alteração no seu mapeamento, assemelhando-se mais às palestras preparadas e passatempos.

Para poder identificar os grupos das características linguísticas mais salientes e coocorrentes do CoTED tanto nos polos positivo e negativo da dimensão 3 de Biber (1988), temos que levar em consideração a estrutura do fator 3, conforme representado na tabela 22:

Estrutura do Fator 3 (BIBER, 1988)			
Referência dependente de situação versus elaborada			
Polo positivo		**Polo negativo**	
oração wh em posição de objeto	0,63	advérbio de tempo	-0,60

oração wh com preposição inicial	0,61	advérbio de lugar	-0,49
oração wh em posição de sujeito	0,45	advérbios	-0,46
coordenação frasal	0,36		
nominalização	0,36		

Tabela 22: Estrutura do Fator 3 da língua inglesa (BIBER, 1988).

Ao considerar o escore médio do CoTED na dimensão 3 **(referência dependente de situação versus elaborada)**, foi separado o exemplo abaixo de um trecho retirado do texto How low-cost *eye care can be world-class* (TED tradicional, T3907_TR_TEDIND09, 2009, média -0.32), considerado como um dos mais representativos no polo negativo – tendo em vista que o escore médio é negativo:

> *And **over time*** [locução adverbial de tempo], *we have grown **into*** [advérbio de lugar] *a network of five hospitals, **predominately*** [advérbio] ***in*** [advérbio de lugar] *the state of Tamil Nadu and Puducherry, and **then*** [advérbio de tempo] *we added several, what we call Vision Centers as a hub-and-spoke model. And **then*** [advérbio de tempo] *more **recently*** [advérbio de tempo] *we started managing hospitals **in*** [advérbio de lugar] *other parts of the country and also setting up hospitals **in*** [advérbio de lugar] *other parts of the world as well. **The last three decades*** [locução adverbial de tempo], *we have done about three-and-a-half million surgeries, a vast majority of them for the poor people. **Now*** [advérbio de tempo], ***each year*** [locução adverbial de tempo] *we perform about 300,000 surgeries. A typical day at Aravind, we would do about a thousand surgeries, maybe see about 6,000 patients, send out teams **into*** [advérbio de lugar] *the villages to examine, bring back patients, lots of telemedicine consultations, and, **on*** [advérbio de lugar] *top of that, do a lot of training, both for doctors and technicians who will become the future staff of Aravind. And **then*** [advérbio de tempo] *doing this day-in and day-out, and doing it well, requires a lot of inspiration and a lot of hard work. And I think this was possible thanks to the building blocks put in place by Dr. V., a value system, an efficient delivery process, and fostering the culture of innovation. (Music) Dr. V: I used to sit with the ordinary village man because I am from a village, and **suddenly*** [advérbio] *you turn around and seem to be in contact with his inner being, you seem to be one with him.*

Podemos ver nesse trecho uma presença significativa de advérbios, principalmente os de lugar e tempo, trazendo referências externas ao próprio texto e dependentes de contexto. Isso significa que existe a necessidade de se explicar a situação ou o contexto para que se entenda o texto em si.

Apesar de pouca distância no mapeamento e pelo baixo valor do desvio-padrão, foi separado um exemplo exclusivo do TED-Ed (média 0.7454412), já que se encontra no polo positivo. Assim, o exemplo abaixo vem do texto *How can you change someone's mind?* (T0817_HM_TEDED, 2018, média 0.75), o qual representa o valor de escore médio da TED-Ed:

> *Three people are at a dinner party. Paul,* ***who*** [oração wh em posição de sujeito] *'s married, is looking at Linda. Meanwhile, Linda is looking at John,* ***who*** [oração wh em posição de sujeito]*'s not married. Is someone* ***who*** [oração wh em posição de sujeito]*'s married looking at someone* ***who*** [oração wh em posição de sujeito]*'s not married? Take a moment to think about it. Most people answer that there's not enough information to tell. And most people are wrong. Linda must be* ***either*** [coordenação frasal] *married* ***or*** [coordenação frasal] *not married—there are no other options. So in either scenario, someone married is looking at someone* ***who*** [oração wh em posição de sujeito]*'s not married. When presented with the explanation, most people change their minds and accept the correct answer, despite being very confident in their first responses. Now let's look at another case. A 2005 study by Brendan Nyhan* ***and*** [coordenação frasal] *Jason Reifler examined American attitudes regarding the justifications for the Iraq War. Researchers presented participants with a news article that showed no weapons of mass destruction had been found. Yet many participants not only continued to believe that WMDs had been found, but they even became more convinced of their original views. So why do arguments change people's minds in some cases and backfire in others? Arguments are more convincing when they rest on a good knowledge of the audience, taking into account* ***what*** [oração wh em posição de objeto] *the audience believes,* ***who*** [oração wh em posição de objeto] *they trust, and* ***what*** [oração wh em posição de objeto] *they value.*

Encontramos nesse trecho características atribuídas ao polo positivo da dimensão 3,

como oração ‘wh’ em posição de sujeito, oração ‘wh’ em posição de objeto e coordenação frasal. Tais traços indicam que existe no texto referências dependente de situação, ou seja, informações e referências que estão explícitas dentro do próprio texto. Esse é um indício de que a TED-Ed tem características um pouco mais distintas em comparação com o TED tradicional e o TEDx na dimensão 2.

4.1.4.1 ANOVA do CoTED – Dimensão 3 (AMD Aditiva)

Considerando os resultados previamente apresentados na seção 3.1.5 das ANOVAs da Análise Funcional Aditiva do CoTED (ver tabela 14), referente à dimensão 3, temos as seguintes considerações: os resultados da ANOVA indicam uma variância não tão significativa em comparação com a dimensão 1, com F = 28.79, p = <.0001 e R2 = 0.016614, indicando que por volta de 1,7% é o percentual de predição de que a variação da linguagem verbal das TED Talks é explicada pela coocorrência das características linguísticas da dimensão 3 de Biber (1988). Apesar de baixa percentagem, ainda assim, podemos classificar a linguagem verbal das TED Talks no geral como mais dependente de referências externas, ou seja, dependente de contexto.

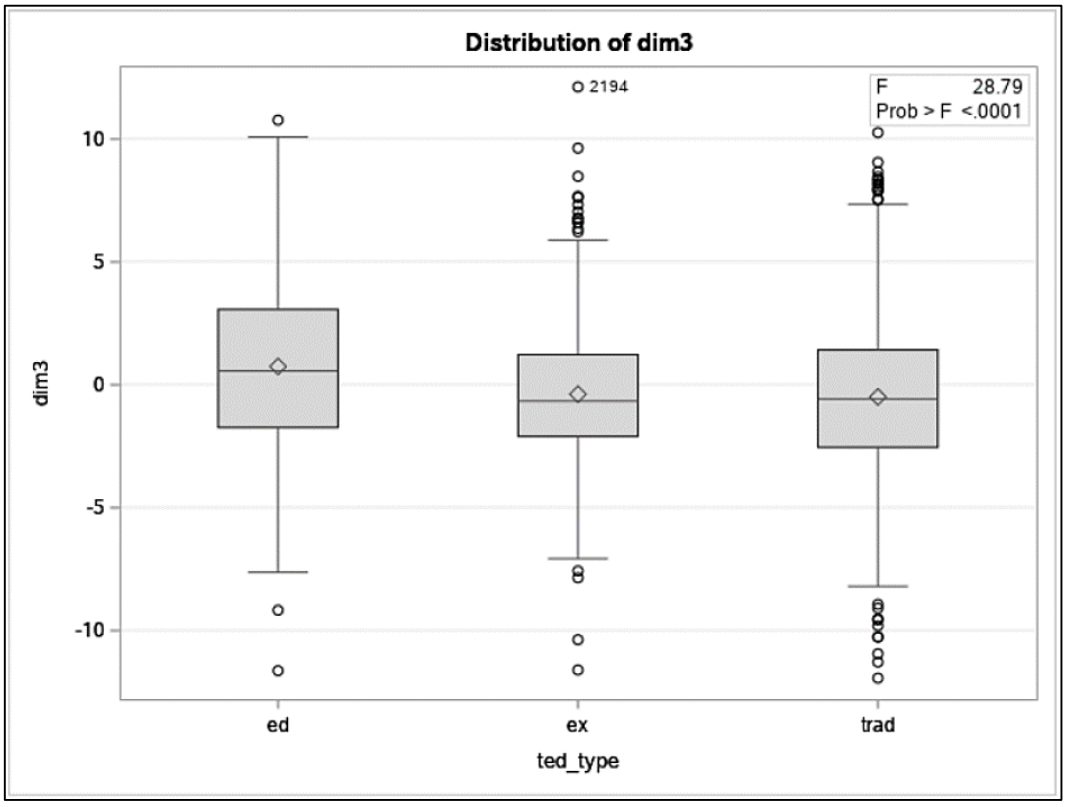


Figura 25: Distribuição – TED tradicional, TEDx e TED-Ed – Dimensão 3 (BIBER, 1988) – Fonte: SAS

OnDemand for Academics.

Na figura 25, temos o gráfico de distribuição – chamado de "gráfico de caixa e bigode" – da TED tradicional, TEDx e TED-Ed na dimensão 3, feito pelo *SAS OnDemand for Academics*. Nele, podemos visualmente perceber uma aproximação entre a Ted tradicional, a TEDx, e a TED-Ed na dimensão 3.

4.1.5 Corpus TED Talks (CoTED) na Dimensão 4 – Argumentação explícita

Ao observarmos a distribuição dos 23 registros considerados por Biber (1988) no mapeamento da dimensão 4 da língua inglesa, alguns deles estão alocados no polo positivo ou no polo negativo, como nas dimensões 1, 2 e 3. Contudo, quanto às características linguísticas, somente o polo positivo da dimensão 4 da língua inglesa é que foi considerado por Biber (1988), e as características carregadas no fator 4 (positivo) estão relacionadas à argumentação explícita. Desta forma, nesse caso, temos uma produção textual com foco persuasivo. Dos 23 registros da língua inglesa considerados por Biber (1988), os registros de cartas profissionais, os editoriais jornalísticos e a ficção romântica são os que melhor representam o polo positivo dessa dimensão. Assim, dos 23 registros da língua inglesa considerados por Biber (1988), os registros de rádio e TV, resenhas jornalísticas e ficção de aventura são os que melhor representam o polo negativo da dimensão 4. Isso significa que, esses registros não possuem em mesma quantidade as características linguísticas presentes nos registros do polo positivo. Logo abaixo, temos o mapeamento do corpus TED Talks geral na Dimensão 4 de Biber (1988) – figura 26:

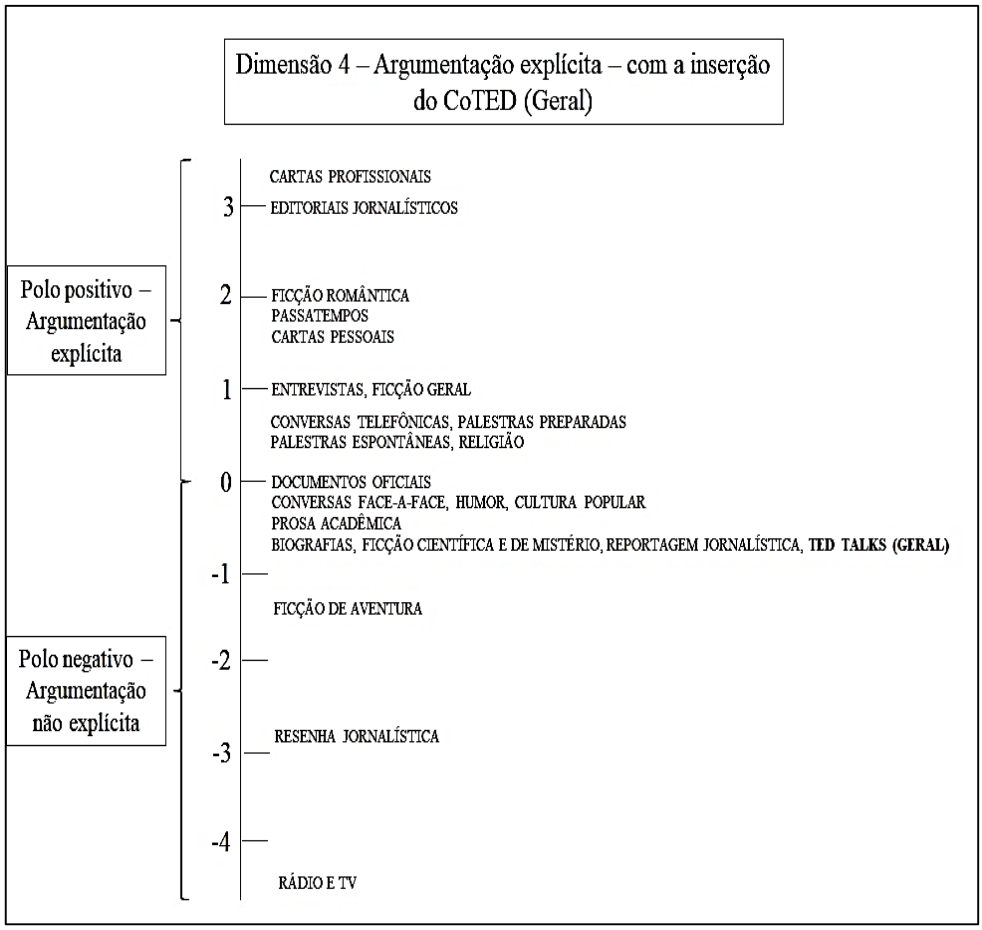


Figura 26: Dimensão 4 – Argumentação explícita – com a inserção do CoTED (Geral).

Segundo o valor de escore médio da dimensão 4, conforme representado no quadro acima, a linguagem verbal das TED Talks geral se localiza na média -0.7345588 da tabela, estando no polo negativo. Essa posição pode nos indicar que as TED Talks têm características semelhantes às biografias, ficções e reportagens jornalísticas. Veremos tais características mais a seguir. Logo abaixo, temos o mapeamento da TED tradicional, da TEDx e da TED-Ed na Dimensão 4 de Biber (1988) – figura 27:

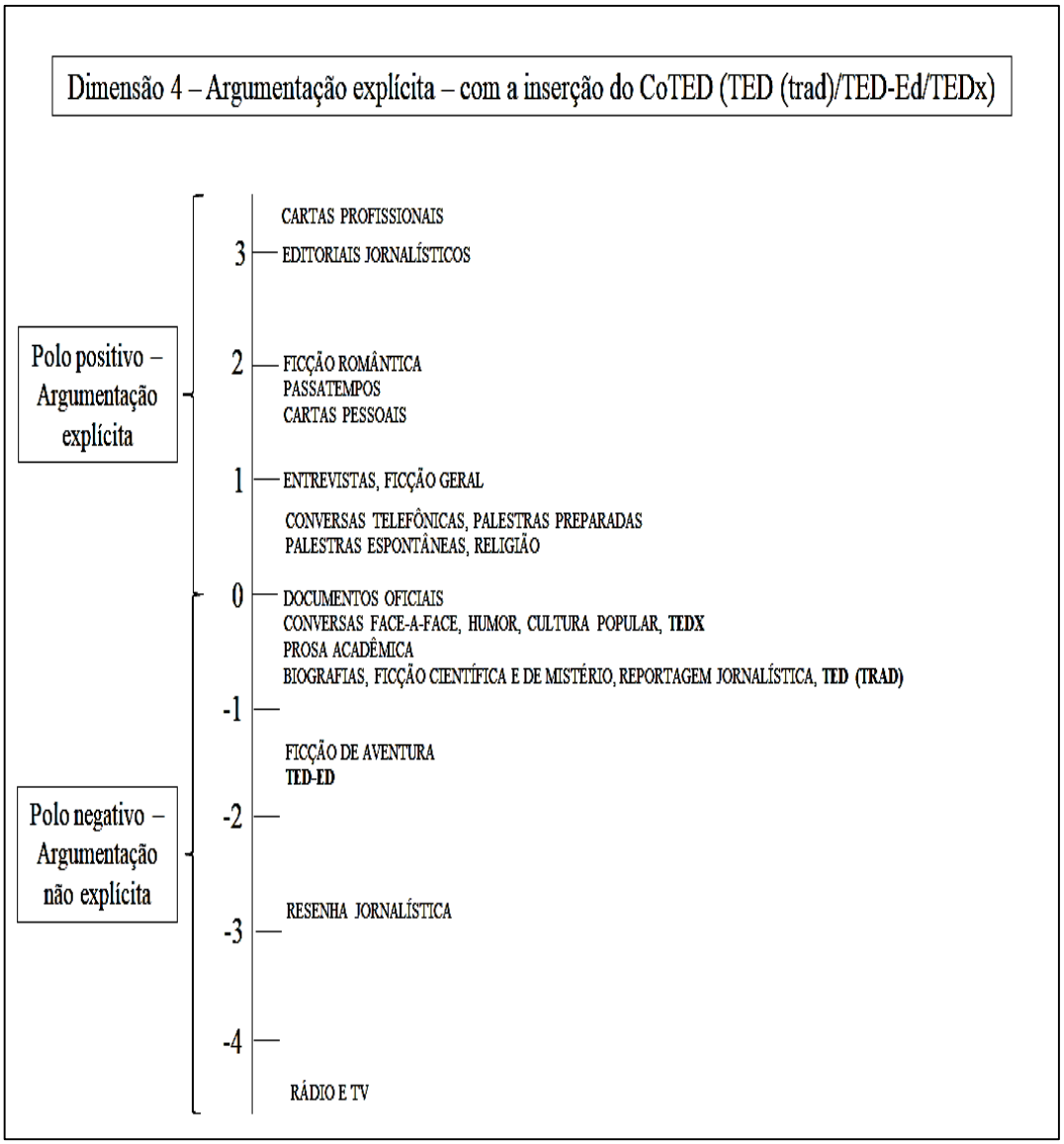


Figura 27: Dimensão 4 – Argumentação explícita – com a inserção do CoTED (TED (trad)/TED-Ed/TEDx).

Considerando o mapeamento das três categorias TED tradicional (média -0.6733208, no polo negativo), TEDx (média -0.3270481, no polo negativo) e TED-Ed (média -1.6970588, no polo negativo), percebe-se que todos eles estão bastante próximos quanto às suas características, apesar de o TED-Ed ter uma maior aproximação à ficção de aventura.

Para poder identificar os grupos das características linguísticas mais salientes e coocorrentes do CoTED tanto nos polos positivo e negativo da dimensão 4 de Biber (1988), temos que levar em consideração a estrutura do fator 4, conforme representado na tabela 23. Contudo, vale relembrar que contamos somente com o polo positivo aqui:

Estrutura do Fator 4 (BIBER 1988)

Argumentação explícita	
Polo positivo	
verbo no infinitivo	0,76
verbo modal de antecipação	0,54
verbo de persuasão	0,49
subordinação condicional	0,47
verbo modal de necessidade	0,46
advérbio encaixado no auxiliar	0,44
(verbo modal de possibilidade	0,37)

Tabela 23: Estrutura do Fator 4 da língua inglesa (BIBER 1988).

Ao considerar o escore médio do CoTED na dimensão 4 **(argumentação explícita)**, foi separado o exemplo abaixo de um trecho retirado do texto *The mathematician who cracked Wall Street* (TED tradicional, T2112_JS_TED15, 2015, média -0.73), considerado como um dos mais representativos no polo negativo – tendo em vista que o escore médio é negativo:

> *Chris Anderson: You were something of a mathematical phenom. You had already taught at Harvard and MIT at a young age. And then the NSA came calling. What was that about? Jim Simons: Well the NSA — that's the National Security Agency — they didn't exactly come calling. They had an operation at Princeton, where they hired mathematicians to attack secret codes and stuff like that. And I knew that existed. And they had a very good policy, because you could do half your time at your own mathematics, and at least half your time working on their stuff. And they paid a lot. So that was an irresistible pull. So, I went there. CA: You were a code-cracker. JS: I was. CA: Until you got fired. JS: Well, I did get fired. Yes. CA: How come? JS: Well, how come? I got fired because, well, the Vietnam War was on, and the boss of bosses in my organization was a big fan of the war and wrote a New York Times article, a magazine section cover story, about how we would win in Vietnam. And I didn't like that war, I thought it was stupid. And I wrote a letter to the Times, which they published, saying not everyone who works for Maxwell Taylor, if anyone remembers that name, agrees with his views. And I gave my own views.*

Como podemos ver, esse texto está categorizado como pertencente ao polo oposto, no caso negativo, da dimensão 4 da língua inglesa. Suas características linguísticas não estão relacionadas à argumentação explícita ou à persuasão. No trecho acima, temos uma forma de

entrevista com relatos sobre eventos marcantes no passado. Isso indica que o texto se comporta como falas de uma história sendo narrada. Como não tivemos um valor muito alto de desvio-padrão, não foi separado um exemplo exclusivo do TED-Ed (ou do TEDx), conforme foi feito para a dimensão 1.

4.1.5.1 ANOVA do CoTED – Dimensão 4 (AMD Aditiva)

Considerando os resultados previamente apresentados na seção 3.1.5 das ANOVAs da Análise Funcional Aditiva do CoTED (tabela 14), referente à dimensão 4, temos as seguintes considerações: os resultados da ANOVA indicam uma variância não tão significativa em comparação com a dimensão 1, com F = 50.57, p = <.0001 e R2 = 0.028823, indicando que 2,9% é o percentual de predição de que a variação da linguagem verbal das TED Talks é explicada pela coocorrência das características linguísticas da dimensão 4 de Biber (1988). Apesar de baixa percentagem, ainda assim, podemos classificar a linguagem verbal das TED Talks no geral como menos relacionada à argumentação explícita ou à persuasão.

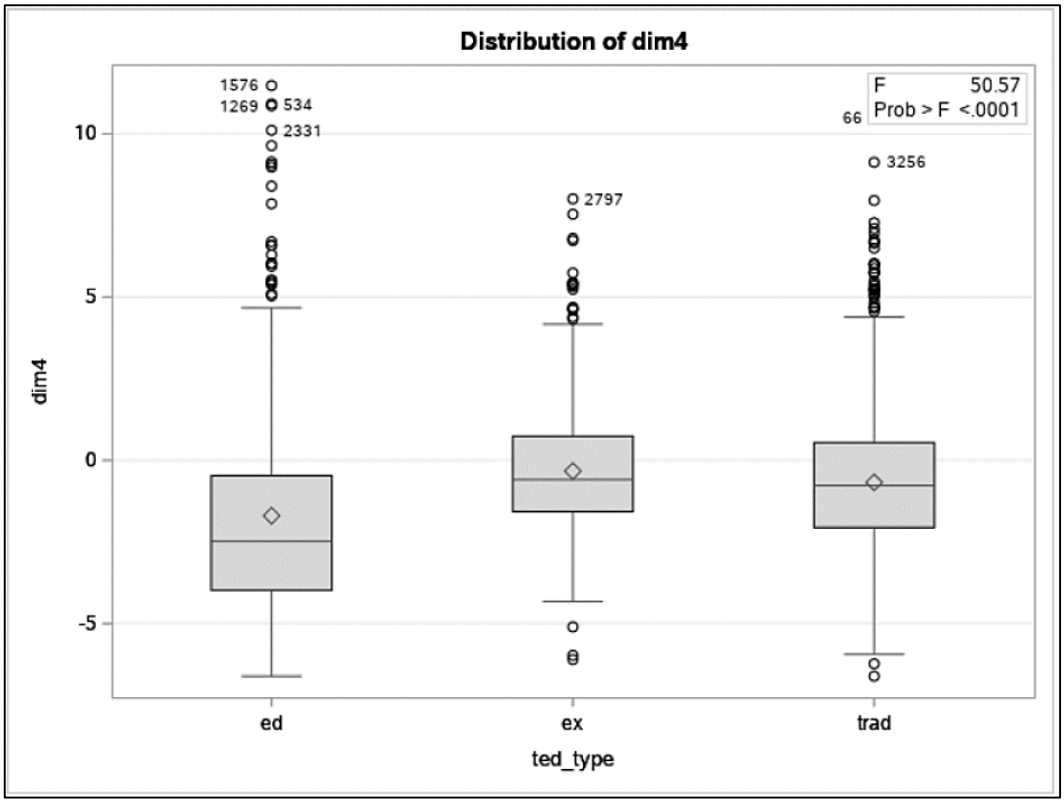


Figura 28: Distribuição – TED tradicional, TEDx e TED-Ed – Dimensão 4 (BIBER, 1988) – Fonte: SAS OnDemand for Academics.

Na figura 28, temos o gráfico de distribuição – chamado de "gráfico de caixa e bigode" – da TED tradicional, TEDx e TED-Ed na dimensão 4, feito pelo *SAS OnDemand for Academics*. Nele, podemos visualmente perceber uma aproximação entre a Ted tradicional, a TEDx, e a TED-Ed na dimensão 4.

4.1.6 Corpus TED Talks (CoTED) na Dimensão 5 – Estilo abstrato versus não abstrato

A dimensão 5 da língua inglesa também é caracterizada pela divisão ou oposição entre os polos positivo e negativo. No polo positivo da dimensão 5, temos uma produção textual com foco em um estilo mais abstrato, ou seja, com foco no uso de uma linguagem com alta densidade de aspectos formais e técnicos. Dos 23 registros da língua inglesa considerados por Biber (1988), os documentos oficiais e os religiosos são os que melhor representam o polo positivo dessa dimensão. Quanto ao polo negativo, encontramos uma produção textual menos abstrata, mais informal e menos técnica. Dos 23 registros considerados por Biber, as conversas telefônicas, conversas face a face e ficção romântica são os que melhor representam o polo negativo da dimensão 5. Logo abaixo temos o mapeamento do corpus TED Talks geral na Dimensão 5 de Biber (1988) – figura 29:

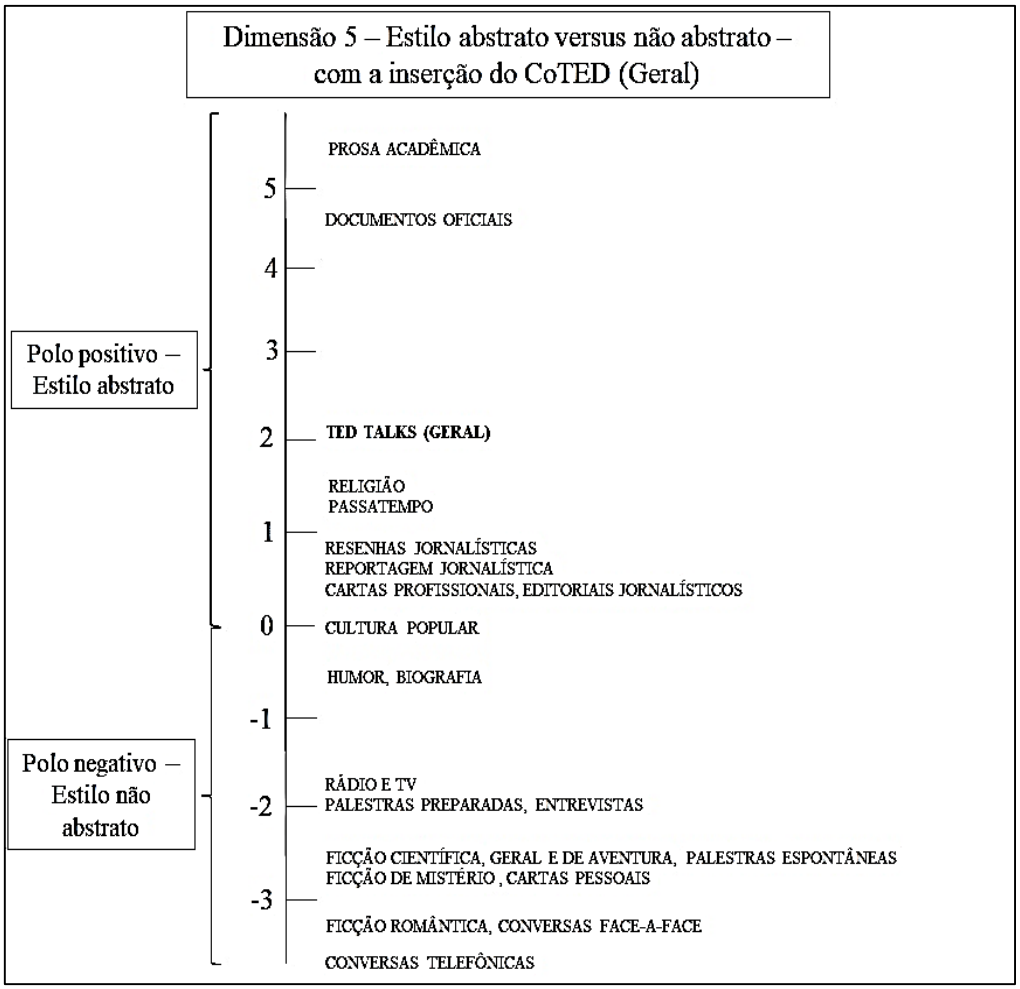


Figura 29: Dimensão 5 – Estilo abstrato versus não abstrato – com a inserção do CoTED (Geral).

Segundo o valor de escore médio na dimensão 5, conforme representado no quadro acima, a linguagem verbal das TED Talks geral se localiza na média 2.1004251 da tabela, estando no polo positivo. Essa posição pode nos indicar que a TED Talks tem características semelhantes aos documentos oficiais e aos textos religiosos. Veremos tais características mais a seguir. Logo abaixo, temos o mapeamento da TED tradicional, da TEDx e da TED-Ed na Dimensão 5 de Biber (1988) – figura 30:

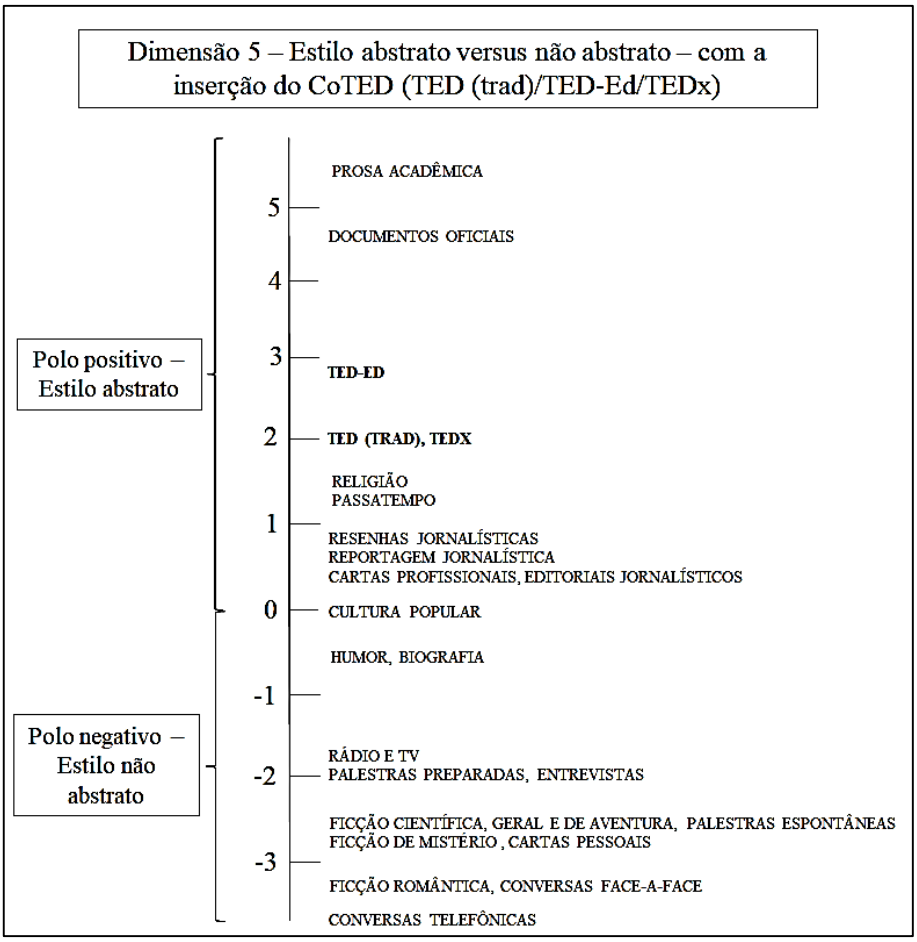


Figura 30: Dimensão 5 – Estilo abstrato versus não abstrato – com a inserção do CoTED (TED (trad)/TED-Ed/TEDx).

Considerando o mapeamento das três categorias TED tradicional (média 2.0073583, no polo positivo), TEDx (média 1.9806302, no polo positivo) e TED-Ed (média 2.8249265, no polo positivo), percebe-se que todos eles estão bastante próximos na tabela. TED-Ed, por sua vez, tende a ser mais relacionado aos documentos oficiais do que aos textos religiosos.

Para poder identificar os grupos das características linguísticas mais salientes e coocorrentes do CoTED tanto nos polos positivo e negativo da dimensão 5 de Biber (1988), temos que levar em consideração a estrutura do fator 5, conforme representado na tabela 24:

Estrutura do Fator 5 (BIBER, 1988)			
Estilo abstrato versus não abstrato			
Polo positivo		**Polo negativo**	
conjuntivos	0,48	(razão forma-ocorrência	-0,31) voz
passiva sem agente	0,43		
orações adjetivas reduzidas de particípio	0,42		
voz passiva com preposição ‘by’	0,41		

modificador pós-nominal	0,40
outros advérbios subordinativos	0,39
(adjetivo em posição predicativa)	0,31

Tabela 24: Estrutura do Fator 5 da língua inglesa (BIBER, 1988)

Ao considerar o escore médio do CoTED na dimensão 5 **(estilo abstrato versus não abstrato)**, foi separado o exemplo abaixo de um trecho retirado do texto *Let's talk crap. Seriously*. (TED tradicional, T2953_RG_TED13, 2013, média 2.1), considerado como um dos mais representativos no polo positivo – tendo em vista que o escore médio é positivo:

> *Let's talk dirty. A few years ago, oddly enough, I needed the bathroom,* ***and*** [conjunção] *I found one, a public bathroom,* ***and*** [conjunção] *I went into the stall,* ***and*** [conjunção] *I prepared to do what I'd done most of my life: use the toilet, flush the toilet, forget about the toilet.* ***And*** [conjunção] ***for*** [conjunção] *some reason that day,* ***instead*** [conjunção], *I asked myself a question,* ***and*** [conjunção] *it was, where does this stuff go?* ***And*** [conjunção] *with that question, I found myself plunged into the world of sanitation — there's more coming — (Laughter) — sanitation, toilets* ***and*** [conjunção] *poop,* ***and*** [conjunção] *I have* ***yet*** [conjunção] *to emerge.* ***And*** [conjunção] *that's* ***because*** [conjunção] *it's such an enraging,* ***yet*** [conjunção] *engaging place to be. To go back to that toilet, it wasn't a particularly fancy toilet, it wasn't* ***as*** [conjunção] *nice* ***as*** [conjunção] *this one from the World Toilet Organization. That's the other WTO. (Laughter)* ***But*** [conjunção] *it had a lockable door, it had privacy, it had water, it had soap* ***so*** [conjunção] *I could wash my hands,* ***and*** [conjunção] *I did because I'm a woman,* ***and*** [conjunção] *we do that. (Laughter) (Applause)* ***But*** [conjunção] *that day, when I asked that question, I learned something,* ***and*** [conjunção] *that was* ***that*** [conjunção] *I'd grown up thinking* ***that*** [conjunção] *a toilet like that was my right, when in fact it's a privilege.*
>
> *[...]*
>
> ***The flush toilet was voted the best medical advance of the last 200 years by the readers of the British Medical Journal*** [voz passiva com preposição 'by'], *and they were choosing over the Pill, anesthesia, and surgery.*
>
> *[...]*
>
> *And she was* ***scared*** [adjetivo em posição predicativa]. *She was* ***scared*** [adjetivo em posição predicativa] *of drunks hanging around. She was* ***scared*** [adjetivo em

posição predicativa] *of snakes. She was* ***scared*** [adjetivo em posição predicativa] *of rape.*

No trecho acima, podemos ver o uso de conjunções, de adjetivos em posição predicativa e da voz passiva, revelando uma produção oral dependente de contexto (externo ao texto) e caracterizada por um estilo mais abstrato, ou seja, existe o uso de uma linguagem com com aspectos de formalidade e tecnicidade, que remete ao uso da informação abstrata. E tal informação costuma exigir termos usados em determinadas áreas de conhecimento, assim como, classificar o vaso sanitário como um avanço da medicina comparável com a pílula, a anestesia e a cirurgia. Como não tivemos um valor muito alto de desvio-padrão, não foi separado um exemplo exclusivo do TED-Ed (ou do TEDx), conforme foi feito para a dimensão 1.

4.1.6.1 ANOVA do CoTED – Dimensão 5 (AMD Aditiva)

Considerando os resultados previamente apresentados na seção 3.1.5 das ANOVAs da Análise Funcional Aditiva do CoTED (tabela 12), referente à dimensão 5, temos as seguintes considerações: os resultados da ANOVA indicam uma variância não tão significativa em comparação com a dimensão 1, com F = 25.31, p = <.0001 e R2 = 0.014636, indicando que 1,5% é o percentual de predição de que a variação da linguagem verbal das TED Talks é explicada pela coocorrência das características linguísticas da dimensão 5 de Biber (1988). Apesar de baixa percentagem, ainda assim, podemos classificar a linguagem verbal das TED Talks no geral como mais focada em informação mais abstrata.

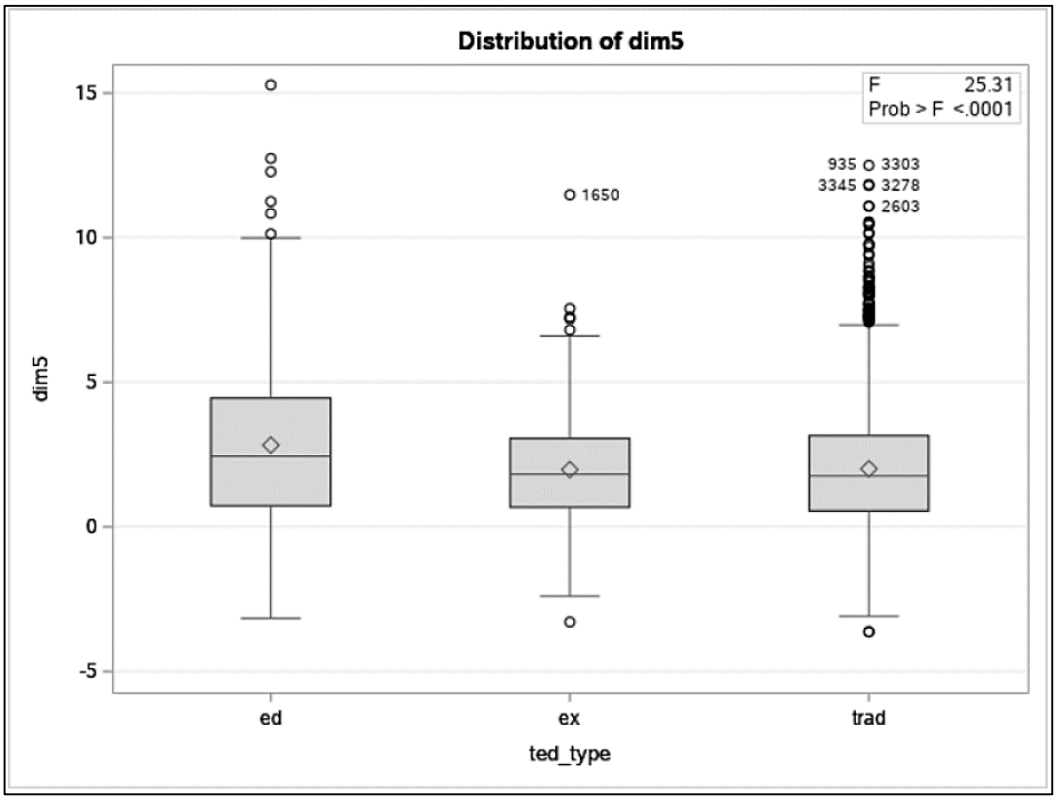


Figura 31: Distribuição – TED tradicional, TEDx e TED-Ed – Dimensão 5 (BIBER, 1988) – Fonte: SAS OnDemand for Academics.

Na figura 31, temos o gráfico de distribuição – chamado de "gráfico de caixa e bigode" – da TED tradicional, TEDx e TED-Ed na dimensão 5, feito pelo *SAS OnDemand for Academics*. Nele, podemos visualmente perceber uma aproximação entre a Ted tradicional, a TEDx, e a TED-Ed na dimensão 5.

4.2 Resultados da Análise Multidimensional Funcional Completa do Corpus TED Talks (CoTED)

Este é o momento da análise em que começamos a responder à segunda pergunta de pesquisa deste trabalho: Quais são as dimensões de variação do Corpus TED Talks (CoTED) sob a perspectiva da AMD Funcional Completa (BIBER, 1988)? As primeiras etapas da AMD Funcional Completa são bastante semelhantes àquelas da AMD Aditiva acima apresentada. Após o desenho, a compilação, a etiquetagem – por meio do *Biber Tagger* – e a contagem das 128 variáveis linguísticas do CoTED – por meio do *Biber Tag Count* –, também foi feito o uso do *SAS OnDemand for Academics* na análise fatorial. Contudo, ao invés de procurar classificar os textos do CoTED ao longo das cinco dimensões de variação dos registros falados e escritos

da língua inglesa considerados por Biber (1988), buscamos encontrar as dimensões de variação do próprio CoTED, ou seja, buscamos encontrar os parâmetros funcionais que caracterizam sua linguagem verbal. Após a Análise Fatorial inicial, com a qual obtivemos os pesos (*loadings*) de cada variável, os demais passos seguidos são (BERBER SARDINHA, 2004, p. 306):

- Determinação do número de fatores por meio da aplicação de técnicas, como observação dos valores eigen (*eigenvalues*) em um gráfico scree (*scree plot*).
- Análise Fatorial posterior, com a rotação dos fatores.
- Cálculo de escores de cada texto por fator, pela padronização dos escores com base na média e no desvio padrão.
- Cálculo dos escores médios de cada variedade por fator.
- Interpretação de cada fator e rotulação das dimensões.

Logo a seguir, na seção 4.2.1, serão apresentados valores eigen (*eigenvalues*) em um gráfico scree (*scree plot*). Logo depois, nas seções 4.2.2 até 4.2.5, serão apresentadas as quatro dimensões encontradas e o processo de interpretação de cada uma delas. Em suma, as dimensões funcionais do CoTED são:

- Dimensão 1: Discurso informacional versus discurso interacional;
- Dimensão 2: Discurso de convencimento ou persuasão;
- Dimensão 3: Discurso assertivo e conjectural;
- Dimensão 4: Discurso baseado em competências.

4.2.1 Gráfico scree (*scree plot*) e valores eigen (*eigenvalues*)

Na figura 32, temos o gráfico de sedimentação (chamado de gráfico scree ou *scree plot*) que representa os valores eigen (*eigenvalues*) – que são os autovalores dos fatores – os quais indicam a quantidade de variação explicada por cada fator. Deste modo, ao analisarmos o gráfico scree, notamos que os primeiros quatro fatores representam a maior parte da variabilidade total nos dados, pois aparecem antes do ponto de ruptura – comumente chamado de "cotovelo", indicando quais fatores contribuem mais para a análise geral. Os demais fatores trazem uma proporção relativamente pequena da variabilidade, não representando uma variação relevante passível de interpretação.

Figura 32: Gráfico scree plot do CoTED.

Logo abaixo, temos a tabela de variação (tabela 25) explicada de acordo com os quatro fatores, que mostra a quantidade de variação capturada por cada um deles. A porcentagem da variabilidade explicada pelo fator 1 é de 9,1%, a do fator 2 é de 4,6%, a do fator 3 é de 3,2% e a do fator 4 é de 2,1% (valores arredondados). A somatória dos índices individuais atinge 19% do total da variação. Em geral, o fator 1 costuma cobrir em maior escala a quantidade de variação do corpus, assim como ocorreu no caso do CoTED.

Variância explicada por cada fator: solução não rotacionada no CoTED			
Fator 1	**Fator 2**	**Fator 3**	**Fator 4**
9.0954709	4.5889020	3.1849807	2.0909929

Tabela 25: Variância explicada por cada fator: solução não rotacionada do CoTED (fatores 1-4)

4.2.2 Dimensão 1 – Discurso informacional versus discurso interacional

A dimensão 1 do CoTED é bastante semelhante à dimensão 1 da língua inglesa – **produção marcada por envolvimento versus produção informacional** (BIBER, 1988). A principal diferença é que os resultados carregaram em polos opostos, ou seja, o polo positivo da língua inglesa é o negativo do CoTED, e vice versa. Isso ocorreu porque as características linguísticas que possuem maior peso na linguagem verbal das TED Talks na dimensão 1 não têm o mesmo peso na dimensão 1 da língua inglesa, ou seja, características como substantivo, tamanho de palavra e preposição não tem o mesmo peso na dimensão 1 da língua inglesa (tabela 26) em comparação com a dimensão 1 do CoTED (tabela 27). Apesar disso, quanto à questão das polaridades invertidas, não existe a possibilidade de influenciar negativamente na análise, pois o que importa é a distribuição dos dois conjuntos de variáveis correlacionadas em um mesmo fator – as características no polo positivo coocorrem em todos os textos e as características do polo negativo também coocorrem em todos os textos, mas quando um texto tem muitas ocorrências de características com pesos positivos, ele tende a ter menos ocorrências de características com pesos negativos, e vice versa (BIBER, 1988, p. 101). Conforme Biber (1988, p. 20) nos explica, a dimensão 1 é uma dimensão mais universal, pois os padrões de variação observados nos estudos de AMD sustentam a probabilidade de parâmetros universais de variação de registro – assim como ocorre com a dimensão 1 da língua inglesa (BIBER, 1988) e a dimensão 1 da língua portuguesa (BERBER SARDINHA, KAUFFMANN e ACUNZO, 2014) – bem como a existência de dimensões únicas de variação em cada idioma e/ou domínio do discurso. (BIBER, 2019, p. 20).[102]

O fator 1 encontrado é o que define a dimensão 1 do CoTED. Neste caso, como tanto os polos positivo e negativo foram carregados, iremos expor seus resultados separadamente e comparando-os com os fatores encontrados por Biber (1988). Foram carregadas 6 características no polo positivo – com valores maiores que .30 – da dimensão 1 do CoTED (tabela 27). As características semelhantes com as do polo negativo da dimensão 1 da língua inglesa (tabela 26, BIBER, 1988) são: tamanho de palavras, preposições, razão forma-ocorrência e adjetivo em posição atributiva. Tais características já nos remetem a avaliar a dimensão 1 do CoTED como uma produção textual de caráter mais informacional, ou seja, ela tem mais características da linguagem escrita, assim como os documentos oficiais e as reportagens jornalísticas.

[102] Original: In sum, the patterns of variation observed across MD studies support the likelihood of universal parameters of register variation as well as the existence of unique dimensions of variation in each language and/or discourse domain. Future MD studies should further our understanding of both.

Estrutura do Fator 1 da língua inglesa - (BIBER, 1988)	
Polo negativo	
substantivo	-0,47
tamanho de palavra	-0,54
preposição	-0,54
razão forma-ocorrência	-0,58
adjetivo em posição atributiva	-0,80
(advérbio de lugar	-0,32)
(voz passiva sem agente	-0,38)
(oração adjetiva reduzida de particípio	-0,39)
(oração adjetiva reduzida de gerúndio	-0,42)

Tabela 26: Estrutura do Fator 1 da língua inglesa – polo negativo (BIBER, 1988).

Estrutura do Fator 1 (CoTED)	
Polo positivo	
tamanho de palavra	0,84417
adjetivos em posição atributiva	0,83555
adjetivos	0,73199
preposição	0,68764
razão forma-ocorrência	0,44152
modificador pós-nominal passivo	0,32734

Tabela 27: Estrutura do Fator 1 (CoTED) – polo positivo.

Ao considerar o escore médio dos textos do CoTED na dimensão 1 no polo positivo, 8.70002830857036, foi separado o exemplo abaixo de um trecho retirado do texto *How a long-forgotten virus could help us solve the antibiotics crisis* (TED tradicional, T0663_AB_TEDBCGT, 2018, média 8.6978079756), considerado como um dos mais representativos no polo positivo – tendo em vista que o escore médio é positivo:

> *The woman had a knee injury, required* ***multiple*** [adjetivo em posição atributiva] *surgeries, and* ***over*** [preposição] *the course* ***of*** [preposição] *these, developed a* ***chronic*** [adjetivo em posição atributiva] ***bacterial*** [adjetivo em posição atributiva] *infection* ***in*** [preposição] *her leg. Unfortunately* ***for*** [preposição] *her, the bacteria causing the infection also did not respond* ***to*** [preposição] *any antibiotic that was* ***available*** [adjetivo]. *So* ***at*** [preposição] *this point, typically, the* ***only*** [adjetivo em posição atributiva] *option left is* ***to*** [preposição] *amputate the leg* ***to*** [preposição] *stop the infection* ***from*** [preposição] *spreading further. Now, my father-in-law was* ***desperate*** [adjetivo] ***for*** [preposição] *a* ***different*** [adjetivo em posição atributiva]

kind ***of*** [preposição] *solution, and he applied* ***for*** [preposição] *an* ***experimental*** [adjetivo em posição atributiva], ***last-resort*** [adjetivo em posição atributiva] *treatment using phages. And guess what? It worked. Within three weeks* ***of*** [preposição] *applying the phages, the* ***chronic*** [adjetivo em posição atributiva] *infection had healed up, where before, no antibiotic was working. I was* ***fascinated*** [adjetivo em posição atributiva] ***by*** [preposição] *this* ***weird*** [adjetivo em posição atributiva] *conception: viruses curing an infection.*

No trecho acima, temos um amplo uso de adjetivos e preposições. Os adjetivos são amplamente usados em linguagem mais formal e elaborada (BIBER, 1995, p. 242). No caso dos adjetivos em posição atributiva, eles são utilizados para melhor elaborar ou explicar as informações nominais de substantivos (principais), ou seja, são formas muito mais integradas e condensadas à forma nominal (sintagma nominal) em comparação com os adjetivos em posição predicativa ou orações relativas (BIBER, 1988, p. 105). Além disso, adjetivos atributivos têm maior ocorrência do que os predicativos (BIBER *et al*., 2002, p. 188). No caso das preposições, os sintagmas preposicionais integram um alto nível de informações ao texto e costumam funcionar como modificadores de substantivos (núcleos da oração). Outra característica que também podemos considerar é o tamanho das palavras que, geralmente, são maiores em discursos não narrativos do que em discursos narrativos, conforme podemos ver na dimensão 1 da língua inglesa (BIBER, 1988). E como Biber explica (1988, p. 104), o tamanho das palavras marcam uma alta densidade de informação além de uma escolha lexical precisa para representar um conteúdo informativo. É interessante notar no trecho acima que, por mais que tenha um certo evento sendo narrado, existe toda uma explicação do que está acontecendo neste evento, sem denotar parecer ser uma história sendo contada, mas descrições explicativas do que aconteceu. Ademais, esse texto pode nos remeter a uma pessoa que tem conhecimento técnico sobre uma área.

No polo negativo da dimensão 1 do CoTED (tabela 28), temos uma produção textual de caráter de envolvimento ou interacional, assim como nas conversas telefônicas e nas conversas face a face. Foram carregadas 12 características no polo negativo – com valores maiores que .30 – na dimensão 1 do CoTED. As características semelhantes com as do polo positivo da dimensão 1 da língua inglesa (tabela 29, BIBER, 1988) são: verbos, contrações, apagamento de *that*, pronome de primeira pessoa, pronome de segunda pessoa, pronome indefinido, pronome *it*, preposição final e oração *wh*. Tais características já nos remetem a avaliar o polo negativo

da dimensão 1 do CoTED como uma produção textual como encontrada na linguagem falada – como nas conversas (simuladas ou não).

Estrutura do Fator 1 (CoTED)	
Polo negativo	
verbos (sem incluir os verbos auxiliares)	-0,81057
contrações	-0,61544
pronome de primeira pessoa/possessivo	-0,57708
pronome de segunda pessoa/possessivo	-0,52173
apagamento de 'that'	-0,52067
pronome nominal/indefinido	-0,39727
verbo modal preditivo	-0,38561
pronome 'it'	-0,36157
preposição final	-0,35897
todas as orações complementares com 'to' controladas por verbo	-0,34089
conjunções	-0,31789
oração 'wh'	-0,31705

Tabela 28: Estrutura do Fator 1 (CoTED) – polo negativo.

Estrutura do Fator 1 da língua inglesa - (BIBER, 1988)	
Polo positivo	
verbo privado	0,96
apagamento de 'that'	0,91
contração	0,90
verbo no tempo presente	0,86
pronome de segunda pessoa	0,86
verbo 'do'	0,82
negação analítica	0,78
pronome demonstrativo	0,76
enfatizador	0,74
pronome de primeira pessoa	0,74
pronome 'it'	0,71
'be' como verbo principal	0,71
subordinação causativa	0,66
partícula discursiva	0,66
pronome indefinido	0,62
advérbio delimitador/atenuador	0,58
advérbio/qualificador-amplificador	0,56
pronome relativo	0,55
pergunta 'wh'	0,52
verbo modal de possibilidade	0,50
coordenação não-frasal	0,48
oração 'wh'	0,47
preposição final	0,43
(advérbio	0,42)
(subordinação condicional	0,32)

Tabela 29: Estrutura do Fator 1 da língua inglesa – polo positivo (BIBER, 1988).

Ao considerar o escore médio dos textos do CoTED na dimensão 1 no polo negativo, -7.46978835312909, foi separado o exemplo abaixo de um trecho retirado do texto *Your brain on improv* (TEDx, T3658_CL_TEDXMA, 2010, média -7.467204931), considerado como um dos mais representativos no polo negativo – tendo em vista que o escore médio é negativo:

> *So **I*** [pronome de primeira pessoa] *am a surgeon **who studies creativity*** [oração-wh], ***and*** [conjunção] ***I*** [pronome de primeira pessoa] *have never **had a patient*** [apagamento de that] ***tell*** *me, "**I*** [pronome de primeira pessoa] *really want **you***

[pronome de segunda pessoa] *to be creative during surgery," **and*** [conjunção] *so **I*** [pronome de primeira pessoa] *guess **there's*** [redução] *a little bit of irony to **it*** [pronome it]. ***I*** [pronome de primeira pessoa] *will say though **that*** [conjunção], *after having done surgery a lot, **it*** [pronome it]***'s*** [redução] *similar to playing a musical instrument. **And*** [conjunção] ***for*** [conjunção] *me, this deep **and*** [conjunção] *enduring fascination with sound is **what led me to both be a surgeon and to study the science of sound*** [oração-wh], *particularly music. **I*** [pronome de primeira pessoa]***'m*** [redução] *going to talk over the next few minutes about **my*** [pronome possessivo de primeira pessoa] *career in terms of **how I*** [pronome de primeira pessoa e oração-wh]***'m*** [redução] ***able to study music and try to grapple with all these questions of how the brain is able to be creative*** [oração-wh]. ***I*** [pronome de primeira pessoa]***'ve*** [redução] *done most of this work at Johns Hopkins University, **and*** [conjunção] *at the National Institute of Health **where I*** [pronome de primeira pessoa e oração-wh] ***was previously**. **I** [pronome de primeira pessoa]**'ll*** [redução] *go over some science experiments **and*** [conjunção] *cover three musical experiments.*

No trecho acima, temos uma alta frequência de pronomes pessoais, principalmente o de primeira pessoa. Segundo Biber (1988), tais pronomes são uma referência direta ao emissor e ao destinatário, sendo frequentemente utilizados em discursos altamente interativos (1988, p. 105). Com relação ao pronome neutro, singular e de terceira pessoa, o *it*, sabe-se que seu uso é para substituir sintagmas ou orações em somente uma palavra (curta), sendo uma forma reduzida para se dizer algo (BIBER, 1988, p. 106). Falando especificamente sobre o pronome de primeira pessoa, Biber (2006, p. 12) explica que esse pronome é usado quando o falante está expressando seus próprios sentimentos e/ou atitudes. Desse modo, tais características apresentadas pelo pronome de primeira pessoa são muito presentes na linguagem oral (BIBER, 1995, p. 242). No caso do uso de orações-wh, é uma outra maneira de se falar sobre algumas questões, sem necessariamente fazer perguntas (BIBER, 1988, p. 107). As conjunções (principalmente o *and*), por sua vez, também são bastante frequentes, como no exemplo acima. Elas são marcadores da complexa relação entre as orações (BIBER, 1988, p. 112) e sua repetição são muito comuns na linguagem oral. Contudo, elas possuem uma variedade lexical consideravelmente baixa quando comparadas com as demais informações do discurso (BIBER, 1988, p. 112). Também, o apagamento de *that* é uma forma de se reduzir ou resumir o que está

sendo dito, ou sobre algo que não se tem certeza (BIBER, 1988, p. 106). Outra forma de redução são as contrações, as quais são relacionadas aos limites impostos em uma produção oral em tempo real. Tais características encontradas denotam mais um caráter interacional, ou seja, possuem um caráter mais próximo da linguagem falada.

Considerando todo o exposto acima, o nome dado à dimensão 1 do CoTED foi "**Discurso informacional versus Discurso interacional**", também semelhante ao nome original da dimensão 1 da língua inglesa, "**Produção marcada por envolvimento versus Produção informacional**" (BIBER, 1988).

4.2.3 Dimensão 2 – Discurso de convencimento ou persuasão

O fator 2 encontrado é o que define a dimensão 2 do CoTED. Neste caso, como o polo negativo não apresentou características significativas para que fosse interpretado como uma dimensão, apenas o polo positivo foi considerado. Foram carregadas 7 características no polo positivo – com valores maiores que .30 – da dimensão 2 do CoTED (tabela 30). Essa dimensão não se assemelha com a dimensão 2 de Biber (1988) – discurso narrativo versus discurso não narrativo –, por isso, não serão feitas comparações.

Estrutura do Fator 2 (CoTED)	
Polo positivo	
advérbio de probabilidade	0,66261
advérbio de certeza	0,63876
advérbio delimitador/atenuador	0,59787
advérbio/qualificador-amplificador	0,54163
advérbio quantificador enfático	0,47102
pronome demonstrativo	0,42658
quantidade de palavras	0,37886

Tabela 30: Estrutura do Fator 2 (CoTED).

Ao considerar o escore médio dos textos do CoTED na dimensão 2 no polo positivo, 3.79795518740931, foram separados dois trechos. O exemplo abaixo vem de um trecho retirado do texto *Terrorism is a failed brand* (TED tradicional, T3087_JMC_TEDGLO12, 2012, média 3.7904510839), considerado como um dos mais representativos no polo positivo – tendo em vista que o escore médio é positivo:

*We **most*** [advérbio quantificador enfático] ***certainly*** [advérbio de certeza] *do talk to terrorists, no question about it. We are at war with a new form of terrorism. It's **sort of*** [locução adverbial de probabilidade] *the good old, traditional form of terrorism, but it's **sort of*** [locução adverbial de probabilidade] *been packaged for the 21st century. One of the big things about countering terrorism is, how do you perceive it? Because perception leads to your response to it. So if you have a traditional perception of terrorism, it would be that it's one of criminality, one of war. So how are you going to respond to it? **Naturally*** [advérbio de certeza], *it would follow that you meet kind with kind. You fight it. If you have a **more*** [advérbio quantificador enfático] *modernist approach, and your perception of terrorism is almost cause-and-effect, then **naturally*** [advérbio de certeza] *from **that*** [pronome demonstrativo], *the responses that come out of it are much **more*** [advérbio quantificador enfático] *asymmetrical. We live in a modern, global world. Terrorists have actually adapted to it. It's something we have to, too, and **that*** [pronome demonstrativo] *means the people who are working on counterterrorism responses have to start, in effect, putting on their Google-tinted glasses, or whatever.*

O exemplo abaixo vem de um trecho retirado do texto *Organic algorithms in architecture* (TED tradicional, T0192_GL_TED05, 2005, média 3.802256549), considerado como um dos mais representativos no polo positivo – tendo em vista que o escore médio é positivo:

*It also is **much*** [advérbio/qualificador-amplificador] ***more*** [advérbio/qualificador-amplificador] *dynamic, so that you can see that the same form opens and closes in a **very*** [advérbio/qualificador-amplificador] *dynamic way as you move across it, because it has **this*** [pronome demonstrativo] *quality of vector in motion built into it. So the same space that appears to be a **kind of*** [locução adverbial de probabilidade] *closed volume, when seen from the other side becomes a **kind of*** [locução adverbial de probabilidade] *open vista. And you also get a sense of visual movement in the space, because every one of the elements is changing in a pattern, so **that*** [pronome demonstrativo] *pattern leads your eye towards the altar. I think that's one of the main changes, also, in architecture: that we're starting to look now not for some ideal form, like a Latin cross for a church, but actually all the traits of*

a church: so, light that comes from behind from an invisible source, directionality that focuses you towards an altar. It turns out it's not rocket science to design a sacred space. You just need to incorporate a certain number of traits in a very ***kind of*** [locução adverbial de probabilidade] *genetic way. So,* ***these*** [pronome demonstrativo] *are the different perspectives of* ***that*** [pronome demonstrativo] *interior, which has a* ***very*** [advérbio/qualificador-amplificador] *complex set of orientations all in a simple form.*

Conforme podemos ver em ambos os exemplos acima, temos um uso significativo de advérbios e muitos desses advérbios, como os de probabilidade e os quantificadores-enfáticos, expressam informações sobre uma ação, acontecimento ou posicionamento pessoal (BIBER et al, 1999, p.553; 2006, p. 12). No caso dos advérbios qualificadores-amplificadores, eles são geralmente usados para enfatizar ao informar determinados sentimentos (BIBER, 1988, p. 106). Desta forma, é possível perceber o quanto que os advérbios exercem um papel importante no convencimento do destinatário sobre algo. Também temos a presença do uso de pronomes demonstrativos os quais representam as referências nominais, podendo ser relacionados a descrições de pessoas, coisas, lugares etc. ou servir de referência entre a proximidade do emissor e do destinatário (animado ou não) (BIBER, 1988, p.106; 2006, p.12; 1999, p.70). Com relação à quantidade de palavras, podemos dizer que é uma medida de especificidade lexical e da extensão dos textos em quantidade de palavras. E essa quantidade é maior nesta dimensão, o que pode nos levar a analisar essa dimensão como mais densa em palavras para se transmitir algo considerado importante. Considerando todo o exposto acima, o nome dado à dimensão 2 do CoTED foi "**Discurso de convencimento ou persuasão**".

4.2.4 Dimensão 3 – Discurso assertivo e conjectural

O fator 3 encontrado é o que define a dimensão 3 do CoTED. Neste caso, como o polo negativo não apresentou características significativas para que fosse interpretado como uma dimensão, apenas o polo positivo foi considerado. Foram carregadas 7 características no polo positivo – com valores maiores que .30 – da dimensão 3 do CoTED (tabela 31). Essa dimensão não se assemelha com a dimensão 3 de Biber (1988) – referência dependente de situação versus elaborada –, por isso, não serão feitas comparações.

Estrutura do Fator 3 (CoTED)	
Polo positivo	
‘t*hat*’ usado em oração complementar controlada por verbo	0,62620
‘*that*’ usado em oração complementar controlada por verbo de probabilidade	0,46168
adjetivo em posição predicativa	0,40132
‘*that*’ usado em oração complementar controlada por substantivo não factivo	0,37973
‘*that*’ usado em oração complementar controlada por verbo factivo	0,37497
todas as categorias de orações complementares com 'that' controladas por adjetivos	0,32950
substantivos de cognição	0,31558

Tabela 31: Estrutura do Fator 3 (CoTED) – polo positivo.

Ao considerar o escore médio dos textos do CoTED na dimensão 3 no polo positivo, 3.16714129791515, foram separados dois trechos. O primeiro é um exemplo de um trecho retirado do texto *Why I turned Chicago’s abandoned homes into art* (TED tradicional, T3250_AW_TEDWOM18, 2018, média 3.1652634022), considerado como um dos mais representativos no polo positivo – tendo em vista que o escore médio é positivo:

> *I really love color. I notice it everywhere and in everything. My family makes fun of me because I like to use colors with elusive-sounding names, like celadon ... (Laughter) ecru ... carmine. Now, if you haven't noticed, I am* ***black*** [adjetivo em posição predicativa], *thank you — (Laughter) and when you grow up in a segregated city as I have, like Chicago, you're conditioned to believe* ***that*** [‘that’ usado em oração complementar controlada por verbo] *color and race can never be separate. There's hardly a day that goes by* ***that*** [‘that’ usado em oração complementar controlada por substantivo] *somebody is not reminding you of your color. Racism is my city's vivid hue. Now, we can all agree* ***that*** [‘that’ usado em oração complementar controlada por verbo] *race is a socially* ***constructed*** [adjetivo em posição predicativa] *phenomenon, but it's often hard to see it in our everyday existence.*
>
> *[...]*
>
> *He argues* ***that*** [‘that’ usado em oração complementar controlada por verbo] *the iconic color of a cola can is red, and that in fact all of us can agree* ***that*** [‘that’ usado em oração complementar controlada por verbo de probabilidade] *it's red but the kinds of reds* ***that*** [‘that’ usado em oração complementar controlada por substantivo não factivo] *we imagine are as varied as the number of people in this*

room. So imagine ***that*** [‘that’ usado em oração complementar controlada por verbo]. *This color* ***that*** [‘that’ usado em oração complementar controlada por substantivo não factivo] *we've all been taught since kindergarten is primary — red, yellow, blue — in fact is not primary, is not irreducible, is not objective but quite subjective.*

O segundo exemplo é de um trecho retirado do texto *Does time exist?* (TED-Ed, T0697_AZJ_TEDED, 2018, média 3.1620692278), considerado como um dos mais representativos no polo positivo – tendo em vista que o escore médio é positivo:

Einstein’s theory seemed to confirm ***that*** [that usado em oração complementar controlada por verbo] *time is woven into the very fabric of the universe. But there’s a big question it didn’t fully resolve: why is it we can move through space in any direction, but through time in only one? No matter what we do, the past is always, stubbornly, behind us. This is called the arrow of time. When a drop of food coloring is dropped into a glass of water, we instinctively know* ***that*** [that usado em oração complementar controlada por verbo] *the coloring will drift out from the drop, eventually filling the glass. Imagine watching the opposite happen. Here, we’d recognize time as unfolding backwards. We live in a universe where the food coloring spreads out in the water, not a universe where it collects together. In physics, this is described by the Second Law of Thermodynamics, which says* ***that*** [that usado em oração complementar controlada por verbo] *systems will gain disorder, or entropy, over time. Systems in our universe move from order to disorder, and it is* ***that*** [that usado em oração complementar controlada por verbo] *property of the universe* ***that*** [‘that’ usado em oração complementar controlada por substantivo não factivo] *defines the direction of time’s arrow. So if time is such a fundamental property, it should be in our most fundamental equations describing the universe, right? We currently have two sets of equations* ***that*** [‘that’ usado em oração complementar controlada por substantivo não factivo] *govern physics.*

[…]

The movement is ***real*** [adjetivo em posição predicativa], *yet also an illusion. Could the physics of time somehow be a similar illusion? Physicists are still exploring these and other questions, so we’re far from a complete explanation. At least for*

the moment.

Nos trechos acima, encontramos o uso de adjetivos em posição predicativa e o uso do *that* em orações complementares. Os adjetivos em posição predicativa caracterizam o sintagma nominal e, geralmente, ocorrem após um verbo de ligação (BIBER *et al*., 2002, p. 188). Ademais, esses adjetivos costumam ser mais frequentes em registros do modo escrito (informativo) do que nos de modo falado (interativo), assim como os adjetivos em posição atributiva (BIBER, 1995, p. 242). Contudo, os adjetivos predicativos têm uma frequência geralmente menor (BIBER *et al*., 2002, p. 188). No geral, os adjetivos funcionam como marcadores de posicionamento do emissor (ou autor do texto) para expressar suas opiniões (BIBER, 1995, p. 242; BIBER et al., 1999; BIBER, 2006a, p. 90). No caso do uso do *that* em orações complementares, isso revela um discurso marcado pela informalidade e o não planejamento das falas. Contudo, esse aspecto contínuo e de ações ainda em andamento são um tanto improvável, considerando que os textos são previamente escritos, o que nos leva a considerar ser um discurso simulado. Tal simulação traz afirmações e conjecturas por parte do falante (ou autor do texto) sobre o assunto tratado. Considerando todo o exposto acima, o nome dado à dimensão 3 do CoTED foi "**Discurso assertivo e conjectural**".

4.2.5 Dimensão 4 – Discurso baseado em competências

O fator 4 encontrado é o que define a dimensão 4 do CoTED. Neste caso, tanto os polos positivo e negativo foram carregados. Contudo, apesar de o polo positivo ter sido carregado, ele apresenta somente 2 características linguísticas com valores maiores que .30 (tabela 32). Por tal motivo, iremos expor ambos os resultados somente por via de comparação, mas não serão considerados aqui na interpretação do fator 4 – isso porque tal resultado não nos possibilita uma análise multidimensional das características linguísticas como proposto neste trabalho. Quanto ao polo negativo da dimensão 4 do CoTED, foram carregadas somente 3 características com valores maiores que .30 (tabela 33). Essa dimensão não se assemelha com a dimensão 4 de Biber (1988) – argumentação explícita –, por isso, não serão feitas comparações.

Estrutura do Fator 4 (CoTED)	
Polo positivo	
pronome de terceira pessoa (exceto 'it')	0,57623

substantivos animados	0,55793

Tabela 32: Estrutura do Fator 4 (CoTED) – polo positivo.

Estrutura do Fator 4 (CoTED)	
Polo negativo	
verbos modais de possibilidade, permissão e habilidade	-0,39801
substantivos concretos	-0,38513
substantivo relacionado a assuntos técnicos ou concretos	-0,32380

Tabela 33: Estrutura do Fator 4 (CoTED) – polo negativo.

Ao considerar o escore médio dos textos do CoTED na dimensão 4 no polo negativo, -2.54031106880997, foram separados dois exemplos. O exemplo abaixo é de um trecho retirado do texto *What your smart devices know (and share) about you* (TED tradicional, T0859_KHSM_TED18, 2018, média -2.545297663), considerado como um dos mais representativos no polo negativo – tendo em vista que o escore médio é negativo:

> *Being smart means the **device*** [substantivo relacionado a assuntos técnicos ou concretos] ***can*** [verbo modal de possibilidade, permissão e habilidade] *connect to the **internet*** [substantivo relacionado a assuntos técnicos ou concretos], *it **can*** [verbo modal de possibilidade, permissão e habilidade] *gather **data*** [substantivo relacionado a assuntos técnicos ou concretos], *and it **can*** [verbo modal de possibilidade, permissão e habilidade] *talk to its owner. But once your **appliances*** [substantivo relacionado a assuntos técnicos ou concretos] **can** [verbo modal de possibilidade, permissão e habilidade] *talk to you, who else are they going to be talking to? I wanted to find out, so I went all-in and turned my one-bedroom apartment in San Francisco into a smart **home*** [substantivo composto relacionado a assuntos técnicos ou concretos]. *I even connected our bed to the **internet*** [substantivo relacionado a assuntos técnicos ou concretos]. *As far as I know, it was just measuring our sleeping habits. I **can*** [verbo modal de possibilidade, permissão e habilidade] *now tell you that the only thing worse than getting a terrible night's sleep is to have your smart **bed*** [substantivo composto relacionado a assuntos técnicos ou concretos] *tell you the next day that you "missed your goal and got a low sleep score." (Laughter) It's like, "Thanks, smart **bed*** [substantivo relacionado a assuntos técnicos ou concretos]. *As if I didn't already feel like shit today."*

(Laughter) All together, I installed 18 internet-connected ***devices*** [substantivo relacionado a assuntos técnicos ou concretos] *in my home. I also installed a Surya. Surya Mattu: Hi, I'm Surya. (Laughter) I monitored everything the smart* ***home*** [substantivo relacionado a assuntos técnicos ou concretos] *did. I built a special* ***router*** [substantivo relacionado a assuntos técnicos ou concretos] *that let me look at all the* ***network*** [substantivo relacionado a assuntos técnicos ou concretos] *activity. You* ***can*** [verbo modal de possibilidade, permissão e habilidade] *think of my* ***router*** [substantivo relacionado a assuntos técnicos ou concretos] *sort of like a security* ***guard*** [substantivo relacionado a assuntos técnicos ou concretos], *compulsively logging all the network* ***packets*** [substantivo relacionado a assuntos técnicos ou concretos] *as they entered and left the smart* ***home*** [substantivo relacionado a assuntos técnicos ou concretos].

O próximo exemplo é de um trecho retirado do texto *How to use data to make a hit TV show* (TEDx, T1987_SW_TEDXCAM, 2015, média -2.547474462), considerado como um dos mais representativos no polo negativo – tendo em vista que o escore médio é negativo:

And now the crucial thing is that ***data*** [substantivo relacionado a assuntos técnicos ou concretos] *and* ***data*** [substantivo relacionado a assuntos técnicos ou concretos] *analysis is only good for the first* ***part*** [substantivo relacionado a assuntos técnicos ou concretos]. ***Data*** [substantivo relacionado a assuntos técnicos ou concretos] *and* ***data*** [substantivo relacionado a assuntos técnicos ou concretos] *analysis, no matter how powerful,* ***can*** [verbo modal de possibilidade, permissão e habilidade] *only help you taking a problem apart and understanding its* ***pieces*** [substantivo relacionado a assuntos técnicos ou concretos]. *It's not suited to put those* ***pieces*** [substantivo relacionado a assuntos técnicos ou concretos] *back together again and then to come to a conclusion. There's another* ***tool*** [substantivo relacionado a assuntos técnicos ou concretos] *that* ***can*** [verbo modal de possibilidade, permissão e habilidade] *do that, and we all have it, and that* ***tool*** [substantivo relacionado a assuntos técnicos ou concretos] *is the* ***brain*** [substantivo relacionado a assuntos técnicos ou concretos]. *If there's one thing a* ***brain*** [substantivo relacionado a assuntos técnicos ou concretos] *is good at, it's taking* ***bits*** [substantivo relacionado a assuntos técnicos ou concretos] *and* ***pieces*** [substantivo relacionado a assuntos técnicos ou concretos]

back together again, even when you have incomplete information, and coming to a good conclusion, especially if it's the ***brain*** [substantivo relacionado a assuntos técnicos ou concretos] *of an expert. And that's why I believe that Netflix was so successful, because they used* ***data*** [substantivo relacionado a assuntos técnicos ou concretos] *and* ***brains*** [substantivo relacionado a assuntos técnicos ou concretos] *where they belong in the process. They use* ***data*** [substantivo relacionado a assuntos técnicos ou concretos] *to first understand lots of* ***pieces*** [substantivo relacionado a assuntos técnicos ou concretos] *about their* ***audience*** [substantivo relacionado a assuntos técnicos ou concretos] *that they otherwise wouldn't have been* ***able to*** [verbo modal de possibilidade, permissão e habilidade] *understand at that* ***depth*** [substantivo relacionado a assuntos técnicos ou concretos], *but then the decision to take all these* ***bits*** [substantivo relacionado a assuntos técnicos ou concretos] *and* ***pieces*** [substantivo relacionado a assuntos técnicos ou concretos] *and put them back together again and make a show* [substantivo relacionado a assuntos técnicos ou concretos] *like "House of Cards," that was nowhere in the* ***data*** [substantivo relacionado a assuntos técnicos ou concretos]. *Ted Sarandos and his* ***team*** [substantivo relacionado a assuntos técnicos ou concretos] *made that decision to license that* ***show*** [substantivo relacionado a assuntos técnicos ou concretos], *which also meant, by the way, that they were taking a pretty big personal risk with that decision. And Amazon, on the other* ***hand*** [substantivo relacionado a assuntos técnicos ou concretos], *they did it the wrong way around. They used* ***data*** [substantivo relacionado a assuntos técnicos ou concretos] *all the way to drive their decision-making, first when they held their competition of* ***TV*** [substantivo relacionado a assuntos técnicos ou concretos] *ideas, then when they selected "Alpha House" to make as a* ***show*** [substantivo relacionado a assuntos técnicos ou concretos]. *Which of course was a very safe decision for them, because they* ***could*** [verbo modal de possibilidade, permissão e habilidade] *always point at the* ***data*** [substantivo relacionado a assuntos técnicos ou concretos], *saying, "This is what the* ***data*** [substantivo relacionado a assuntos técnicos ou concretos] *tells us."*

Nos exemplos acima, temos o uso bastante presente de substantivos, os quais carregam grande densidade de informações, sendo os principais portadores de significado referencial em um texto (BIBER, 1988, p. 104). A linguagem escrita, ou a linguagem formal e padronizada, é

geralmente caracterizada pelo uso de substantivos de forma a integrar ou elaborar informações (BIBER, 1995, p. 242). Assim, percebe-se que, nesta dimensão existe uma certa importância dada ao uso de substantivos concretos e/ou técnicos, trazendo o foco no conhecimento específico em determinadas áreas. Outra característica mais presente é o uso de verbos modais de possibilidade, permissão e/ou habilidade. Tais verbos podem sinalizar incerteza ou falta de precisão na apresentação de algum tipo de informação; ou em declarações sobre habilidades ou sobre possibilidades com relação a certos prováveis eventos (BIBER, 1988, p.111). Neste caso, a questão da habilidade está muito mais presente, o que corrobora com o foco no conhecimento concreto e/ou técnico. Considerando todo o exposto acima, o nome dado à dimensão 4 do CoTED foi "**Discurso baseado em competências**".

4.3 Análise do Modelo Linear Geral (GLM) das TED Talks (CoTED)

Este é o momento em que respondemos a terceira questão da pesquisa: Como se dá a variação multidimensional funcional em termos das variáveis independentes "apresentador" e "evento"? Antes mesmo de obter todos os resultados das AMD Funcional Aditiva e AMD Funcional Completa, havia uma pressuposição de que variáveis externas à linguagem verbal das TED Talks pudessem exercer influência sob ela, ou seja, havia uma pressuposição de que variáveis independentes como "apresentador" – que corresponde aos 2.845 apresentadores/ autores dos vídeos/textos das TED Talks analisadas –, e "evento" – que corresponde aos 412 títulos de eventos encontrados – poderiam exercer influência na variação linguística da linguagem verbal das TED Talks. Até o momento, viu-se que a divisão das TED Talks tem mais significação na dimensão 1 da língua inglesa, que postula a segmentação entre linguagem falada e linguagem escrita. Porém, os valores baixos de R2 demonstravam que ainda existia algo a mais a ser explicado. Desta forma, ao se efetuar a Análise do Modelo Linear Geral (GLM) das TED Talks (CoTED) – considerando as 4 dimensões da linguagem verbal das TED Talks com a AMD Funcional Completa –, foi constatado que tanto os apresentadores quanto os eventos exercem sim significativa influência na variação, conforme podemos ver nos resultados da tabela 15, descritos a seguir:

- Dimensão 1: os resultados da ANOVA indicam uma variância significativa, com F = 4.26, p = <.0001 e R2 = 0.365712, indicando que 36,6% da variação é influenciada pela variável "evento".

- Dimensão 2: os resultados da ANOVA indicam uma variância significativa, com F = 4.60, p = <.0001 e R2 = 0.383898, indicando que 38,4% da variação é influenciada pela variável "evento".
- Dimensão 3: os resultados da ANOVA indicam uma variância significativa, com F = 1.38, p = <.0001 e R2 = 0.157239, indicando que 15,7% da variação é influenciada pela variável "evento".
- Dimensão 4: os resultados da ANOVA indicam uma variância significativa, com F = 1.49, p = <.0001 e R2 = 0.167496, indicando que 16,7% da variação é influenciada pela variável "evento".

Considerando que outras variáveis estejam exercendo essa influência, não é surpresa que, no caso da variável independente "apresentador", também obtivemos um resultado significativo (ver tabela 15) – descrito logo a seguir:

- Dimensão 1: os resultados da ANOVA indicam uma variância significativa, com F = 1.49, p = <.0001 e R2 = 0.342302, indicando que 34,2% da variação é influenciada pela variável "apresentador".
- Dimensão 2: os resultados da ANOVA indicam uma variância significativa, com F = 1.37, p = <.0001 e R2 = 0.323247, indicando que 32,3% da variação é influenciada pela variável "apresentador".
- Dimensão 3: os resultados da ANOVA indicam uma variância significativa, com F = 1.26, p = <.0001 e R2 = 0.303327, indicando que 30,3% da variação é influenciada pela variável "apresentador".
- Dimensão 4: os resultados da ANOVA indicam uma variância significativa, com F = 1.33, p = <.0001 e R2 = 0.315607, indicando que 31,6% da variação é influenciada pela variável "apresentador".

Podemos assim, perceber que a linguagem verbal das TED Talks tem uma complexidade maior na sua descrição e definição. As 5 dimensões da língua inglesa explicam parcialmente como ela funciona, mas os fatores externos a ela têm considerável relevância que precisa ser considerada. Talvez, seja por isso que existem afirmações de que existe um formato ou estilo TED, distinto de qualquer outro formato de evento, de palestra e até de vídeo educativo.

5. Considerações Finais

Chegamos na parte do trabalho em que trazemos algumas considerações sobre a pesquisa e seus resultados. Desta forma, serão apresentadas considerações sobre a Análise Multidimensional Funcional Aditiva (AMD Aditiva) do Corpus das TED Talks – o CoTED – e sobre a Análise de Variância (ANOVA), tanto das TED Talks em geral, quanto das suas três categorias – TED tradicional, TEDx e TED-Ed, consideradas no presente trabalho; de forma comparativa, serão apresentadas considerações sobre a Análise Multidimensional Funcional Completa (AMD Completa) do Corpus das TED Talks – o CoTED – e sobre a Análise do Modelo Linear Geral (GLM) do CoTED. Em acréscimo, também serão feitas algumas considerações a respeito dos resultados desta pesquisa em relação aos das pesquisas sobre as TED Talks anteriormente citadas, e serão apresentadas as limitações e as possibilidades em pesquisas futuras a partir dos achados deste trabalho.

Conforme previamente discutido, é possível encontrar pessoas fazendo referências, usando, estudando, pesquisando, analisando e coletando (seção 2.1) a linguagem verbal das TED Talks. Contudo, ainda não existem pesquisas sobre essa linguagem sob a perspectiva da AMD. Deste modo, entender o funcionamento da linguagem usada pelas TED Talks significa entender a criação de um modo contemporâneo de pensar o mundo que, além de disseminar ideias, se multiplicou, criou comportamentos e meios de agir. E por ser considerado como um molde em uma nova onda de influenciadores, existia uma lacuna nos estudos sobre sua linguagem verbal. Portanto, foi proposto nesta pesquisa verificar quais são as características linguísticas – gramático-funcionais – usadas, e como essas características coocorrem para criar o discurso típico das TED Talks. Para tal empreitada, foram elencados três objetivos principais: 1) verificar se as TED Talks podem ser classificadas como um registro distinto composto por três sub-registros: TED Tradicional, TEDx e TED-Ed; 2) comparar as TED Talks e seus três sub-registros com os parâmetros de variação da língua inglesa encontrados por Biber (1988); e 3) verificar se há variação na linguagem verbal das TED Talks e, se houver, quais são os parâmetros que movem tal variação. Para alcançar tais objetivos, foram levantadas três perguntas: 1) como o corpus das TED Talks (CoTED) se encaixa nas dimensões de variação da língua inglesa encontradas por Biber (1988)? – sendo respondida por meio da AMD Funcional Aditiva; 2) quais são as dimensões de variação do corpus das TED Talks (CoTED) sob a perspectiva da AMD Funcional Completa? – sendo respondida por meio da AMD Funcional

Completa; e 3) Como se dá a variação multidimensional funcional em termos das variáveis independentes "apresentador" e "evento" do corpus das TED Talks (CoTED)? – sendo respondida por meio da AMD Funcional Completa. Com o intuito de responder às questões de pesquisa e alcançar os objetivos propostos, a presente pesquisa adotou a Análise Multidimensional Funcional, nas versões Aditiva e Completa (BIBER, 1988; 2009) – que se baseiam na pesquisa de corpora utilizando ferramentas computacionais especializadas. Para tal propósito, foram coletadas transcrições de 3.411 vídeos TED – formando o corpus chamado CoTED.

Com a AMD Funcional Aditiva, os textos do corpus CoTED foram mensurados segundo as dimensões da língua inglesa encontradas por Biber (1988). Considerando os resultados obtidos, respondemos a primeira pergunta da pesquisa: como o corpus das TED Talks (CoTED) se encaixa nas dimensões de variação da língua inglesa encontradas por Biber (1988)?

Levando em conta a dimensão 1 da língua inglesa – Produção marcada por envolvimento versus informacional –, podemos dizer que, de acordo com o seu mapeamento, as TED Talks – no geral – se aproximam mais de entrevistas, palestras espontâneas e cartas pessoais. Isso significa que temos uma produção textual de caráter de envolvimento ou interacional, como encontrado na linguagem falada – assim como nas conversas (simuladas ou não). Por sua vez, considerando o mapeamento das três categorias, TED tradicional, TEDx e TED-Ed, a princípio, vê-se uma clara distinção entre TED-Ed com relação ao TED tradicional e ao TEDx na dimensão 1. De início, isso apontaria que, por mais que os vídeos TED-Ed sejam indicados como vídeos TED Talks e que sigam o formato ou estilo TED, existe sim uma diferença nas características linguísticas por eles apresentados. Tais características se aproximam das palestras preparadas e das ficções. Isso nos levaria a concluir que, existe muito pouco envolvimento ou situação interacional na sua linguagem. Porém, conforme anteriormente visto, a variação das TED-Ed – além das Ted tradicionais e TEDx – na dimensão 1 também é bastante considerável, ou seja, tanto as TED Talks no geral quanto as 3 categorias aqui definidas apresentam uma variação no mapeamento da dimensão 1 da língua inglesa, sendo ora mais interacional e ora mais informacional. É por isso que foi descartada a possibilidade de se considerar as 3 categorias como sub-registros das TED Talks. Em acréscimo, com a ANOVA da dimensão 1 da língua inglesa sob as TED Talks (no geral), foi possível predizer que, por volta de 20% da variação é explicada pela dimensão 1 da língua inglesa. Contudo, apesar de ser um valor considerável, ainda existem outras questões presentes que influenciam na linguagem

verbal das TED Talks. Esse pode ser um fator para se pensar as TED Talks como um registro adaptável, sendo ora mais interacional e ora mais informacional.

Considerando a dimensão 2 da língua inglesa – Discurso narrativo versus não narrativo –, podemos dizer que, no geral, de acordo com o mapeamento das TED Talks, elas se aproximam mais de resenhas jornalísticas e conversas ao telefone. Isso significa que temos uma produção textual com foco não narrativo. Por sua vez, considerando o mapeamento das três categorias, TED tradicional, TEDx e TED-Ed, percebe-se que tanto a TED Talks Geral quanto às suas três categorias se encontram muito próximas umas das outras. Isso indica que elas se assemelham aos registros das resenhas jornalísticas, conversas ao telefone e cartas profissionais, todos os quais, apresentam uma produção textual com foco no discurso não narrativo. Assim, temos outro indício de que os três sub-registros não caberiam aqui. Em acréscimo, com a ANOVA da dimensão 2 da língua inglesa sob as TED Talks (no geral), foi possível predizer que menos de 2,0% da variação da linguagem verbal das TED Talks é explicada pela coocorrência das características linguísticas da dimensão 2 da língua inglesa. Apesar de ser um valor baixo, ainda assim, podemos classificar a linguagem verbal das TED Talks no geral como menos narrativa. Contudo, talvez, esse pode ser mais um fator para se pensar as TED Talks como um registro adaptável, podendo ser ora mais narrativo ora menos narrativo.

Com relação à dimensão 3 da língua inglesa – Referência dependente de situação versus elaborada –, considerando o mapeamento das TED Talks no geral, podemos dizer que elas têm características semelhantes às entrevistas e reportagens jornalísticas. Isso significa que, temos uma produção textual de situação elaborada, ou seja, as referências são dependentes de contexto, sendo elas externas ao próprio texto. Por sua vez, considerando o mapeamento das três categorias, TED tradicional, TEDx e TED-Ed, percebe-se que, tanto a TED tradicional quanto a TEDx estão praticamente no mesmo mapeamento da TED geral, ou seja, elas possuem características semelhantes às entrevistas e reportagens jornalísticas. Somente a TED-Ed sofreu uma pequena alteração no seu mapeamento, assemelhando-se mais às palestras preparadas e passatempos. Em acréscimo, com a ANOVA da dimensão 3 da língua inglesa sob as TED Talks (no geral), por volta de 1,7% é o percentual de predição de que a variação da linguagem verbal das TED Talks é explicada pela coocorrência das características linguísticas da dimensão 3 da língua inglesa. Mais uma vez, o valor é baixo, mas ainda sim considerável. Porém, talvez, esse pode ser mais um fator para se pensar as TED Talks como um registro adaptável, podendo ser ora mais dependente de contexto ora menos dependente de contexto.

Levando em conta a dimensão 4 da língua inglesa – Argumentação explícita –, podemos dizer que, no geral, as TED Talks têm características semelhantes às biografias, ficções e reportagens jornalísticas. Isso significa que, temos uma produção textual menos relacionada à argumentação explícita. Por sua vez, considerando o mapeamento das três categorias, TED tradicional, TEDx e TED-Ed, percebe-se que tanto a TED tradicional quanto a TEDx estão bastante próximas quanto às suas características, apesar de o TED-Ed ter uma maior aproximação à ficção de aventura. Em acréscimo, com a ANOVA da dimensão 4 da língua inglesa sob as TED Talks (no geral), por volta de 2,9% é o percentual de predição de que a variação da linguagem verbal das TED Talks é explicada pela coocorrência das características linguísticas da dimensão 4 da língua inglesa. O valor é relativamente baixo, porém, ainda considerável. Talvez, esse pode ser mais um fator para se pensar as TED Talks como um registro adaptável, sendo ora mais persuasivo ora menos persuasivo.

Considerando a dimensão 5 da língua inglesa – Estilo abstrato versus não abstrato –, podemos dizer que, no geral, as TED Talks têm características semelhantes aos documentos oficiais e aos textos religiosos. Isso significa que, temos uma produção textual com foco em um estilo mais abstrato, ou seja, com foco no uso de uma linguagem com alta densidade de aspectos formais e técnicos. Por sua vez, considerando o mapeamento das três categorias, TED tradicional, TEDx e TED-Ed, percebe-se que todos eles estão bastante próximos. TED-Ed, por sua vez, tende a ser mais relacionado aos documentos oficiais do que aos textos religiosos. Em acréscimo, com a ANOVA da dimensão 5 da língua inglesa sob as TED Talks (no geral), foi possível predizer que, por volta de 1,5% é o percentual de predição de que a variação da linguagem verbal das TED Talks é explicada pela coocorrência das características linguísticas da dimensão 5 da língua inglesa. O valor é relativamente baixo, porém, ainda considerável. Talvez, esse pode ser mais um fator para se pensar as TED Talks como um registro adaptável, sendo ora mais de estilo abstrato ora menos de estilo abstrato.

Com a AMD Funcional Completa, foi feita a extração fatorial completa e, considerando os resultados obtidos, respondemos a segunda pergunta da pesquisa: Quais são as dimensões de variação do corpus das TED Talks (CoTED) sob a perspectiva da AMD Funcional Completa?

A dimensão 1 do CoTED – Discurso informacional versus discurso interacional – é bastante semelhante à dimensão 1 da língua inglesa – Produção marcada por envolvimento versus informacional. A principal diferença é que os resultados da extração fatorial do CoTED carregaram em polos opostos, ou seja, o polo positivo da língua inglesa é o negativo do CoTED,

e vice-versa. Isso ocorreu porque as características linguísticas que possuem maior peso na linguagem verbal das TED Talks na dimensão 1 não têm o mesmo peso na dimensão 1 da língua inglesa, ou seja, características como substantivo, tamanho de palavra e preposição não tem o mesmo peso na dimensão 1 da língua inglesa em comparação com a dimensão 1 do CoTED. A dimensão 1 do CoTED – no polo positivo – se caracteriza como uma produção textual de caráter mais informacional, ou seja, ela tem mais características da linguagem escrita, assim como os documentos oficiais e as reportagens jornalísticas. No polo negativo da dimensão 1 do CoTED, temos uma produção textual de caráter de envolvimento ou interacional, assim como nas conversas telefônicas e nas conversas face a face. Por tal motivo, o nome dado à dimensão 1 do CoTED foi "**discurso informacional versus discurso interacional**". Mas é interessante notar que, nesse caso, o caráter informacional das TED Talks foi mais saliente, em contrapartida com os resultados da AMD Aditiva. Esse pode ser um fator para se pensar as TED Talks como um registro adaptável, sendo ora mais informacional/interacional ora menos informacional/interacional. Em acréscimo, com a GLM da dimensão 1 do CoTED, – respondendo a terceira questão da pesquisa: como se dá a variação multidimensional funcional em termos das variáveis independentes "apresentador" e "evento"? – foi constatado que as variáveis externas exercem significativa influência na variação. Na dimensão 1 do CoTED, os resultados indicam uma variância significativa, indicando que 36,6% da variação é influenciada pela variável "evento" e que 34,2% da variação é influenciada pela variável "apresentador".

No caso da dimensão 2 do CoTED – Discurso de convencimento ou persuasão –, ela não se assemelha com a dimensão 2 da língua inglesa – Discurso narrativo versus não narrativo. Temos nessa dimensão o uso extensivo de advérbios, que geralmente expressam informações sobre uma ação, acontecimento ou posicionamento pessoal, ou são utilizados para enfatizar ao informar determinados sentimentos. Desta forma, é possível perceber o quanto que os advérbios são utilizados no convencimento do destinatário sobre algo. Também temos a presença do uso de pronomes demonstrativos os quais representam as referências nominais, podendo ser relacionados a descrições de pessoas, coisas, lugares etc. ou servir de referência entre a proximidade do emissor e do destinatário (animado ou não). Com relação à quantidade de palavras, ela é maior nesta dimensão, o que pode nos levar a analisá-la como mais densa em palavras para se transmitir algo considerado importante. Deste modo, o nome dado à dimensão 2 do CoTED foi "**discurso de convencimento ou persuasão**". Mas é interessante notar que, nesse caso, o caráter mais argumentativo ou persuasivo das TED Talks foram mais salientes, em contrapartida com os resultados da AMD Aditiva, considerando a presença da dimensão 4

da língua inglesa no CoTED. Esse pode ser mais um fator para se pensar as TED Talks como um registro adaptável, sendo ora mais persuasivo ora menos persuasivo. Talvez, seja por isso que a ANOVA da dimensão 4 da língua inglesa sob as TED Talks (no geral) represente menos de 2,9% como percentual de predição de que a variação da linguagem verbal das TED Talks é explicada pela coocorrência das características linguísticas da dimensão 4 da língua inglesa. Em acréscimo, com a GLM da dimensão 2 do CoTED, – respondendo a terceira questão da pesquisa: como se dá a variação multidimensional funcional em termos das variáveis independentes "apresentador" e "evento"? – foi constatado que as variáveis externas exercem significativa influência na variação. Na dimensão 2 do CoTED, 38,4% da variação é influenciada pela variável "evento", e 32,3% da variação é influenciada pela variável "apresentador".

Por sua vez, a dimensão 3 do CoTED – Discurso assertivo e conjectural – também não se assemelha com a dimensão 3 da língua inglesa – Referência dependente de situação versus elaborada. Temos nessa dimensão o uso de adjetivos em posição predicativa, que costumam ser mais frequentes em registros do modo escrito (informativo) do que nos de modo falado (interativo). No geral, os adjetivos funcionam como marcadores de posicionamento do emissor (ou autor do texto) para expressar suas opiniões. Também temos o uso do *that* em orações complementares. Isso revela um discurso marcado pela informalidade e o não planejamento das falas. Contudo, esse aspecto seria um tanto improvável, considerando que os textos são previamente escritos, o que nos leva a considerar ser um discurso simulado. Tal simulação traz afirmações e conjecturas por parte do falante (ou autor do texto) sobre o assunto tratado. Considerando todo o exposto acima, o nome dado à dimensão 3 do CoTED foi "**discurso assertivo e conjectural**". Em acréscimo, com a GLM da dimensão 3 do CoTED, – respondendo a terceira questão da pesquisa: como se dá a variação multidimensional funcional em termos das variáveis independentes "apresentador" e "evento"? – foi constatado que as variáveis externas exercem significativa influência na variação. Na dimensão 3 do CoTED, 15,7% da variação é influenciada pela variável "evento" e 30,3% da variação é influenciada pela variável "apresentador".

Por fim, a dimensão 4 do CoTED – Discurso baseado em competências; última dimensão encontrada – não se assemelha com a dimensão 4 da língua inglesa – Argumentação explícita. Temos nessa dimensão o uso extensivo de substantivos, os quais carregam grande densidade de informações, sendo os principais portadores de significado referencial em um texto. A linguagem escrita, ou a linguagem formal e padronizada, é geralmente caracterizada

pelo uso de substantivos de forma a integrar ou elaborar informações. Nesta dimensão existe uma certa importância dada ao uso de substantivos concretos e/ou técnicos, trazendo o foco no conhecimento específico em determinadas áreas. Outra característica mais presente é o uso de verbos modais de possibilidade, permissão e/ou habilidade. Tais verbos podem sinalizar incerteza ou falta de precisão na apresentação de algum tipo de informação ou em declarações sobre habilidades ou sobre possibilidades com relação a certos prováveis eventos. Neste caso, a questão da habilidade está muito mais presente, o que corrobora com o foco no conhecimento concreto e/ou técnico. Considerando todo o exposto acima, o nome dado à dimensão 4 do CoTED foi "**discurso baseado em competências**" – características que condizem com a AMD Aditiva da dimensão 5.

Agora, de acordo com o que foi anteriormente apresentado – na Introdução deste trabalho –, existem estudos bastante válidos sobre as TED Talks, mas que geralmente são de menor porte – visando principalmente ao ensino da língua inglesa, à análise do discurso ou ao estudo da prosódia. Muitas dessas pesquisas trazem modelos pedagógicos na criação de atividades didáticas em língua inglesa, focando no ensino de vocabulário de áreas específicas e de estratégias que desenvolvam a percepção oral dos estudantes. Pesquisadores como Sabota e Almeida Filho (2017) consideraram as TED Talks como uma das ferramentas tecnológicas classificadas como potenciais mediadoras no processo de aprimoramento da competência teórica do professor de idiomas. Outros pesquisadores – como Rousseau *et al.* (2012) e Hasebe (2015) – buscaram disponibilizar corpora das TED Talks como uma ferramenta que fosse útil para o público em geral e alunos de língua inglesa como língua estrangeira. De acordo com Chang e Huang (2015), por sua vez, identificaram padrões entre os movimentos encontrados nas TED Talks que retratam traços de reprodução ou adaptação de outros gêneros como discursos de formatura, apresentações em conferências, discursos políticos e apresentações comerciais. Contudo, eles também afirmam que existe uma flexibilidade nessa ordem ou estrutura, cujos cenário e o propósito são o que diferenciam as TED Talks dos demais gêneros. Assim sendo, segundo eles, as TED Talks apresentam uma natureza heterogênea, podendo ser resultado de sua missão proposta de "informar, inspirar, surpreender e encantar" seus ouvintes (CHANG; HUANG, p. 50, 2015). Camiciottoli e Bonsignori (2015) alocaram as TED Talks em uma escala, juntamente com as palestras acadêmicas de disciplinas específicas, que se caracterizam como mais autênticos, monológicos, formais e científicos. Porém, assim como alguns gêneros apresentam diferentes colocações na escala, as TED Talks são também consideradas como um exemplo disso. Caliendo e Compagnone (2014) explicam que as

palestras TED possuem um caráter informativo, assim como as palestras de universidades, porém o que as diferencia das demais palestras é o fato de as palestras TED exercerem um papel de espaço pragmático alternativo, onde acadêmicos constroem sua imagem ao dar ênfase na sua afiliação a uma comunidade de especialistas e ao promover suas pesquisas, e seus resultados, como algo tangível e extremamente confiável. Outra pesquisa interessante é a da pesquisadora Miranda (2016), a qual classificou as TED Talks como um novo gênero discursivo ou um hipergênero.

Assim, não é de se surpreender que a linguagem verbal das TED Talks seja de grande interesse, afinal, ela é adaptável tanto para a linguagem falada quanto para a linguagem escrita, por exemplo – o que é de grande valia na área da educação, especialmente no ensino do inglês como língua estrangeira. O seu discurso abstrato ou baseado em competências, por sua vez, também pode ser um grande aliado na área da educação, especialmente quando o foco é em cursos que focam em áreas mais específicas. Também, a presença do apresentador com suas opiniões e conjecturas se fazem muito presentes. Mas a sua adaptabilidade pode ser o que tem instigado tanto os pesquisadores ao tentar definir e explicar as TED Talks, chegando a conclusões de que seja um novo gênero ou até um hipergênero. E, considerando os resultados obtidos nesta pesquisa, postula-se a possibilidade de classificar as TED Talks como um **registro híbrido**.

Contudo, para uma compreensão mais completa, sem dúvidas temos que considerar os possíveis fatores externos à linguagem verbal que a influenciam, como é o caso dos "eventos" e dos "apresentadores" – o que certamente demandaria futuras pesquisas. Outra questão a ser considerada é que os registros do corpus utilizado por Biber (1988) datam da década de 1960 e, apesar de sua pesquisa descrever a língua inglesa de modo geral, não se sabe até que ponto os mesmos textos usados apresentariam resultados semelhantes ou não, se sua pesquisa fosse realizada atualmente. Talvez seja por isso que encontramos algumas divergências entre os resultados da AMD Aditiva em comparação com a AMD Completa. Ainda assim, considerando o grande número de estudos que já foram realizados com essa abordagem, podemos dizer que as dimensões ainda são, de fato, construtos válidos (BERBER SARDINHA, 2014).

Por fim, conforme previamente dito, como resultado desta pesquisa, espera-se contribuir para que possamos conhecer de modo detalhado o funcionamento da linguagem verbal dessa influente modalidade de comunicação contemporânea, e contribuir de forma direta ou indireta para futuras pesquisas e possíveis aplicações.

6. Referências

ANDERSON, C. **O guia oficial do TED para falar em público**. Tradução de Donaldson Garschagen e Renata Guerra. Rio de Janeiro: Editora Intrínseca, 2016.

ANDERSON, C. **TED talks: the official TED guide to public speaking**. Boston: Houghton Mifflin Harcourt, 2016.

ALY, B. **The history of American public address as a research field**. In: Quarterly Journal of Speech, volume 29, 3° edição, 1943.
Disponível em: https://www.tandfonline.com/doi/abs/10.1080/00335634309380896.

ARAÚJO, R. F. **A linguagem dos reality TV shows norte-americanos: análise e classificação**. 2017. 242 f. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) - Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem, Pontifícia Universidade Católica de São Paulo, São Paulo, 2017.

BASKERVILLE, B. **The People's Voice: The Orator in American Society**. University of Kentucky, 1979.

BERBER SARDINHA, T.; VEIRANO PINTO, M. **A linguistic typology of American television**. In: International Journal of Corpus Linguistics, Volume 26, edição 1, 2021, p. 127-160.

BERBER SARDINHA, T.; PINTO, VEIRANO PINTO, M. **Multi-dimensional analysis: research methods and current issues**. London: Bloomsbury Academic, 2019.

BERBER SARDINHA, T. **A corpus-based history of Applied Linguistics**. Paper presented at the 18th World Congress of Applied Linguistics (AILA), Rio de Janeiro, RJ, Brazil, 2017.

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

BERBER SARDINHA, T.; KAUFFMANN, C.; ACUNZO, C. **Dimensions of register variation in Brazilian Portuguese**. In BERBER SARDINHA, T.; VEIRANO PINTO, M. (Eds.), Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber. Amsterdam/Philadelphia, PA: John Benjamins, 2014.

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

BERBER SARDINHA, T. **25 years later: Comparing Internet and pre-Internet registers.** In: T. Berber Sardinha; M. Veirano Pinto (Orgs.); Multi-Dimensional Analysis, 25 years on. A tribute to Douglas Biber, Studies in Corpus Linguistics. v. 60, Amsterdam; Philadelphia: John Benjamins, p.81–105, 2014b.

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

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

BÉRTOLI-DUTRA, P. **Multi-dimensional analysis of pop songs**. In BERBER SARDINHA, T.; VEIRANO PINTO, M. (Eds.). MultiDimensional Analysis, 25 years A tribute to Douglas Biber. Amsterdam and Philadelphia: John Benjamins Company, 2014.

BÉRTOLI-DUTRA, P. **Linguagem da música popular anglo-americana de 1940 a 2009**. 2010. 290f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo, São Paulo, 2010.

BESNIER, N. **The linguistic relationships of spoken and written Nukulaelae registers**. Language, v. 64, p. 707-736, 1988.

BIBER, D. **A model of textual relations within the written and spoken modes**. University of S. California. (PHD Thesis). LING WRIL, 1984.

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

BIBER, D. **Representativeness in Corpus Design**. Literary and Linguistic computing, v. 8, p. 243–257, 1993.

BIBER, D.; HARED, M. **Linguistic correlates of the transition to literacy in Somali: Language adaptation in six press registers**. In: BIBER, D.; FINEGAN, E. (Org.). Sociolinguistic perspectives on register. Oxford: Oxford University Press, 1994. p. 182-216.

BIBER, D. **Dimensions of Register Variation: A Cross-Linguistic Comparison**. Cambridge: Cambridge University Press, 1995.

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

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

BIBER, D. et al. **Longman Student Grammar of Spoken and Written English**. Harlow, Essex, England: Pearson Education Limited, 2002.

BIBER, D. et al. **Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus**. In: Monograph Series, Princeton/ Nova Jersey: Educational Testing Service, 2004. (Fonte: https://origin-www.ets.org/Media/Research/pdf/RM-04-03.pdf).

BIBER, D. **Conversation text types: A multi-dimensional analysis**. JDT 2004: 7 Journées internationales d'Analyses statistique des Donmées Textualles, 2004.

BIBER, D.; DAVIES, M.; JONES, J. K.; TRACY-VENTURA, N. **Spoken and written register variation in Spanish: A multi-dimensional analysis**. Corpora, v. 1, 2006, p. 1-37.

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

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

BIBER, D. **Multidimensional approaches**. In: Corpus linguistics: An international handbook, Anke Lüdeling e Merja Kytö (eds.). Berlin: Walter de Gruyter, 2009, p. 822-855.

BIBER, D. **Multi-Dimensional Analysis: A Historical Synopsis**. In: BERBER SARDINHA, T. PINTO, Márcia Veirano. **Multi-dimensional analysis: research methods and current issues**. London: Bloomsbury Academic, 2019, p. 11-26.

BÍBLIA. Português. **Bíblia On-line**. Disponível em: https://bibliaportugues.com. (Acesso em: 23 set. 2020).

BONINI, Adair. **Os gêneros do jornal: o que aponta a literatura da área de comunicação no Brasil?** In: Linguagem em (Dis)curso, v.4, n.1, p.205-231, 2003.

BROOKES, G.; MCENERY, T. **The Routledge handbook of English language and digital humanities**. Nova York: Routledge, 2020.

CALIENDO, G.; COMPAGNONE, A. **Expressing epistemic stance in university lectures and TED Talks**. Lingue e Linguaggi. 2014, Vol. 11, p105-122. 18 p.

CAMICIOTTOLI, B. C.; BONSIGNORI, V. **The Pisa audiovisual corpus project: a multimodal approach to ESP research and teaching**. ESP Today, 2015. Disponível em: http://www.esptodayjournal.org/pdf/current_issue/8.12.2015/BELINDA&VERONICA-full-text.pdf.

CANTOS-GOMEZ, P. **Multivariate Statistics Commonly Used in Multi-Dimensional Analysis**. In: Multi-dimensional analysis: research methods and current issues. London: Bloomsbury Academic, 2019.

CELANI, M. A. A. **Transdisciplinaridade na Linguística Aplicada no Brasil**. In: Inês Signorini e Marilda C. Cavalcanti. (Org.). Linguística Aplicada e Transdisciplinaridade: questões e perspectivas. Campinas: Mercado de Letras, 1998, p. 115-127.

CHAGAS, P. **A mudança linguística.** In: FIORIN, J. L. (org.). Introdução à lingüística I. Objetos teóricos. São Paulo: Contexto, 2002, p. 141-163.

CHANG, Y; HUANG, H. **Exploring TED Talks as a pedagogical resource for oral presentations**. English Teaching & Learning 39.4 (Special Issue 2015): 29-62.

CHOMSKY, Noam. **Syntactic Structures**. The Hague/Paris: Mouton, 1957.

CONRAD, S. **Academic discourse in two disciplines: professional writing and student development in biology and history.** Dissertation (Doctor in Philosophy) – Department of English, Northern Arizona University, Flagstaff, 1996.

CORREIA, R. P. S. **Automatic Classification of Metadiscourse**. Tese de doutorado, Universidade Técnica de Liboa e Universidade Carnegie Mellon, 2018.

CROSSLEY, S.; LOUWERSE, M. M. Multi-dimensional register classification using bigrams. **International Journal of Corpus Linguistics**, 2007. Disponível em: http://dx.doi.org/10.1075/ijcl.12.4.02cro.

CYRANKA, L. F. M. **Evolução dos Estudos Linguísticos**. In: Revista Práticas de Linguagem. v. 4, n. 2. UFJF: jul./dez. 2014. Disponível em: https://www.ufjf.br/praticasdelinguagem/edicoes-2/edicoes/volume-4-n-2-juldez-2014/ e https://www.ufjf.br/praticasdelinguagem/files/2014/09/160-198-Evolu%c3%a7%c3%a3o-dos-estudos-lingu%c3%adsticos.pdf.

DE MÖNNINK, I.; BROM, N.; OOSTDIJK, N. Using the MF/MD method for automatic text classification. **Extending the scope of corpus-based research**. p.13–25, 2003. Amsterdam: Brill Rodopi.

DELFINO, M. C. N. **Uso de música para o ensino de inglês como língua estrangeira em um ambiente baseado em corpus**. 2016. 159 f. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) - Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem, Pontifícia Universidade Católica de São Paulo, São Paulo, 2016.

EASTMAN, C. **Oratory and Platform Culture in Britain and North America, 1740–1900**. In: Oxford Handbooks Online (www.oxfordhandbooks.com). Oxford: Oxford University Press, 2018.

EGBERT, J. **Corpus Design and Representativeness**. In: Multi-dimensional analysis: research methods and current issues. London: Bloomsbury Academic, 2019.

FARRELL, J. M. **"Above all Greek, above all Roman fame": Classical Rhetoric in America during the Colonial and Early National Periods**. International Journal of the Classical Tradition 18:3, 2011, 415-436.

FRANCO SILVA, L..; PINTO, P. T.; NAZZI-LARANJA, L. A ted talk corpus for teaching lexical bundles for Brazilians EAP students. TaLC2020 Abstracts: **John Benjamins Publishing Company, 2020.** Disponível em: https://f-origin.hypotheses.org/wp-content/blogs.dir/6396/files/2020/07/Abstracts130720.pdf.

FREIRE, P. **Pedagogia da Indignação: cartas pedagógicas e outros escritos**. São Paulo, UNESP, 2000.

GRABE, W. **Applied Linguistics: A Twenty-First-Century Discipline**. In: The Oxford Handbook of Applied Linguistics (2 ed.). Oxford University Press: 2012.

HALLIDAY, M. A. K. **An Introduction to Functional Grammar**. 1° edição. London: Edward Arnold, 1985.

HALLIDAY, M. A. K. **An Introduction to Functional Grammar**. 4° edição. Londres/Nova Yorke: Routledge, 2014.

HASEBE, Y**. Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks**. In: Procedia: Social and Behavioral Sciences, v. 198: 2015, p. 174–182. Disponível em: https://www.sciencedirect.com/journal/procedia-social-and-behavioral-sciences/vol/198/suppl/C.

KAUFFMANN, C. H. **Linguística de corpus e estilo: análises multidimensional e canônica na ficção de Machado de Assis**. 2020. 277 f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) - Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem, Pontifícia Universidade Católica de São Paulo, São Paulo, 2020.

KIM, Y.-J.; BIBER, D. **A corpus-based analysis of register variation in Korean**. In: BIBER, D.; FINEGAN, E. (Org.). Sociolinguistic perspectives on register. Oxford: Oxford University Press, 1994. p. 157-181.

KUCERA, H.; FRANCIS, W. N. **Computational analysis of present-day American English**. Providence: Brown University Press, 1967.

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

LEE, D. Y. W. **Modelling Variation in Spoken and Written Language: the Multi-Dimensional Approach Revisited**, 1999. Doutorado em Linguística e Língua Inglesa, Lancaster University.

MATOS(a), M. S. P. B. **A curiosidade e o conhecimento: a praxiologia em debate**. In: Educon, Aracaju, Volume 08, n. 01, p.1-8, 2014. Disponível em: http://anais.educonse.com.br/2014/a_curiosidade_e_o_conhecimento_a_praxiologia_em_debate.pdf.

MATOS(b), D. A. S.; RODRIGUES, E. C. **Análise fatorial**. Brasília: Enap, 2019.

MAYER, C. **O que e como escrevemos na web: um estudo multidimensional de variação**

de registro em língua inglesa. 2018. 129 f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) - Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem, Pontifícia Universidade Católica de São Paulo, São Paulo, 2018.

MEGID, S. B. C. Núcleo básico: linguagem, trabalho e tecnologia. Fernanda Mello Demai (rev.); André Müller de Mello (coord.) – São Paulo: Fundação Padre Anchieta, 2011.

MIRANDA, A. A. P. B. **Palestras TED: um novo gênero do discurso?** 2016. 157 f. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) - Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem, Pontifícia Universidade Católica de São Paulo, São Paulo, 2016.

MIYAMOTO, M. R. T. **O Impacto do uso das Novas Tecnologias em Aulas de Inglês para cursos Tecnológicos**. Revista CBTecLE, v. 1, n. 1, 2017.

MOITA LOPES, L. P. **Linguística aplicada e vida contemporânea: problematização dos construtos que tem orientado a pesquisa.** In: MOITA LOPES, L. P. (Org.) Por uma linguística aplicada indisciplinar. São Paulo: Parábola Editorial, 2006, p. 85-107.

NUTTALL, P. A. **A Classical and Archaeological Dictionary of the Manners, Customs, Laws, Institutions, Arts, Etc. of the Celebrated Nations of Antiquity, and of the Middle Ages: To which is Prefixed A Synoptical and Chronological View of Ancient History**. Whittaker and Company, 1840, p. 358. (Disponível em: <https://archive.org/details/classicalarchaeo00nuttuoft>. (Acesso em: 23 set. 2020). Citado em: https://en.wikipedia.org/wiki/Palaestra.

O'KEEFFE, A.; MCCARTHY, M. **The Routledge Handbook of Corpus Linguistics**. Abingdon: Routledge, 1 ed., 2010.

PARODI, G. **Variation across registers in Spanish: Exploring the El-Grial PUCV Corpus**. In: PARODI, G. (Ed.). **Working with Spanish Corpora**. London: Continuum, 2007.

PEROZA, J.; RESENDE, M. A. **A dialética da curiosidade: pressupostos para uma praxiologia do conhecimento em Paulo Freire**. In: X Congresso Nacional de Educação – EDUCERE. Paraná: PUCPR, 2011. Disponível em: https://educere.bruc.com.br/CD2011/pdf/4348_2600.pdf e https://periodicos.uninove.br/eccos/article/view/3157/2154.

PETTER, M. **Linguagem, língua, linguística.** In: FIORIN, J. L. (org.). Introdução à lingüística I. Objetos teóricos. São Paulo: Contexto, 2002, p. 11-24.

PINHEIRO, E. G. **Combinação Linear de Classificadores para Análise de Risco na Evasão da UFC**. Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Estatística, Fortaleza, 2018.

RAINE, P. **Talk Corpus: A Web-based Corpus of TED Talks for English Language Teachers and Learners**. In: Accents Asia, 12(1), 2019, p. 1-38.

RATANAKUL, S. **A Study of Problem-Solution Discourse: Examining TED Talks through the Lens of Move Analysis**. In: LEARN Journal: Language Education and Acquisition Research Network, Volume 10, número 2, 2017, p. 25-46.

RESENDE, S V. **Dimensões de variação do texto traduzido: uma abordagem multidimensional.** 2019. 295 f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) - Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem, Pontifícia Universidade Católica de São Paulo, São Paulo, 2019.

ROUSSEAU, A.; DELÉGLISE, P.; ESTÈVE, Y. **TED-LIUM: An automatic speech recognition dedicated corpus**. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation. (LREC'12), Turquia, Maio de 2012.

SABOTA, B.; ALMEIDA FILHO, J. C. P. **Análise do potencial da mediação tecnológica para o enriquecimento da competência teórica de professores de línguas**. Acta Scientiarum. Language and Culture, vol. 39, no. 4, 2017.

SÁTIRO, T. F.; SILVA, A. X. S. **A sabedoria das TED Talks aplicada a aulas de física**. Encontros Universitários da UFC, Fortaleza, v. 3, n. 1, 2018.

SILVA, L. F; PINTO, P. T.; DIAS, E. **Atividades de compreensão oral com base em corpora de TED Talks: um estudo piloto**. In: Linguística de corpus: perspectivas. Organizadoras: Maria José Bocorny Finatto, Rozane Rodrigues Rebechi, Simone Sarmento, Ana Eliza Pereira Bocorny. — Porto Alegre: Instituto de Letras, UFRGS, 2018.

SILVA(c), A. F**. A compreensão oral de vocabulário técnico em contexto de ESP**. Revista CBTecLE, v. 1, n. 2, 2017.
Disponível em: https://revista.cbtecle.com.br/index.php/CBTecLE/article/view/78.

SINCLAIR, J. M. **Beginning the study of lexis**. ln: BAZELL, C. E. (Org.). In Memory of J. R. Firth. Londres: Longman, 1966, p. 410-30.

SINCLAIR, J. M. **Corpus evidence in language description**. In: WICHMANN et al. Teaching and language corpora. Londres/Nova York: Longman, 1997. p. 27 -39.

SWALES, J. M. **Genre Analysis: English in academic and research settings**. Cambridge: Cambridge University Press, 1990.

TANVEER, M.; HASSAN, M.; GILDEA, D.; HOQUE, E. (2019). **Predicting TED Talk Ratings from Language and Prosody**. In: arXiv:1906.03940. University of Rochester, 2019. Disponível em: https://arxiv.org/pdf/1906.03940.pdf.

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

TOLEDO, Y. M. **A linguagem verbal das artes visuais: uma análise multidimensional do discurso sobre a fotografia de Sally Mann**. 2020. 182 f. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) - Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem, Pontifícia Universidade Católica de São Paulo, São Paulo, 2020.

TSAI, T. J. **Are You TED Talk Material? Comparing Prosody in Professors and TED Speakers**. In: INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association. Dresden, 2015, p. 2534-2538.

TSOU, A; THELWALL, M; MONGEON, P; SUGIMOTO, C R. **A Community of Curious Souls: An Analysis of Commenting Behavior on TED Talks Videos**. In: PLoS ONE, volume 9, número 4, 2014.

VEIRANO PINTO, M. **A linguagem dos filmes norte-americanos ao longo dos anos: uma abordagem multidimensional**. 2013. 488f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo, São Paulo, 2013.

VILELA, P. C. S. **Classificação de Sentimento para Notícias sobre a Petrobras no Mercado Financeiro**. Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2011.

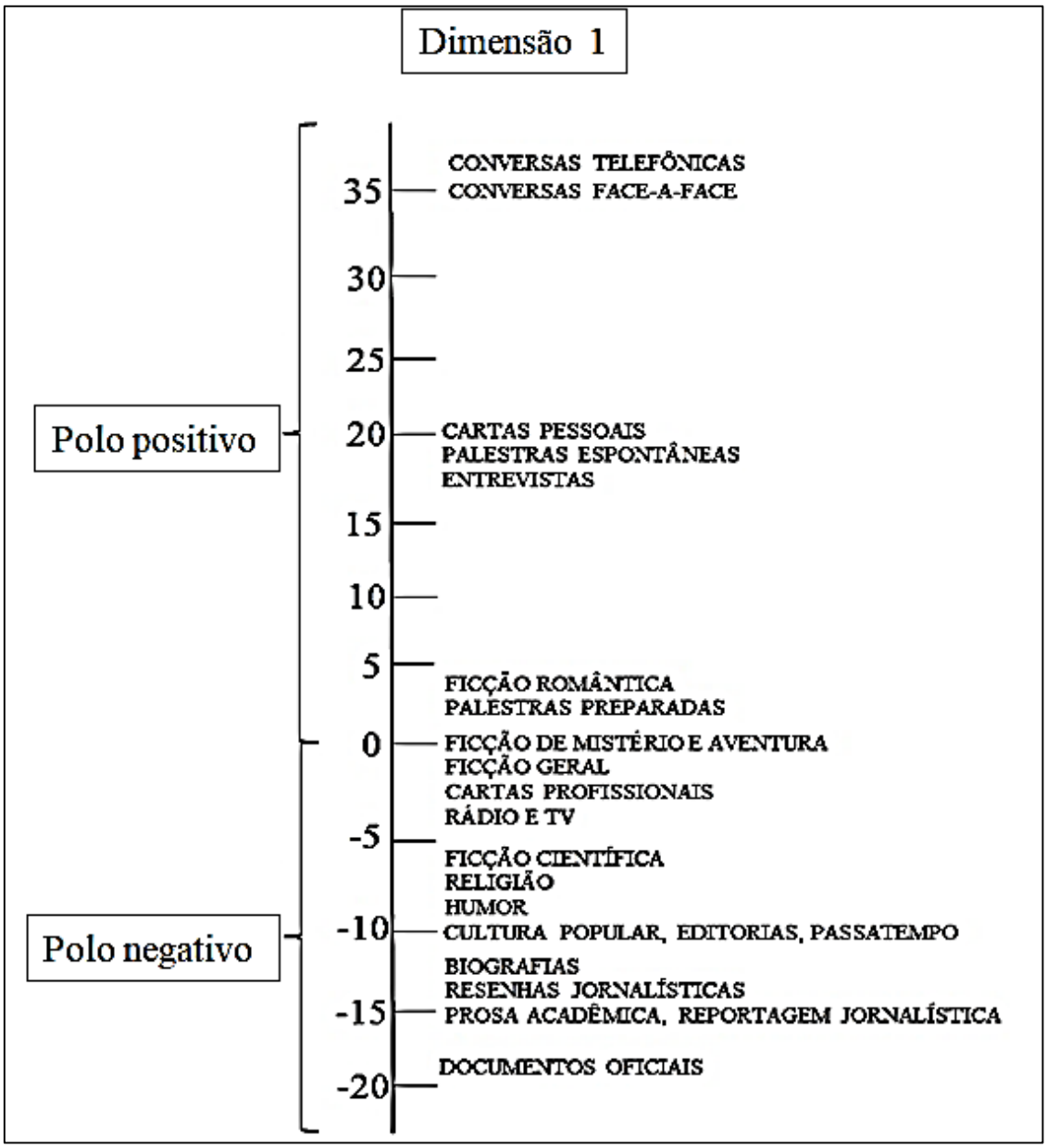
WESTIN, I.; GEISLER, C. **A multi-dimensional study of diachronic variation in British newspaper editorials**. ICAME, v. 26, 2002, p. 133-152.

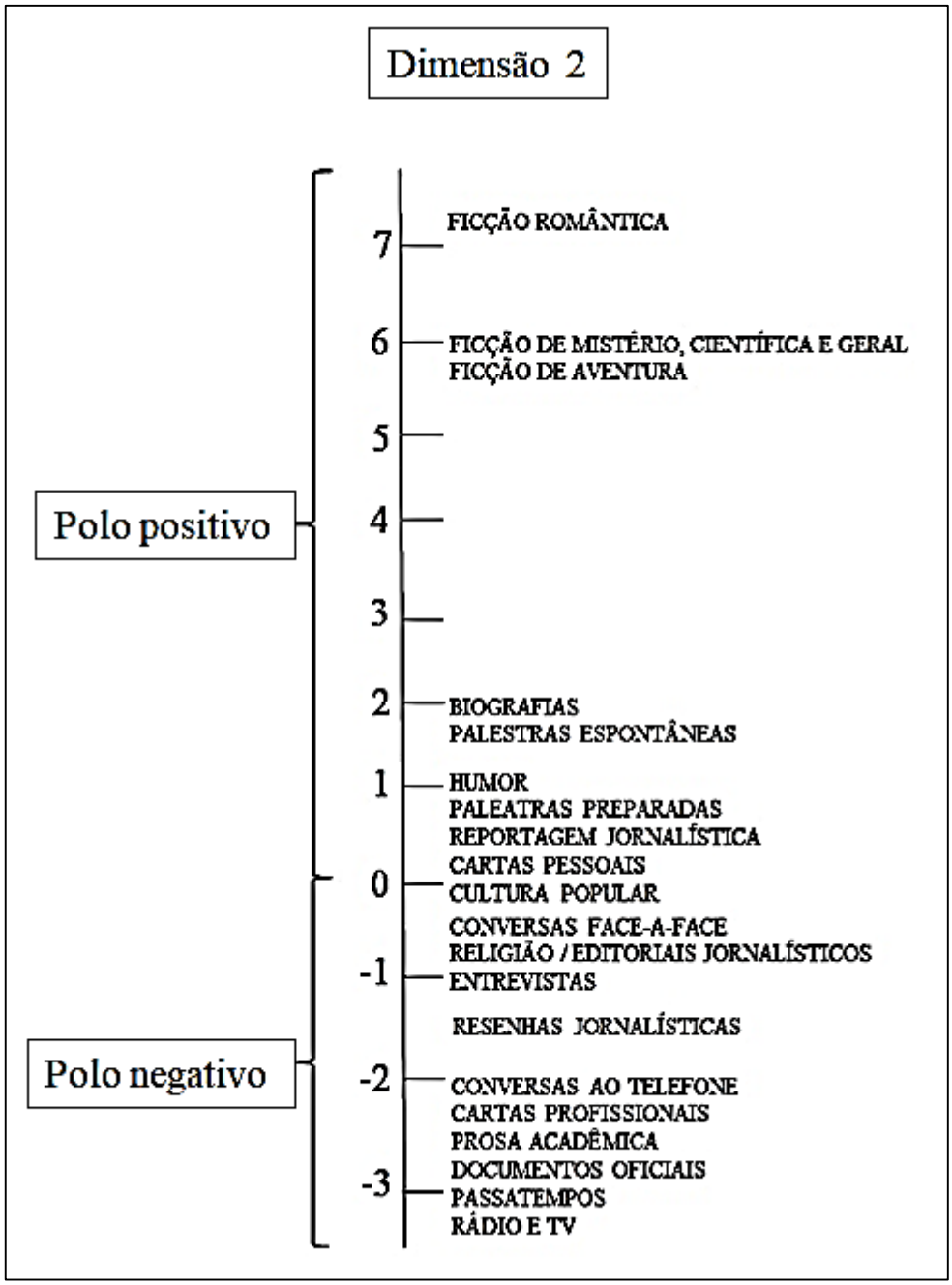
WHITE, M. **Language in job interviews: differences relating to success and socioeconomic variables.** Dissertation (Doctor in Philosophy) – Department of English. Northern Arizona University, Flagstaff, 1994.

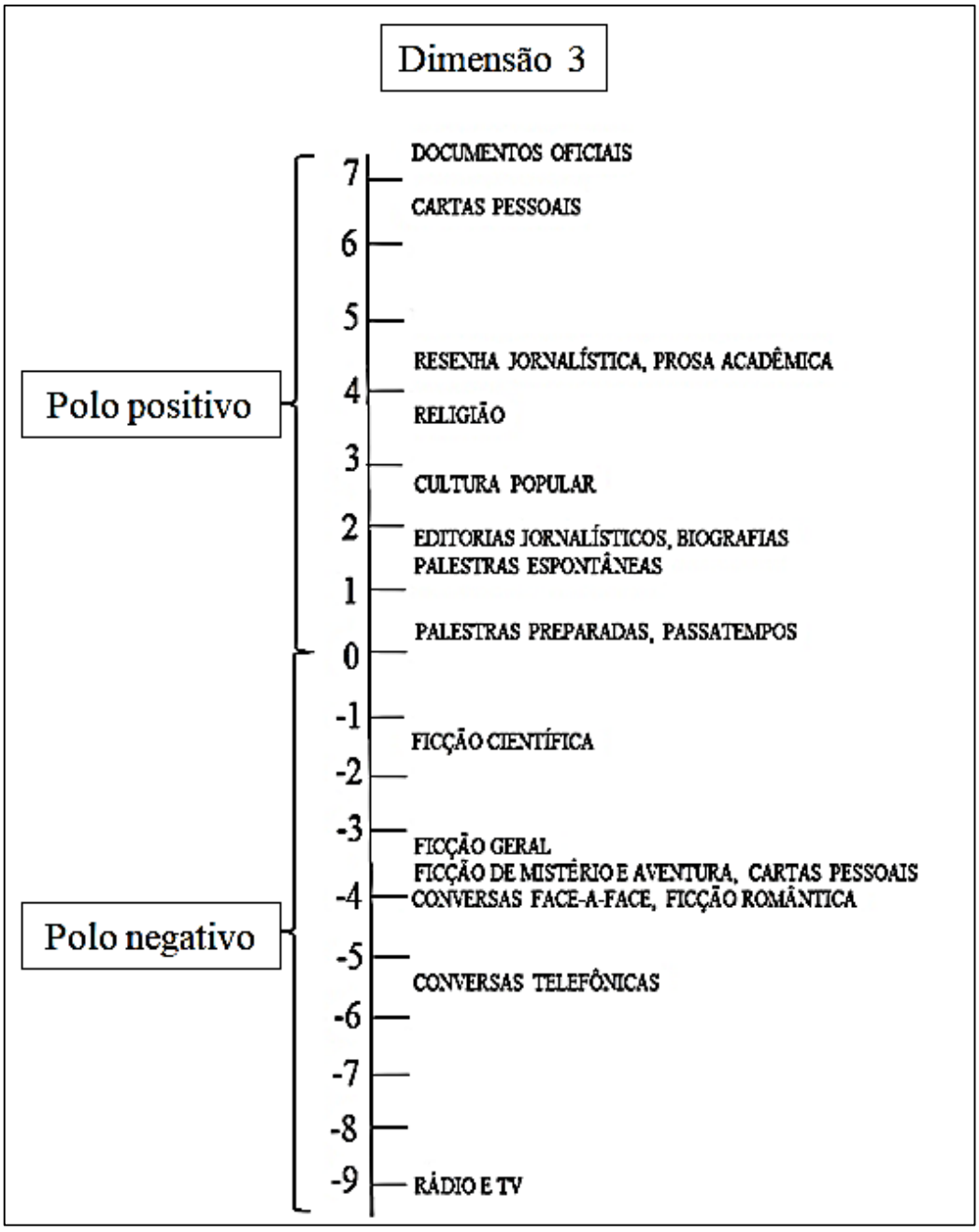
ZUPPARDO, M. C. **Dimensões de variação em manuais aeronáuticos: um estudo baseado na análise multidimensional.** 2014. 154 f. Dissertação (Mestrado em Linguística) - Pontifícia Universidade Católica de São Paulo, São Paulo, 2014.

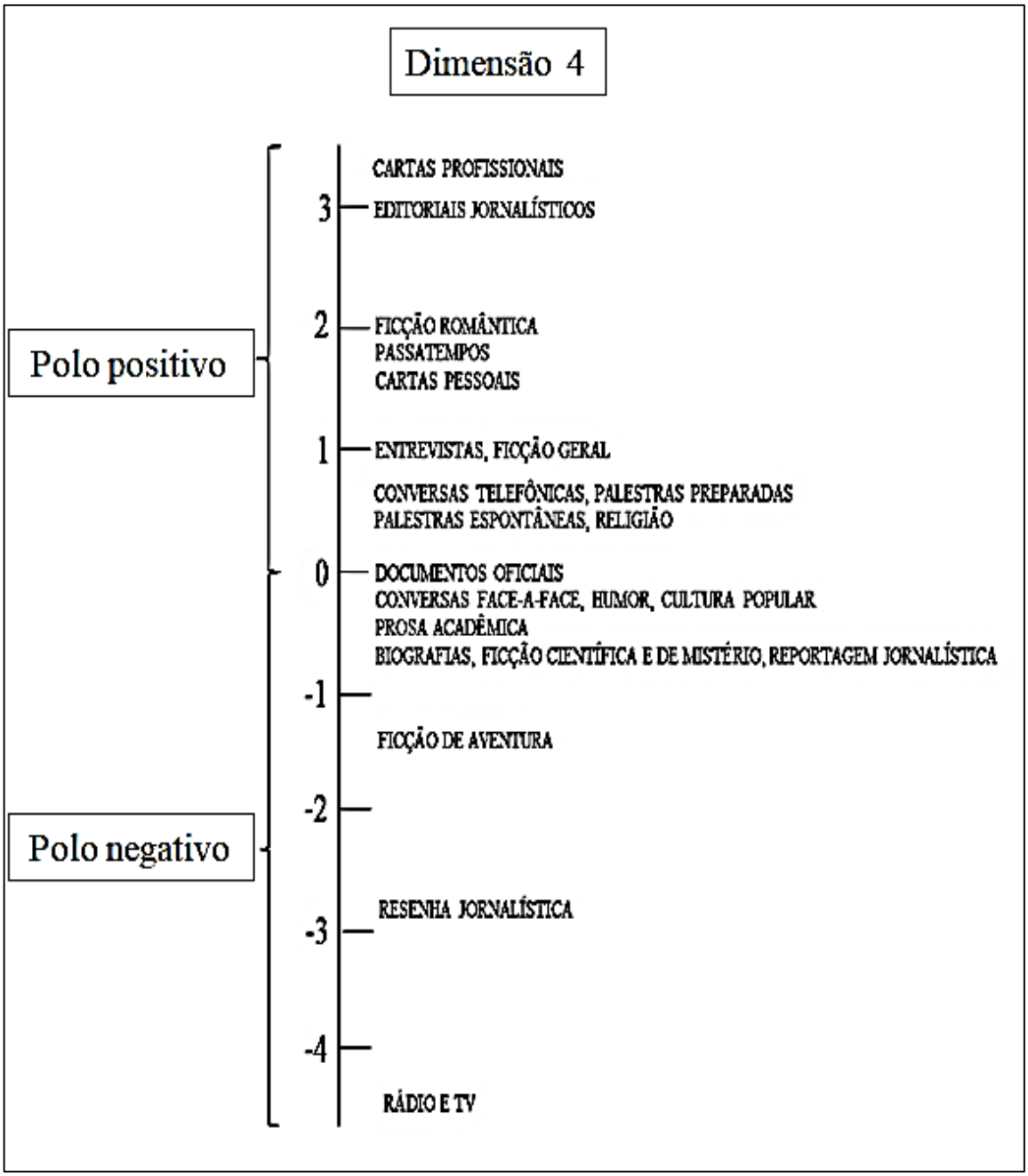
Anexos:

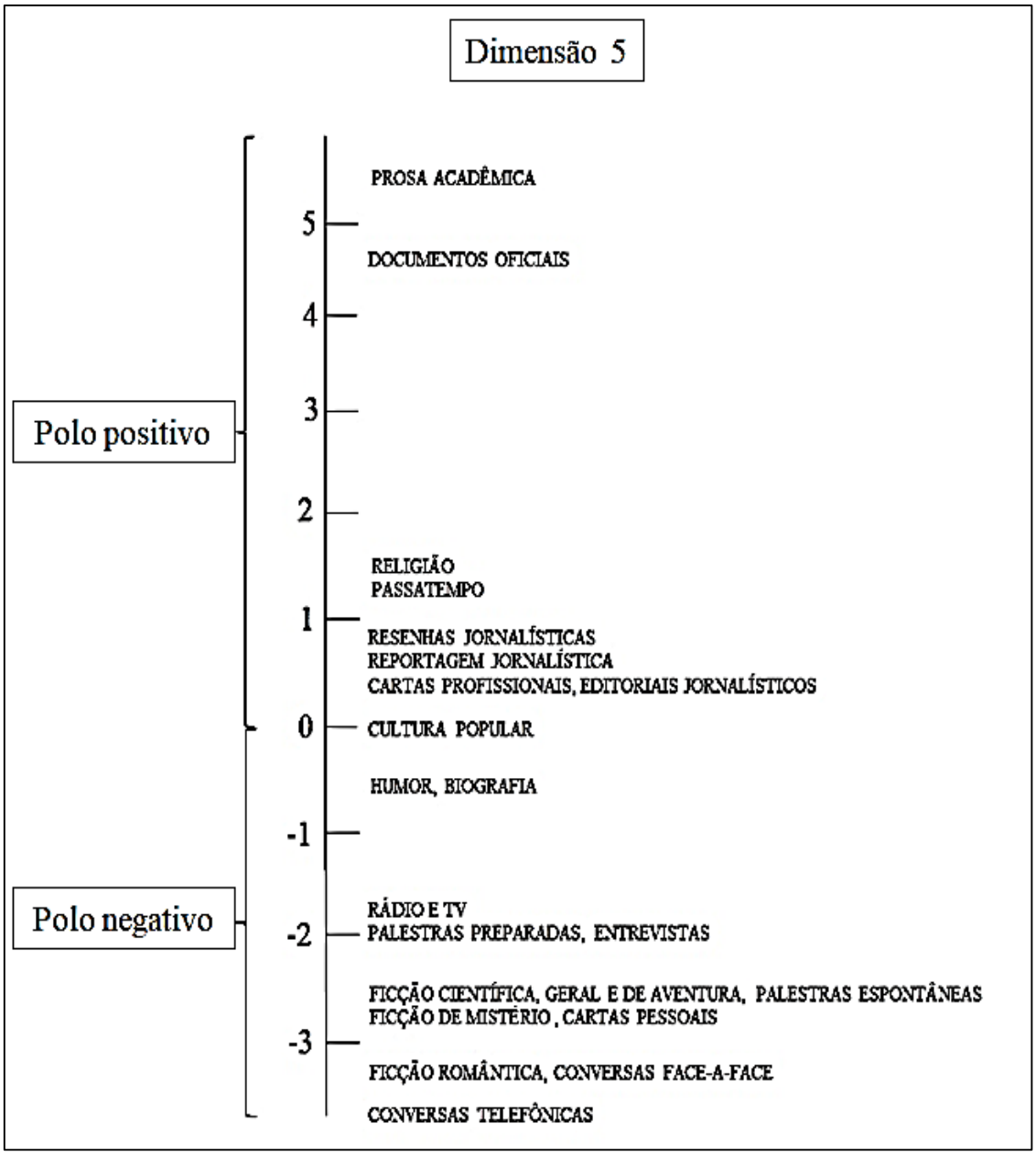
Anexo 1




Dimensão 2
Polo positivo
Polo negativo
7
FICÇÃO ROMÂNTICA
6
FICÇÃO DE MISTÉRIO, CIENTÍFICA E GERAL
FICÇÃO DE AVENTURA
5
4
3
2
BIOGRAFIAS
PALESTRAS ESPONTÂNEAS
1
HUMOR
PALEATRAS PREPARADAS
REPORTAGEM JORNALÍSTICA
CARTAS PESSOAIS
0
CULTURA POPULAR
CONVERSAS FACE-A-FACE
RELIGIÃO / EDITORIAIS JORNALÍSTICOS
-1
ENTREVISTAS
RESENHAS JORNALÍSTICAS
-2
CONVERSAS AO TELEFONE
CARTAS PROFISSIONAIS
PROSA ACADÊMICA
DOCUMENTOS OFICIAIS
-3
PASSATEMPOS
RÁDIO E TV


Dimensão 3
Polo positivo
Polo negativo
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
DOCUMENTOS OFICIAIS
CARTAS PESSOAIS
RESENHA JORNALÍSTICA, PROSA ACADÊMICA
RELIGIÃO
CULTURA POPULAR
EDITORIAS JORNALÍSTICOS, BIOGRAFIAS
PALESTRAS ESPONTÂNEAS
PALESTRAS PREPARADAS, PASSATEMPOS
FICÇÃO CIENTÍFICA
FICÇÃO GERAL
FICÇÃO DE MISTÉRIO E AVENTURA, CARTAS PESSOAIS
CONVERSAS FACE-A-FACE, FICÇÃO ROMÂNTICA
CONVERSAS TELEFÔNICAS
RÁDIO E TV


Dimensão 4
Polo positivo
Polo negativo
3
2
1
0
-1
-2
-3
-4
CARTAS PROFISSIONAIS
EDITORIAIS JORNALÍSTICOS
FICÇÃO ROMÂNTICA
PASSATEMPOS
CARTAS PESSOAIS
ENTREVISTAS, FICÇÃO GERAL
CONVERSAS TELEFÔNICAS, PALESTRAS PREPARADAS
PALESTRAS ESPONTÂNEAS, RELIGIÃO
DOCUMENTOS OFICIAIS
CONVERSAS FACE-A-FACE, HUMOR, CULTURA POPULAR
PROSA ACADÊMICA
BIOGRAFIAS, FICÇÃO CIENTÍFICA E DE MISTÉRIO, REPORTAGEM JORNALÍSTICA
FICÇÃO DE AVENTURA
RESENHA JORNALÍSTICA
RÁDIO E TV


Dimensão 5
Polo positivo
Polo negativo
5
4
3
2
1
0
-1
-2
-3
PROSA ACADÊMICA
DOCUMENTOS OFICIAIS
RELIGIÃO
PASSATEMPO
RESENHAS JORNALÍSTICAS
REPORTAGEM JORNALÍSTICA
CARTAS PROFISSIONAIS, EDITORIAIS JORNALÍSTICOS
CULTURA POPULAR
HUMOR, BIOGRAFIA
RÁDIO E TV
PALESTRAS PREPARADAS, ENTREVISTAS
FICÇÃO CIENTÍFICA, GERAL E DE AVENTURA, PALESTRAS ESPONTÂNEAS
FICÇÃO DE MISTÉRIO, CARTAS PESSOAIS
FICÇÃO ROMÂNTICA, CONVERSAS FACE-A-FACE
CONVERSAS TELEFÔNICAS

Anexo 2

Descrições das etiquetas do Biber Tagger		
1	:+clp+++	colon + clause punctuation
2	;+clp+++	semi-colon + clause punctuation
3	?+clp+++	question mark + clause punctuation
4	!+clp+++	exclamation mark + clause punctuation
5	,++++	comma
6	-++++	dash
7	++++ double quote mark 8 '++++ single quote mark 9 (++++ left parenthesis 10)++++ right parenthesis 11 $++++ dollar sign 12 %++++ percent sign 13 &fo++++ formula symbols 14 &fw++++ foreign word 15 abl++++ pre-	multi-word coordinating conjunction (as well as)

	qualifier (rather, such) 16 abn++++ pre-quantifier (all, half) 17 abx++++ pre-quantifier/double conjunction (both) 18 ap++++ post-determiner (many, more, most, only, other, own, same, ...) 19 aps++++ (others) 20 at++++ singular indefinite article (a, an) 21 ati++++ singular definite article (the, no) 22 cc++++ coordinating conjunction (and, but, or) 23 cc+phrs+++ coordinating conjunction + phrasal connector 24 cc++++	

25	cc++neg++	coordinating conjunction + + negation (nor)
26	cd++++	cardinal number (2, 3, 4, two, three, four, hundred, ...)
27	cd+date+++	cardinal number + date (year only)
28	cdl++++	cardinal number: 1, one
29	cd1s++++	cardinal number: ones
30	cds++++	cardinal plural (tens, hundreds, thousands)
31	od++++	ordinal number (1st, 2nd, first, second, ...)
32	cs+cnd+++	subordinating conjunction + conditional (if, unless)
33	cs+con+++	subordinating conjunction + concessive (although, though)
34	cs+cos+++	subordinating conjunction + causative (because)
35	cs+who+++	subordinating conjunction + WH word (whether)
36	cs+sub+++	subordinating conjunction + other (as, except, until, ...)
37	cs"++++	multi-word subordinating conjunction (in that, so that, ...)
38	dt+dem+++	determiner + demonstrative (this, that, these, those modifying N)
39	dt+pdem+++	determiner + demonstrative pronoun (this, that, these, those)
40	dti++++	singular or plural determiner (any, enough, some)
41	dt++++	other singular determiner (another, each)
42	dtx++++	determiner/double conjunction (either)
43	ex+pex+++	existential there
44	in++++	preposition
45	in+ppvb+++	preposition + prepositional verb (account for, join in, ...)
46	in+pl+++	preposition + place marker (above, behind, beside, ...)
47	in"++++	multi-word preposition (as to, away from, instead of, ...)
48	in+strn+++	preposition + stranded
49	jj+atrb+++	adjective + attributive function
50	jj+atrb++xvbg+	adjective + attributive function + + -ing form
51	jj+atrb++xvbn+	adjective + attributive function + + past participle form
52	jj+pred+++	adjective + predicative function
53	jj++++	adjective + indeterminate function

54	jjb+atrb+++	attributive-only adjective + attributive (chief, entire)
55	jjr+atrb+++	comparative adjective + attributive function
56	jjr+pred+++	comparative adjective + predicative function
57	jjt+atrb+++	superlative adjective + attributive function
58	md+nec+++	modal + necessity (ought, should, must)
59	md+pos+++	modal + possibility (can, may, might, could)
60	md+prd+++	modal + prediction (will, would, shall)
61	md"++pmd"++	modal + + multi-word periphrastic modal (e.g., be going to)
62	nn++++	singular common noun
63	nn+nom+++	singular noun + nominalization
64	nvbg+++xvbg+	singular noun + + + -ing form
65	nns++++	plural common noun
66	nns+nom+++	plural noun + nominalization
67	nnu++++	unit of measurement (lb, kg, ...)
68	np++++	singular proper noun
69	nps++++	plural proper noun
70	npl++++	locative noun
71	npt++++	singular titular noun
72	npts++++	plural titular noun
73	nr++++	singular adverbial noun (east, west, today, home, ...)
74	nrs++++	plural adverbial noun
75	pp1a+pp1+++	first person subject pronoun + first person pronoun
76	pp1a+pp1+++0	first person subject pronoun + 1st person pro. + contracted
77	pp1o+pp1+++	first person object pronoun + first person pronoun
78	pp$+pp1+++	possessive determiner + first person pronoun (my, our)
79	ppl+pp1+++	singular reflexive pronoun + first person pronoun (myself)
80	ppls+pp1+++	plural reflexive pronoun + first person pronoun (ourselves)
81	pp2+pp2+++	second person pronoun + second person pronoun

82	pp$+pp2+++	possessive determiner + second person pronoun (your)
83	ppl+pp2+++	singular reflexive pronoun + second person pronoun (yourself)
84	pp3a+pp3+++	third person subject pronoun + third person personal pronoun
85	pp3o+pp3+++	third person object pronoun + third person personal pronoun
86	pp3+pp3+++0	third person pronoun + 3rd person personal pro. + contracted
87	pp$+pp3+++	possessive + 3rd pers. personal pro. (his, her, their)
88	ppl+pp3+++	sg. reflexive pronoun + 3rd pers. personal pro. (her/himself)
89	ppls+pp3+++	pl. reflexive pronoun + 3rd pers. personal pro. (themselves)
90	pp3+it+++	third person pronoun + third person impersonal pronoun (it)
91	pp$+it+++	possessive determiner + third person impersonal pronoun (its)
92	pp$$++++	possessive pronoun (mine, yours, ...)
93	pn"++++	multi-word nominal pronoun (no one, ...)
94	pn++++	nominal pronoun (someone, everything, ...)
95	ql++++	qualifier + (as, less, more, too)
96	ql+amp+++	qualifier + amplifier (very)
97	ql+emph+++	qualifier + emphatic (most)
98	qlp++++	post-qualifier (enough, indeed)
99	rb++++	general adverb
100	rb"++++	multi-word adverb (at last, in general)
101	rb+cnj	adverb + conjunct (however, therefore, thus, ...)
102	rb++neg++	neither
103	rb+amp+++	adverb + amplifier (absolutely, completely, entirely, ...)
104	rb+down+++	adverb + downtoner (nearly, only, merely, ...)
105	rb+emph+++	adverb + emphatic (just, really, so, ...)

106	rb+hdg+++	adverb + hedge (almost, maybe, ...)
107	rb"+hdg"+++	multi-word adverb + hedge (kind of, sort of)
108	rb+phrv+++	adverb + phrasal verb (get in, wrap up, ...)
109	rb+tm+++	adverb + time marker (afterwards, again, immediately, ...)
110	rb+dspt+++	adverb + discourse particle (anyway, well, ...)
111	rbr++++	comparative adverb (better, quicker)
112	rbr+tm+++	comparative adverb + time marker (earlier, later, sooner, ...)
113	rn+pl+++	nominal adverb + place marker (here, there)
114	rn+tm+++	nominal adverb + time marker (now, then)
115	rn+dspt+++	nominal adverb + discourse particle (now)
116	rp++++	adverbial particle (back, in, round, up, ...)
117	rp+pl+++	adverbial particle + place marker (away, behind, out, ...)
118	tht+jcmp+++	that as dependent clause head + adjective complement
119	tht+ncmp+++	that as dependent clause head + noun complement
120	tht+vcmp+++	that as dependent clause head + verb complement
121	tht+rel+++	that as dependent clause head + relative clause
122	tht+rel+obj++	that as dep. clause head + relative clause + object position
123	tht+rel+subj++	that as dep. clause head + relative clause + subject position
124	to++++	infinitive marker
125	to"++++	multi-word infinitive marker (in order to)
126	uh++++	interjection/filler (hey, oh, ok, yes, erm ...)
127	vb++++	base form of verb, excluding verbs in infinitive clauses
128	vb+++xvbn+	base form of verb + + + past participle form
129	vb+be+aux++	base form of verb + be + auxiliary verb
130	vb+be+vrb++	base form of verb + be + main verb
131	vb+bem+aux++	verb + am + auxiliary verb
132	vb+bem+aux++0	verb + am + auxiliary verb + + contracted (m)
133	vb+bem+vrb++	verb + am + main verb

134	vb+bem+vrb++0	verb + am + main verb + + contracted (m)
135	vb+ber+aux++	verb + are + auxiliary verb
136	vb+ber+aux++0	verb + are + auxiliary verb + + contracted (re)
137	vb+ber+vrb++	verb + are + main verb
138	vb+ber+vrb++0	verb + are + main verb + + contracted (re)
139	vb+do+aux++	verb + do + auxiliary verb
140	vb+do+vrb++	verb + do + main verb
141	vb+hv+aux++	verb + have + auxiliary verb
142	vb+hv+aux++0	verb + have + auxiliary verb + + contracted (ve)
143	vb+hv+vrb++	verb + have + main verb
144	vb+hv+vrb++0	verb + have + main verb + + contracted (ve)
145	vb+seem+++	base form of verb + seem/appear
146	vb+vprv+++	base form of verb + private verb (believe, feel, think, ...)
147	vb+vprv+tht0++	base form of verb + private verb + that deletion **
148	vb+vpub+++	base form of verb + public verb (assert, complain, say, ...)
149	vb+vpub+tht0++	base form of verb + public verb + that deletion **
150	vb+vsua+++	base form of verb + suasive verb (ask, command, insist, ...)
151	vbd+++xvbn+	past tense verb + + + past participle form
152	vbd+bed+aux++	past tense verb + were + auxiliary verb
153	vbd+bed+vrb++	past tense verb + were + main verb
154	vbd+bedz+aux++	past tense verb + was + auxiliary verb
155	vbd+bedz+vrb++	past tense verb + was + main verb
156	vbd+dod+aux++	past tense verb + did + auxiliary verb
157	vbd+dod+vrb++	past tense verb + did + main verb
158	vbd+hvd+aux++	past tense verb + had + auxiliary verb
159	vbd+hvd+vrb++	past tense verb + had + main verb
160	vbd+seem++xvbn+	past tense verb + seem/appear
161	vbd+vprv++xvbn+	past tense + private verb (believe, feel, think, ...)
162	vbd+vprv+tht0+xvbn+	past tense + private verb + that deletion **
163	vbd+vpub++xvbn+	past tense + public verb (assert, complain, say, ...)
164	vbd+vpub+tht0+xvbn+	past tense + public verb + that deletion **

165	vbd+vsua++xvbn+	past tense + suasive verb (ask, command, insist, ...)
166	vbg+++xvbg+	present progressive verb + + + -ing form
167	vbg+beg++xvbg+	present progressive verb + being
168	vbg+beg+aux+xvbg+	present progressive verb + being + auxiliary verb
169	vbg+hvg++xvbg+	present progressive verb + having
170	vbg+vprv++xvbg+	pres. prog. + private verb (believe, feel, think, ...)
171	vbg+vprv+thtO+xvbg+	present progressive + private verb + that deletion **
172	vbg+vpub++xvbg+	pres. prog. + public verb (assert, complain, say, ...)
173	vbg+vpub+thtO+xvbg+	present progressive + public verb + that deletion**
174	vbg+vsua++xvbg+	pres. prog. + suasive verb (ask, command, insist, ...)
175	vwbg+++xvbg+	present progressive postnominal modifier
176	vwbg+beg++xvbg+	present progressive postnominal modifier + being
177	vwbg+hvg++xvbg+	present progressive postnominal modifier + having
178	vwbg+vprv++xvbg+	present prog. postnom. modifier + private verb
179	vwbg+vpub++xvbg+	present prog. postnom. modifier + public verb
180	vbi++++	base form of verb in infinitive clause
181	vbi+vprv+++	infinitive verb + private verb (believe, feel, think, ...)
182	vbi+vprv+tht0++	infinitive verb + private verb + that deletion **
183	vbi+vpub+++	infinitive verb + public verb (assert, complain, say, ...)
184	vbi+vpub+tht0++	infinitive verb + public verb + that deletion **
185	vbi+vsua+++	infinitive verb + suasive verb (ask, command, insist, ...)
186	vbz	++++ 3rd person singular verb
187	vbz+bez+aux++	3rd person sg. verb + is + auxiliary verb
188	vbz+bez+aux++0	3rd person sg. + is + auxiliary verb + + contracted (s)
189	vbz+bez+vrb++	3rd person sg. verb + is + main verb
190	vbz+bez+vrb++0	3rd person sg. + is + main verb + + contracted (s)
191	vbz+doz+aux++	3rd person sg. verb + does + auxiliary verb
192	vbz+doz+vrb++	3rd person sg. verb + does + main verb
193	vbz+hvz+vrb++	3rd person sg. verb + has + main verb
194	vbz+seem+++	3rd person sg. verb + seem/appear
195	vbz+vprv+++	3rd person sg. + private verb (believe, feel, think, ...)
196	vbz+vprv+tht0++	3rd person sg. + private verb + that deletion **

197	vbz+vpub+++	3rd person sg. + public verb (assert, complain, say, ...)
198	vbz+vpub+tht0++	3rd person sg. + public verb + that deletion **
199	vbz+vsua+++	3rd person sg. + suasive verb (ask, command, insist, ...)
200	vprf+++xvbn+	perfect aspect verb + + + past participle form
201	vprf++tht0+xvbn+	perfect aspect verb + + that deletion **
202	vprf+ben+aux+xvbn+	perfect aspect verb + been + auxiliary verb
203	vprf+ben+vrb+xvbn+	perfect aspect verb + been + main verb
204	vpsv++agls+xvbn+	main clause passive verb + + agentless passive
205	vpsv++by+xvbn+	main clause passive verb + + by passive
206	vwbn+++xvbn+	passive postnominal modifier + + + past participle form
207	vwbn+vprv++xvbn+	passive postnominal modifier + private verb
208	vwbn+vpub++xvbn+	passive postnominal modifier + public verb
209	vwbn+vsua++xvbn+	passive postnominal modifier + suasive verb
210	wdt+who+++	WH determiner + WH word (what, whatever, whichever, ...)
211	wdt+who+whcl++	WH determiner + WH word + WH clause
212	wdt+who+whq++	WH determiner + WH word + WH question
213	whp+rel+obj++	WH pronoun + relative clause + object position
214	whp+rel+pied++	WH pronoun + relative clause + object position with
215	whp+rel+subj++	WH pronoun + relative clause + subject position
216	whp+who+++	WH pronoun + WH word (not a relative clause)
217	whp+who+whq++	WH pronoun + WH word + WH question
218	wrb+who+++	WH adverb (how, when, where, ...) + WH word
219	wrb+who+whcl++	WH adverb + WH word + WH clause
220	wrb+who+whq++	WH adverb + WH word + WH question
221	xnot++not++	not + + negation
222	xnot++not++0	not + + negation + + contracted form (nt)
223	xvbn+++xvbn+	past participle form - indeterminate grammatical function
224	xvbg+++xvbg+	present participle form - indeterminate grammatical
225	zz++++	letter of the alphabet