



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Câmpus de São José do Rio Preto

LUCIANO FRANCO DA SILVA

A LINGUAGEM FORMULAICA NAS TED TALKS:
UMA INVESTIGAÇÃO SOBRE A VARIAÇÃO DISCIPLINAR DE PACOTES LEXICAIS

São José do Rio Preto
2023

LUCIANO FRANCO DA SILVA

A LINGUAGEM FORMULAICA NAS TED TALKS:

UMA INVESTIGAÇÃO SOBRE A VARIAÇÃO DISCIPLINAR DE PACOTES LEXICAIS

Tese apresentada como parte dos requisitos para a obtenção do título de Doutor em Estudos Linguísticos, junto ao Programa de Pós-Graduação em Estudos Linguísticos da Universidade Estadual “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Financiadora: CAPES

Orientadora: Prof.^a. Dr^a. Paula Tavares Pinto

São José do Rio Preto
2023

S586l Silva, Luciano Franco da
A linguagem formulaica nas TED Talks : uma investigação sobre a
variação disciplinar de pacotes lexicais / Luciano Franco da Silva. --
São José do Rio Preto, 2023
199 f. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp),
Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientadora: Paula Tavares Pinto

1. Pacotes lexicais. 2. Linguística de corpus. 3. Linguagem
formulaica. 4. TED Talks. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de
Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

LUCIANO FRANCO DA SILVA

A LINGUAGEM FORMULAICA NAS TED TALKS:
UMA INVESTIGAÇÃO SOBRE A VARIAÇÃO DISCIPLINAR DE PACOTES LEXICAIS

Tese apresentada como parte dos requisitos para a obtenção do título de Doutor em Estudos Linguísticos, junto ao Programa de Pós-Graduação em Estudos Linguísticos da Universidade Estadual “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Financiadora: CAPES

Comissão Examinadora

Prof.^a. Dr.^a. Paula Tavares Pinto
Universidade Estadual Paulista (UNESP)
Orientadora

Prof. Dr. Vander Viana
Universidade de Edimburgo

Prof.^a. Dr.^a. Deise Prina Dutra
Universidade Federal de Minas Gerais (UFMG)

Prof.^a. Dr.^a. Patrícia Bértoli
Universidade Estadual do Rio de Janeiro (UERJ)

Prof.^a. Dr.^a. Talita Serpa
Universidade Estadual Paulista (UNESP)

São José do Rio Preto
22 de setembro de 2023

*Aos meus pais, Mario J. Ribeiro da Silva (in memoriam) e Maria Franco Guimarães
À minha esposa, Luana*

AGRADECIMENTOS

A Deus, pela força e serenidade concedidas em todos os momentos desta jornada.

À minha família, meus pais, Mario (*in memoriam*) e Maria, minha esposa, Luana, pelo respaldo, compreensão e estímulo ao longo deste trabalho. Sem o apoio de vocês, esta tese não teria se concretizado.

Minha eterna gratidão e carinho à minha orientadora, Prof.^a Dr.^a Paula Tavares Pinto, pela competência, tranquilidade e delicadeza. Agradeço especialmente pela confiança depositada em mim, pela liberdade que me foi dada e pelas palavras amigas nos momentos desafiadores.

Ao Prof. Dr. Vander Viana, agradeço por todos os ensinamentos e contribuições ao longo desta jornada acadêmica. Muito obrigado por me ajudar a crescer, pelos conselhos e por ser uma fonte de inspiração e admiração tanto pessoal quanto profissionalmente.

Agradeço às Prof.^a Dr.^a Deise Prina Dutra e Prof.^a Dr.^a Patrícia Bértoli pelas valiosas contribuições ao meu trabalho e aos demais professores da pós-graduação do Ibilce: Prof.^a Dr.^a Adriane Orenha-Ottaiano, Prof.^a Dr.^a Diva Cardoso de Camargo, Prof. Dr. Douglas Consolo, Prof.^a Dr.^a Fabiana Komesu, Prof. Dr. Lauro Maia Amorin, Prof.^a Dr.^a Lília Abreu-Tardeli, Prof.^a Dr.^a Solange Aranha e Prof.^a Dr.^a Suzi Cavalari. Suas disciplinas foram fundamentais para meu crescimento acadêmico e profissional.

A todos os meus familiares e amigos que contribuíram para a realização deste trabalho, um agradecimento especial a Adriano Boaretti, cujo apoio e incentivo durante mais de uma década foram inestimáveis, e à Elen Dias, pela amizade, conselhos e ensinamentos. Sua confiança em mim, mesmo nos momentos em que duvidei de mim mesmo, foi crucial. Sem você, não teria chegado tão longe.

À Amanda, Alex, Ana, Dorival, Fernanda, Juliana, Lucas, Paula e todos os queridos amigos, obrigado pela amizade sincera, confiança e companheirismo.

Aos companheiros da pós-graduação, Aline Cantarotti, Carol Tavares, Jeane Cardoso, José Victor de Souza, Luana Viana, Talita Serpa, Tamiris Destro, Silmara Moscatelli e Willian Garcia: compartilhamos muitos momentos juntos, desde os cafés até os congressos e grupos de estudos.

Por fim, expresso minha gratidão aos estudantes e servidores do IFPR-Goioerê.

Este trabalho foi realizado com o valioso apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001*, à qual agradeço.

RESUMO

A tese ora apresentada delinea uma investigação acerca da linguagem formulaica em TED Talks, um gênero discursivo híbrido de palestras voltadas a popularização do conhecimento. A pesquisa centra-se na identificação e análise de marcadores de variação disciplinar presentes nos padrões linguísticos mais recorrentes em um conjunto de 1.519 transcrições de TED Talks. Estas foram associadas a 34 disciplinas acadêmicas e organizadas em dois *corpora* de estudo de áreas disciplinares distintas: TED-STEM [Ciências, Tecnologia, Engenharia e Matemática] e TED-SHAPE [Ciências Sociais, Humanidades e Artes para o Povo e para Economia]. A base teórica deste trabalho está ancorada em três vertentes: variação disciplinar, linguagem formulaica e linguística de *corpus*. A perspectiva de registro embasou a compreensão de como características linguísticas recorrentes se relacionam ao contexto situacional e ao propósito comunicativo subjacentes às TED Talks. A concepção da natureza formulaica da língua e as implicações envolvendo sua investigação por meio de *corpora* viabilizaram um estudo com foco nos pacotes lexicais – sequências fixas de três ou mais palavras, altamente frequentes e distribuídas em um registro. Utilizando parâmetros distribucionais (frequência relativa e dispersão), este estudo abordou pacotes lexicais formados por quatro palavras, com frequência relativa mínima de 20 ocorrências por milhão de palavras e presentes em, pelo menos, 4% dos textos de cada *corpora*. Foram identificados 138 pacotes lexicais distintos no *corpus* TED-STEM e 114 no *corpus* TED-SHAPE. Os resultados sugerem que palestrantes vinculados ao domínio STEM tendem a mobilizar um conjunto maior de expressões padronizadas para expressar posicionamento, aludir a referências e organizar e reformular o discurso quando comparados aos palestrantes que trabalham no domínio SHAPE. Mais especificamente, as funções discursivas de direcionamentos, habilidades e elaboração/clarificação de tópicos revelaram-se como altamente padronizadas no *corpus* TED-STEM. Variações estruturais também foram evidenciadas, com pacotes lexicais que incorporam orações subordinadas (*e.g. what we do is*) apresentando uma proporção ligeiramente maior no *corpus* TED-STEM, ao passo que pacotes lexicais compostos por fragmentos de sintagmas nominais e preposicionais (*e.g. at the end of*) ocorreram em maior proporção no *corpus* TED-SHAPE. Os achados evidenciam diferenças no modo como palestrantes de ambos os domínios disciplinares interagem com o público-alvo, expressam opiniões e desejos e organizam seus discursos. Na conclusão desta tese, são apresentadas as principais características do discurso disciplinar das TED Talks, e esboçadas sugestões para o aproveitamento dos dados em futuras investigações sobre a linguagem formulaica em discursos de popularização do conhecimento.

Palavras-chave: Pacotes lexicais. Linguística de *corpus*. Linguagem formulaica. TED Talks.

ABSTRACT

The thesis presented herein outlines an investigation into formulaic language in TED Talks, a hybrid genre of lectures aimed at popularizing knowledge. The research focuses on identifying and analyzing markers of disciplinary variation present in recurrent linguistic patterns in a collection of 1,519 transcripts of TED Talks. These were associated with 34 academic disciplines and organized into two study corpora of distinct disciplinary areas: TED-STEM [Science, Technology, Engineering, and Mathematics] and TED-SHAPE [Social Sciences, Humanities, and Arts for People and the Economy]. The theoretical framework is based on three key aspects: register variation, formulaic language, and corpus linguistics. The perspective of disciplinary variation underpinned the understanding of how recurrent linguistic features relate to the situational context and the underlying communicative purpose of TED Talks. The notion of formulaic language and the implications surrounding its investigation through corpora enabled a study focused on lexical bundles – fixed sequences of three or more words, highly frequent and distributed across a register. Employing distributional parameters (relative frequency and dispersion), this study focused on four-word lexical bundles, occurring at least 20 times per million words, across 4% of the texts in each corpus. A total of 138 distinct lexical bundles were identified in the TED-STEM corpus and 114 in the TED-SHAPE corpus. The findings suggest that speakers affiliated with the STEM domain tend to deploy a broader set of standardized expressions to signal stance, time/place references, and organize and reformulate their discourse compared to their SHAPE counterparts. More specifically, discursive functions of directives, abilities, and topic elaboration/clarification emerged as highly standardized in the TED-STEM corpus. Structural variations were also highlighted: lexical bundles incorporating dependent clauses fragments (e.g. what we do is) were more frequent in the TED-STEM corpus, while lexical bundles consisting of noun and prepositional phrase fragments (e.g. at the end of) occurred more frequently in the TED-SHAPE corpus. The findings highlight differences in how speakers from both disciplinary domains engage with their target audience, express opinions, and desires, and organize their discourse. In the concluding remarks, the primary features regarding the disciplinary discourse of TED Talks are presented, and suggestions are outlined for leveraging the data in future investigations on formulaic language in science popularization discourses.

Keywords: Lexical bundles. Corpus linguistics. Formulaic language. TED Talks.

LISTA DE FIGURAS E GRÁFICOS

Figura 1 – Principais características situacionais das TED Talks	27
Figura 2 – Caminho metodológico em estudos dirigidos por <i>corpus</i>	35
Figura 3 – Relação teoria/dados empíricos em linguística de <i>corpus</i>	36
Figura 4 – Diferentes prioridades na identificação de linguagem formulaica	40
Figura 5 – Principais características relacionadas aos pacotes lexicais	50
Figura 6 – Relação de representatividade em um <i>corpus</i> linguístico	59
Figura 7 – Diferenças de estilos entre as TED Talks e as TED-Ed	66
Figura 8 – Etiquetas com a letra A na plataforma <i>TED Talks</i>	67
Figura 9 – Organização dos agrupamentos multidisciplinares	69
Figura 10 – Modelo de arquivamento dos metadados	71
Figura 11 – Modelo de extração das URLs	73
Figura 12 – Segmentação dos textos pela plataforma Lexos®	74
Figura 13 – Modelo de cabeçalho do Sketch Engine® em SGML	74
Figura 14 – Organização dos arquivos dos <i>corpora</i> de estudo	75
Figura 15 – Ferramenta <i>N-Gram</i>	80
Figura 16 – Visualização da <i>Key Word in Context</i> pela ferramenta <i>Concordance</i>	81
Gráfico 1 – Distribuição dos textos por ano	76
Gráfico 2 – Distribuição da FR e dispersão dos dados entre os <i>corpora</i> de estudo	93
Gráfico 3 – Porcentagem das ocorrências dos tipos de pacotes lexicais	94
Gráfico 4 – Distribuição de ocorrências de pacotes lexicais de posicionamento entre os <i>corpora</i> de estudo	96
Gráfico 5 – Pacotes lexicais de posicionamento epistêmico	98
Gráfico 6 – Pacotes lexicais de desejos e motivações pessoais	103
Gráfico 7 – Pacotes lexicais de necessidades/obrigações	107
Gráfico 8 – Pacotes lexicais de direcionamentos	111
Gráfico 9 – Pacotes lexicais de intenções/predições	115
Gráfico 10 – Pacotes lexicais de habilidades	119
Gráfico 11 – Distribuição de ocorrências de pacotes lexicais de organização discursiva entre os <i>corpora</i> de estudo	129

Gráfico 12 – Pacotes lexicais de introdução de tópicos	131
Gráfico 13 – Pacotes lexicais de elaboração/clarificação	135
Gráfico 14 – Pacotes lexicais de identificação/foco	148
Gráfico 15 – Distribuição dos pacotes lexicais de expressão de referências	158
Gráfico 16 – Pacotes lexicais de quantificações	159
Gráfico 17 – Pacotes lexicais de atributos tangíveis	164
Gráfico 18 – Pacotes lexicais de atributos intangíveis	166
Gráfico 19 – Pacotes lexicais de referência temporal, espacial e multifuncional	169

LISTAS DE QUADROS

Quadro 1 – Principais características situacionais dos destinatários das TED Talks	23
Quadro 2 – Movimentos/passos retóricos típicos em TED Talks	28
Quadro 3 – Taxonomia estrutural proposta por Biber, Conrad e Cortes (2004)	47
Quadro 4 – Taxonomia funcional proposta por Biber, Conrad e Cortes (2004)	48
Quadro 5 – Taxonomia funcional proposta por Hyland (2008a)	49
Quadro 6 – Descrição dos metadados	71
Quadro 7 – Descrição da codificação SGML utilizada	73
Quadro 8 – Taxonomia de <i>corpora</i> linguísticos	76
Quadro 9 – Principais particularidades do discurso disciplinar das TED Talks evidenciadas por pacotes lexicais	180

LISTA DE TABELAS

Tabela 1 – Coletânea de trabalhos relacionados a pacotes lexicais na literatura acadêmica	51
Tabela 2 – Etiquetas disciplinares no site TED Talks	68
Tabela 3 – Distribuição do <i>corpus</i> TED-STEM	77
Tabela 4 – Distribuição do <i>corpus</i> TED-SHAPE	78
Tabela 5 – Pacotes lexicais em contexto de sobreposição no <i>corpus</i> TED-STEM	86
Tabela 6 – Pacotes lexicais em contexto de sobreposição no <i>corpus</i> TED-SHAPE	88

LISTAS DE ABREVIATURAS E SIGLAS

BASE – *British Academic Spoken English Corpus*

BAWE – *British Academic Written English Corpus*

BNC – *British National Corpus*

COBUILD – *Collins Birmingham University International Language Database*

DiA – Dispersão absoluta nos textos do *corpus*

DiR% – Dispersão relativa nos textos do *corpus* (em porcentagem)

EAP – *English for Academic Purposes*

FA – Frequência Absoluta

FR – Frequência Relativa

KWIC – *Key Word in Context*

L2 – Segunda Língua

LC – Linguística de *Corpus*

LE – Língua Estrangeira

LI – Língua Inglesa

LinFE – Línguas para Fins Específicos

MICASE – *Michigan Corpus of Academic Spoken English*

OS – Oração Subordinada

SGML – *Standard Generalized Markup Language*

SHAPE – *Social Sciences, Humanities and the Arts for People and the Economy*

STEM – *Science, Technology, Engineering and Mathematics*

SN – Sintagma Nominal

SP – Sintagma Preposicional

SV – Sintagma Verbal

TCSE – *TED Corpus Search Engine*

TED – *Technology, Engineering and Design*

URL – *Uniform Resource Locator*

WIT³ – *Web Inventory of Transcribed and Translated Talks*

SUMÁRIO

1	INTRODUÇÃO	16
2	FUNDAMENTAÇÃO TEÓRICA	21
2.1	Análise situacional de TED Talks	21
2.2	Linguística de <i>corpus</i> : concepções centrais, tipos de estudos e estatuto do campo	30
2.3	Linguagem formulaica: conceitos e contribuições para os estudos linguísticos	38
2.4	Pacotes lexicais	43
2.5	Fundamentação teórica: Considerações finais	55
3	MATERIAL E MÉTODOS	57
3.1	Princípios envolvidos na coleta de textos e compilação de <i>corpora</i> linguísticos	57
3.2	Coleta dos textos, organização dos metadados e tipologia dos <i>corpora</i> de estudo	67
3.3	Descrição dos <i>corpora</i> TED-STEM e TED-SHAPE	75
3.4	Ferramentas computacionais adotadas para exploração dos <i>corpora</i> de estudo	78
3.5	CrITÉrios para extração dos pacotes lexicais	81
3.6	Refinamento dos pacotes lexicais relacionados ao tÓpico e em contexto de sobreposição	84
3.7	Material e métodos: Considerações finais	90
4	APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS	92
4.1	Diferenças na recorrência e estrutura dos pacotes lexicais extraídos dos <i>corpora</i> TED-STEM e TED-SHAPE	92
4.2	Pacotes lexicais de expressão de posicionamento	95
4.2.1	Posicionamento epistêmico	97
4.2.2	Posicionamento atitudinal: desejos e motivações pessoais	103
4.2.3	Posicionamento atitudinal: necessidades/obrigações	107
4.2.4	Posicionamento atitudinal: direcionamentos	110
4.2.5	Posicionamento atitudinal: intenções/predições	114
4.2.6	Posicionamento atitudinal: habilidades	118
4.2.7	Expressão de posicionamento: o que os dados nos revelam?	123
4.3	Pacotes lexicais organizadores discursivos	128
4.3.1	Introdução de tÓpicos	130
4.3.2	Elaboração/clarificação de tÓpico	134

4.3.3 Identificação/foco	147
4.3.4 Organizadores discursivos: o que os dados nos revelam?	154
4.4 Pacotes lexicais de expressões de referências	157
4.4.1 Especificação de atributos	159
4.4.2 Referências de tempo, espaço e multifuncional	168
4.4.3 Expressões de referência: o que os dados nos dados revelam?	174
5 CONSIDERAÇÕES FINAIS E ENCAMINHAMENTOS	176
REFERÊNCIAS	184
APÊNDICE A – Pacotes lexicais extraídos do <i>corpus</i> TED-STEM	195
APÊNDICE B – Pacotes lexicais extraídos do <i>corpus</i> TED-SHAPE	198

1 INTRODUÇÃO

A linguagem formulaica constitui um tema de interesse crescente e tem sido explorada por meio de diferentes abordagens e métodos de investigação no campo dos estudos linguísticos. De acordo com Wray (2002), o conceito de linguagem formulaica está relacionado à noção de união, costumes e hábitos intrínsecos ao falante, estando, assim, virtualmente presente em todos os aspectos da comunicação humana. Contudo, na literatura acadêmica, ainda persiste uma certa ambiguidade da definição do que é ou não formulaico na comunicação humana (Granger; Paquot, 2008; Schmitt; Carter, 2004; Siyanova-Chanturia, 2015; Wray, 2002). Um exemplo disso é a grande quantidade de termos na literatura que denotam a natureza formulaica da língua inglesa, LI, tais como: expressões idiomáticas (*e.g. kick the bucket*), colocações (*e.g. fast food*), fórmulas (*e.g. nice to see you again!*), entre outros. Apesar de estarem relacionados à formulaicidade da língua, Conrad e Biber (2004) esclarecem que existem certas diferenças fundamentais entre tipos de sequências formulaicas no que tange aos seus componentes internos, funções discursivas e método de identificação.

A presente tese investigará a linguagem formulaica no gênero palestras TED – acrônimo de *Technology, Entertainment, Design*¹, sob a perspectiva da abordagem baseada em frequência e da exploração de *corpora* eletrônicos. Para tanto, adotei como objeto de estudo os pacotes lexicais (*lexical bundles*), termo introduzido por Biber et al. (1999) para descrever padrões lexicogramaticais fixos e altamente recorrentes em *corpus*, que operam como blocos pré-fabricados utilizados pelos falantes em diversos registros linguísticos. Na literatura acadêmica, a identificação de pacotes lexicais em um *corpus* é comumente realizada com base em dois critérios: frequência relativa mínima (por milhão de palavras) e dispersão em um número mínimo de textos do *corpus*. O primeiro critério assegura a seleção de sequências recorrentes o suficiente para serem consideradas típicas de um registro específico da língua, enquanto o segundo previne a idiosincrasia, ou seja, corrobora que o pacote lexical identificado seja um padrão lexicogramatical presente no *corpus* como um todo, e não restrito a alguns textos.

¹ Tecnologia, Entretenimento, Planejamento, em português.

Certamente, os estudos pioneiros de John Sinclair pavimentaram o caminho para diversas abordagens de investigação da linguagem formulaica no âmbito da linguística de *corpus*, LC. Em especial, o princípio idiomático propõe que “grande parte da produção linguística de um falante ocorre por meio do uso recorrente de frases pré-construídas, codificadas e decodificadas como entidades únicas na língua, e não como um segmento de palavras com significados individuais²” (Sinclair, 1991, p. 110). Nessa perspectiva, a LC se destaca por ser um modo de investigação que prioriza observações empíricas e sistemáticas das variações e padrões de uso da língua natural (Biber; Reppen, 2015). Tal abordagem se apresenta como uma alternativa prática e objetiva de identificação, análise e descrição de sequências formulaicas recorrentes em grandes quantidades de dados.

Embora a literatura sobre variação disciplinar se concentre majoritariamente nos registros escritos da língua (Hyland, 2004a; 2006; 2011a; Hyland; Jiang, 2017; 2019), o discurso acadêmico oral tem recebido crescente atenção nos últimos anos, especialmente com o surgimento de *corpora* especializados como o *British Academic Spoken English* (BASE) (Nesi; Thompson, 2006) e o *Michigan Corpus of Academic Spoken English* (MICASE) (Simpson; Lee; Leicher, 2002). Além disso, estudos têm mostrado que, assim como nos registros escritos, os discursos acadêmicos orais também apresentam marcas de variações entre diferentes disciplinas e gêneros.

No tocante aos objetivos, a presente tese busca investigar o que o uso de pacotes lexicais pode revelar acerca das características do discurso disciplinar entre dois *corpora* de estudo compilados a partir das transcrições de TED Talks, um gênero textual recente voltado à popularização e divulgação do conhecimento científico para o público leigo. Os *corpora* de estudo são denominados conforme a seguintes temáticas: Ciência, Tecnologia, Engenharia e Matemática (TED-STEM³), e Ciências Sociais, Humanidades e Artes para o Povo e para Economia (TED-SHAPE⁴). Já a escolha deste tema é motivada pela relevância sociocultural do movimento de popularização e divulgação científica em nossa atual era da informação, caracterizado como a mobilização de debates, recursos e produtos que facilitem a veiculação

² No original: *The principle of idiom is that language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments.*

³ Em inglês: *Science, Technology, Engineering, and Mathematics.*

⁴ Em inglês: *Social Sciences, Humanities and the Arts for People and the Economy.*

e acesso às informações a fim de reduzir barreiras entre as esferas de atividade acadêmica, científica e tecnológica e o restante da sociedade (Beacco et al. 2002; Calsamiglia, 2003; Motta-Roth; Scherer, 2016).

Becher e Trowler (2001) argumentam que o conhecimento científico possui natureza social e é construído dentro de diferentes comunidades/tribos disciplinares com interesses específicos, literatura, normas e convenções que moldam a maneira como os pesquisadores percebem e interpretam a realidade. Nesse contexto, a recodificação do discurso especializado com o objetivo de facilitar o acesso a conteúdos de divulgação científica assume papel crucial na democratização do conhecimento. Logo, a investigação do discurso em TED Talks se justifica pelo reconhecimento global desses eventos em produzirem conteúdos de fácil acesso, promover debates sobre temas especializados e de alto impacto, além de fomentar o desenvolvimento do espírito científico, epistemológico e empreendedor em nossa sociedade contemporânea (Caliendo, 2014; Compagnone, 2015; Mattiello, 2017).

A principal questão que orienta este trabalho é:

- **Como os pacotes lexicais ajudam a evidenciar características do discurso das TED Talks de diferentes domínios disciplinares?**

Para responder a essa questão macro de pesquisa, será necessário investigar a linguagem formulaica nos *corpora* de estudo pela perspectiva da abordagem baseada em frequência, que prioriza aspectos tais como: alta recorrência e dispersão no *corpus*, fixidez e tamanho da sequência (Conrad; Biber, 2004). Assim, apresento a seguir as questões específicas que guiam a pesquisa:

- i. Como a análise de pacotes lexicais de quatro palavras pode distinguir o discurso das TED Talks associadas aos domínios disciplinares STEM e SHAPE?
- ii. Como o conjunto de pacotes lexicais mais recorrentes entre esses domínios disciplinares varia em termos de frequência, dispersão e estrutura gramatical?
- iii. Como as funções linguísticas de *posicionamento*, *organização discursiva* e *expressão referencial* podem elucidar diferenças na estruturação do discurso entre TED Talks das áreas STEM e SHAPE?

Apesar das características discursivas das TED Talks já terem sido investigadas em trabalhos anteriores (Caliendo; Compagnone, 2014; Chang; Huang, 2015; Liu; Chen, H. 2020b), o presente estudo busca contribuir no preenchimento de uma lacuna apontada por Caliendo (2014), na qual a autora destaca a necessidade de mais estudos que explorem as particularidades disciplinares presentes no discurso da TED Talks.

Hyland (2004a; 2006) caracteriza *disciplinas acadêmicas* como instituições humanas, na qual a criação de conhecimentos e uso da língua são influenciados por fatores de natureza pessoal e interpessoal de seus membros, bem como por normas institucionais e socioculturais da comunidade à qual pertencem. Nesse âmbito, investigações sobre o discurso formulaico acadêmico tornam-se altamente relevantes, pois, para ser aceito como membro de uma comunidade disciplinar, o falante deve demonstrar não apenas competência técnica e teórica em seu campo de atuação, mas também empregar convenções linguísticas fundamentais na criação e manutenção de identidade cultural de seus membros (Becher; Trowler, 2001; Coxhead; Byrd, 2007; Hyland, 2004a; 2008a; 2012).

Diversos estudos têm mostrado que diferentes disciplinas apresentam convenções linguísticas específicas na elaboração de argumentos, justificativas e resultados de suas pesquisas. Poos e Simpson (2002), por exemplo, relatam que tanto em registros orais quanto escritos, os anguladores (*hedges*) *kind of* e *sort of* são mais recorrentes nas disciplinas de humanidades e ciências sociais do que em disciplinas das ciências exatas e biológicas. A influência disciplinar na produção do discurso acadêmico é tão marcante que a premissa de um vocabulário acadêmico comum a todas as disciplinas seria algo improvável (Hyland; Tse, 2007).

A fim de contribuir para reduzir a lacuna apresentada por Caliendo (2014), parto da hipótese de que a investigação de pacotes lexicais possui potencial de revelar características linguístico-discursivas típicas de TED Talks pertencentes aos grupos disciplinares STEM e SHAPE em TED Talks. Tal possibilidade, até onde foi possível descobrir, foi pouco explorada na literatura acerca deste gênero textual.

Tendo apresentado a contextualização e motivações desta pesquisa, apresento abaixo os objetivos específicos desta tese:

- Identificar e categorizar os pacotes lexicais de quatro palavras mais recorrentes e dispersos em ambos os *corpora* de estudo, de acordo com a taxonomia estrutural proposta por Biber, Conrad e Cortes (2004).

- Analisar e comparar o uso de tais pacotes lexicais entre TED Talks para realizar as funções de posicionamento, organização discursiva e referência, propostas por Biber, Conrad e Cortes (2004).

A estrutura desta tese está organizada em mais quatro capítulos. No Capítulo 2, **Fundamentação Teórica**, apresento o arcabouço teórico utilizado para a realização deste estudo, em especial, abordo os fundamentos teóricos relacionados à análise situacional das TED Talks (Biber; Conrad, 2009), linguagem formulaica (Conrad; Biber, 2004; Sinclair, 1991; Wray, 2002), com foco nos pacotes lexicais, e das implicações envolvendo o estudo da língua por meio de *corpora* (Biber, 2006a; Biber; Barbieri, 2007; Biber; Conrad; Cortes, 2004; Hyland, 2008a; Liu; Chen, H. 2020b; Nesi; Basturkmen, 2006). No terceiro capítulo, **Material e Métodos**, descrevo os passos metodológicos envolvidos na compilação dos *corpora* de estudo (Sinclair, 2005), assim como os critérios utilizados na identificação e extração dos pacotes lexicais (Biber, 2006a; Biber; Conrad; Cortes, 2004; Chen, Y. 2009; Grondona, 2015; Hyland, 2008a). No quarto capítulo, **Apresentação e Discussão dos Resultados**, apresento e discuto o que os dados extraídos revelam acerca do discurso disciplinar das TED Talks. Após este percurso, encerro esta tese com as **Considerações Finais e Encaminhamentos** relacionados ao presente estudo.

2 FUNDAMENTAÇÃO TEÓRICA

O presente capítulo está organizado da seguinte maneira: na Seção 2.1, apresento uma análise situacional das TED Talks seguindo a proposta de Biber e Conrad (2009). Na sequência, na Seção 2.2, discorro sobre o status da LC na presente tese, assim como os principais conceitos e paradigmas de pesquisa envolvendo a exploração de *corpora* nos estudos linguísticos. Na Seção 2.3, apresento o conceito de linguagem formulaica com base nos estudos de Wray (2002), focando na descrição dos critérios utilizados para identificar a linguagem formulaica nos *corpora* de estudo (Biber, 2009; Conrad; Biber, 2004). Na Seção 2.4, abordo o que são os pacotes lexicais e quais são os critérios para a sua extração, além de discutir acerca de pesquisas utilizaram essa unidade de análise para investigar variação disciplinar no discurso acadêmico. Por último, na Seção 2.5, apresento as considerações finais deste capítulo.

2.1 Análise situacional de TED Talks

No Capítulo 1, expliquei o que são as TED Talks, suas origens e o papel importante que desempenham como um dos principais canais de popularização do conhecimento na contemporaneidade. Agora, proponho uma discussão acerca dos aspectos situacionais associados à produção das TED Talks, utilizando o Quadro de Análise Situacional proposto por Biber e Conrad (2009). Segundo esses autores, as diversas variedades de textos podem ser analisadas a partir de três perspectivas: registro, gênero e estilo.

A perspectiva do registro concentra-se nas características linguísticas recorrentes em uma variedade textual que atendem a um determinado propósito comunicativo específico, determinado pela situação de uso em que o texto foi criado. Vale ressaltar que essas características não são restritas a um determinado registro. Por exemplo, o uso de voz passiva (*e.g. was based on, it was found that*) é considerado uma característica típica da escrita acadêmica; no entanto, esta estrutura também pode ser encontrada em virtualmente qualquer registro linguístico.

A perspectiva de gênero, por outro lado, está interessada na maneira como uma variedade textual é convencionalmente estruturada em uma comunidade linguística, ou seja, nos elementos estruturais necessários para se produzir um e-mail, currículo ou artigo

científico, dentre outros. De acordo com Biber e Conrad (2009), a perspectiva de gênero se concentra nos aspectos linguísticos que distinguem um determinado gênero de outros. Por exemplo, as expressões “*Dear Sir*” e “*Best regards*”, embora não sejam particularmente recorrentes, são um traço distintivo do gênero textual e-mail.

Por fim, a perspectiva de estilo também está interessada nos aspectos linguísticos recorrentes em um texto; no entanto, o foco agora está nas particularidades estilísticas, e não funcionais, do texto. Padrões de escrita associados a um determinado autor ou período histórico são exemplos de estudos focados na perspectiva de estilo.

Considerando que o objetivo deste estudo é investigar padrões linguísticos funcionais em TED Talks, focarei minha discussão com base na perspectiva de registro, já que abordarei as características linguísticas mais pervasivas em TED Talks. Sendo assim, nos parágrafos a seguir, discuto as características situacionais das TED Talks com base na proposta de Biber e Conrad (2009).

O primeiro aspecto a ser considerado são os *participantes*, aqueles que produzem o texto e aqueles a quem o texto é destinado. No caso das TED Talks, elas são, em sua maioria, produzidas por um falante individual, sempre identificado publicamente. No que diz respeito à especialização dos palestrantes, é comum que as TED Talks sejam apresentadas por pessoas vinculadas a universidades e grupos de pesquisa; empreendedores, empresários e consultores corporativos, políticos, artistas e escritores. De fato, qualquer pessoa com uma “ideia que mereça ser disseminada⁵” pode se candidatar⁶ a apresentar uma TED Talk.

Em relação aos destinatários, as TED Talks apresentam três perfis de plateia bem distintos: o participante copresente, o remoto, e o usuário da *web*. O Quadro 1 apresenta as principais características associadas a cada um destes perfis.

⁵ Referência ao slogan das TED Talks. Em inglês: “*ideas worth spreading*”.

⁶ Mais informações sobre o processo de candidatura podem ser encontradas em <https://www.ted.com/participate/nominate>. Acesso em maio, 2021.

Quadro 1 – Principais características situacionais dos destinatários das TED Talks

Participante copresente	• Presente fisicamente no evento; • Possibilidade de interação direta com o palestrante; • Deslocamento entre diferentes auditórios/palcos durante a conferência; • Maior tempo de interação ocorre entre os intervalos das palestras; • Devem lidar com diversos tipos de problemas, tais como: cansaço físico e mental, choque de horários, custos com viagens, hotéis e alimentação, dentre outros.
Participante remoto	• Acompanha o evento ao vivo, mas não está fisicamente presente; • Interações são mediadas por computador (e.g. videoconferências, salas de bate-papo on-line); • A troca de salas de videoconferência ocorre por meio de links diretos; • Os participantes podem interagir pelo bate papo sem interferirem na apresentação do palestrante; • O aplicativo TEDConnect⁷ permite aos participantes remotos trocarem mensagens entre si e personalizarem suas biografias, informações de contatos e palestras de interesses; • Podem sofrer problemas técnicos relacionados à perda de conexão com a internet, ou mau funcionamento de hardware e software; • Podem criar <i>playlists</i> para (re)assistirem às TED Talks em qualquer horário.
Usuários da web (on-lookers)	• São os principais consumidores das TED Talks; • Não possuem acesso a todas as palestras, somente aquelas disponibilizadas pelo site; • Os vídeos são disponibilizados no site após semanas ou meses de as apresentações terem sido realizadas; • Não é possível deixarem comentários ou interagirem em tempo real com outros usuários; • Têm acesso a uma versão editada de cada palestra; • Possuem informações limitadas acerca do evento como um todo uma vez que os vídeos podem ser acessados isoladamente; • Podem sofrer problemas técnicos relacionados à perda de conexão com a internet, ou mau funcionamento de hardware e software; • Podem criar <i>playlists</i> para (re)assistirem às TED Talks em qualquer horário.

Fonte: Organizado pelo autor.

Essa análise dos perfis de público evidencia a diversidade e a abrangência das TED Talks, bem como destaca a importância de considerar as várias formas de interação entre os participantes ao investigar aspectos linguísticos funcionais neste contexto. Um aspecto interessante em relação aos usuários da *web* foi observado por Sugimoto et al. (2013). Os autores perceberam que as TED Talks apresentadas por profissionais vinculados a instituições acadêmicas recebem mais comentários e curtidas em redes sociais como o Youtube® em comparação com os conteúdos produzidos por palestrantes não acadêmicos, todavia, a quantidade de visualizações entre esses grupos apresenta números similares.

⁷ Aplicativo na qual os participantes das TED Talks têm acesso aos cronogramas, lista de eventos, lista de convidados e participantes, além de permitir aos usuários trocas de mensagens, perfil personalizado e até mesmo envio de localização durante o evento. Disponível em <https://www.ted.com/attend/conferences/tedconnect>. Acesso em setembro, 2021.

Em relação à compreensão dos textos pelos destinatários, eles devem lidar com aspectos típicos da oralidade, tais como: processamento das informações em tempo real, velocidade de fala e sotaque do palestrante. Contudo, os participantes copresentes estão limitados a receber informações de maneira linear, que não podem ser retomadas a menos que o palestrante repita algo que ele julgue ser relevante. Por outro lado, aqueles que acessam as TED Talks disponibilizadas online possuem diversas ferramentas que auxiliam a compreensão, tais como: possibilidade de pausa, controle de velocidade e legendas. Assim, os participantes remotos e os usuários da *web* não estão limitados a receber informações de forma linear, e podem revisitar trechos do texto quantas vezes forem necessárias e da maneira que lhes for mais conveniente.

A segunda característica situacional envolve o tipo de *relacionamento entre os participantes*. Por serem predominantemente monólogos, quase não há interação entre palestrante e público-alvo durante as TED Talks, salvo ocorrências de perguntas retóricas, curtos momentos de interação ou perguntas feitas pelo mediador do evento ao final da palestra. Por outro lado, uma prática comum empregada por palestrantes é o compartilhamento de informações de eventos relacionados às suas vidas pessoais de modo a se aproximarem de sua plateia e estabelecer engajamento (Caliendo, 2014). No que diz respeito às relações de poder, o palestrante controla a produção do texto e assume o papel de especialista/autoridade no assunto diante do público-alvo. Entretanto, não é possível avaliar o quão diferente é o nível de conhecimento especializado entre o palestrante e os membros da plateia.

Sobre o *canal de produção*, as TED Talks são comumente descritas na literatura como um gênero híbrido mediado pela *web* (Caliendo; Compagnone, 2014; D'Avanzo, 2015; Mattiello, 2017). Tal descrição implica que o ambiente virtual não é apenas uma característica contextual de produção das TED Talks, mas uma parte integrante que as distingue de outros gêneros relacionados à oratória, como palestras, aulas expositivas e sermões. No caso desses gêneros, a *web* é comumente um meio no qual o conteúdo do texto é apresentado. Um professor pode, por exemplo, gravar suas aulas e disponibilizá-las no Youtube® para que seus alunos revisem os conteúdos posteriormente. Nesses casos, a *web* opera simplesmente como um local/meio de reprodução de um conteúdo, originalmente produzido em uma sala de aula, por esse professor.

Por outro lado, as TED Talks possuem uma dimensão midiática bastante particular, isso porque elas utilizam diversos tipos de suportes, como projetores, microfones, iluminação, filmagem e edição de vídeo realizadas por equipes especializadas, tudo para que a transposição das palestras para as mídias digitais ocorra da forma mais natural possível. Além disso, a incorporação de *hyperlinks*, notificações por aplicativos, criação de lista de reprodução personalizadas, compartilhamento de conteúdo e tantas outras funcionalidades disponibilizadas, tornam a dissociação entre TED Talks e ambientes virtuais uma tarefa extremamente difícil ou até mesmo impossível de ser realizada.

Em relação às *circunstâncias de produção*, Caliendo (2014) e Caliendo e Compagnone (2014) descrevem as TED Talks como eventos de fala planejados que apresentam traços tanto da língua oral quanto escrita. Ou seja, os textos são produzidos oralmente com apoio de recursos típicos do modo escrito (anotações, ponto texto e *scripts* das palestras) como guias na produção dos textos. Além disso, é comum os palestrantes optarem pelo estilo de palestras roteirizadas em suas apresentações, conforme explicado por Anderson (2016, p. 138), “na verdade, a maior parte dos oradores do TED roteiriza e decora a palestra inteira e faz de tudo para evitar que *pareça* decorada”. Portanto, as TED Talks são um tipo de texto oral, mas que, por serem previamente planejadas, e até mesmo roteirizadas, também carregam marcas semelhantes aos textos escritos.

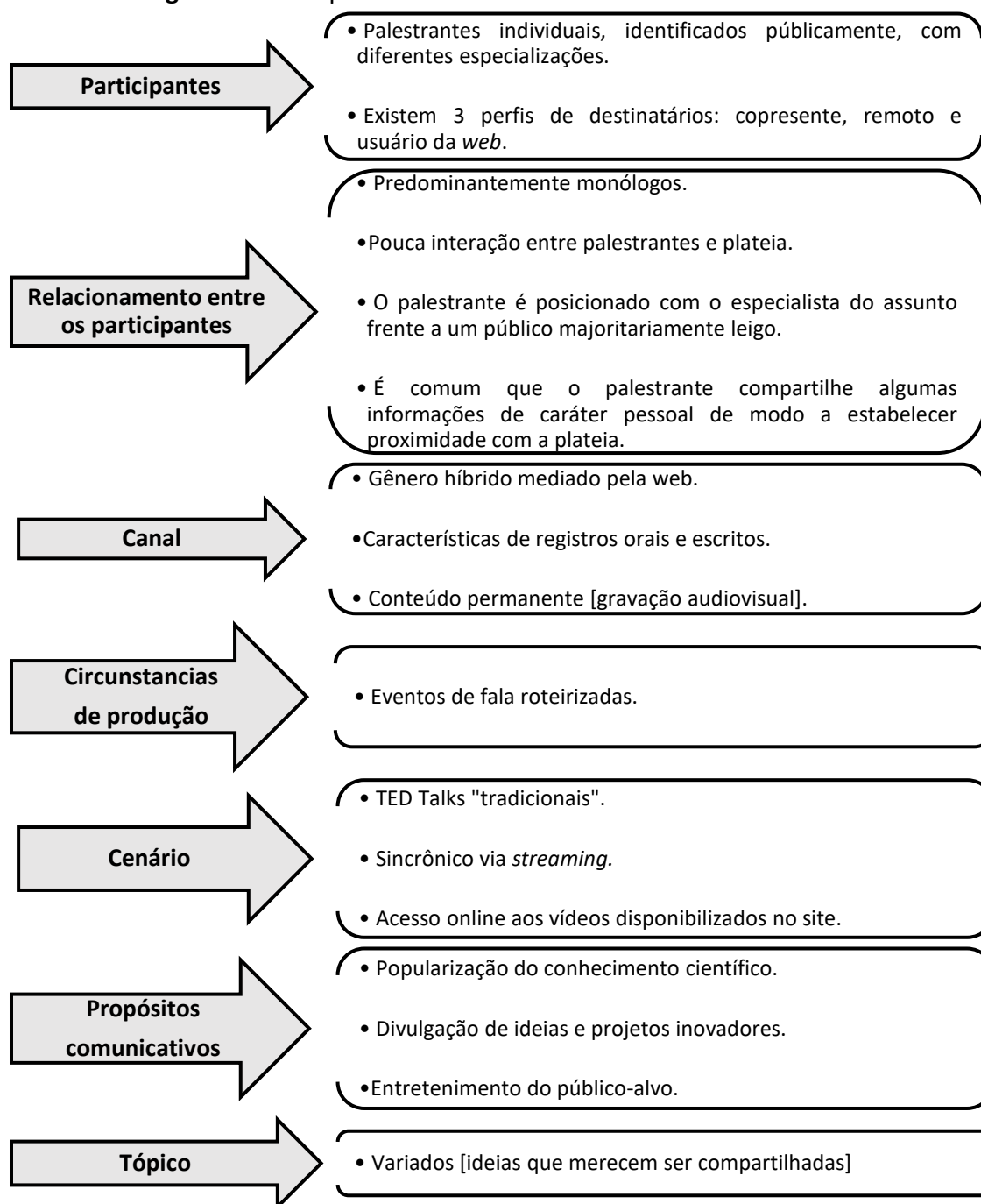
O *cenário*, segundo Biber e Conrad (2009), refere-se ao contexto físico (tempo e espaço) em que a comunicação ocorre. No caso das TED Talks, é possível distinguir entre três tipos de cenários. O primeiro engloba as TED Talks ‘tradicionais’, no qual palestrantes e plateia compartilham o mesmo tempo e espaço físico. O segundo cenário ocorre quando as TED Talks são transmitidas ao vivo por *streaming*. Nestes casos, o local compartilhado entre os participantes passa a ser salas virtuais que podem ou não permitir o uso de *webcams*, microfones ou envio de mensagens da plateia ao palestrante. O último tipo de cenário descreve o modo como os usuários da *web* recebem os textos. Nestes casos, não há compartilhamento de tempo e espaço entre plateia e o participante, e os usuários podem ter acesso a conteúdos produzidos há décadas atrás a qualquer hora e em qualquer local que desejarem.

Já os *propósitos comunicativos* das TED Talks são, em linhas gerais: a popularização do conhecimento, apresentação de projetos inovadores, entretenimento do público-alvo e persuasão da relevância social que tal projeto oferece ou poderá oferecer quando completado

(Ludewig, 2017; Miranda, J., 2018). A última categoria do quadro situacional é referente ao *tópico* abordado. As TED Talks são extremamente heterogêneas e apresentam uma ampla variedade de assuntos, tanto gerais (*e.g.* budismo, mudanças climáticas, segurança cibernética) quanto acadêmicos (*e.g.* arquitetura, matemática, psicologia). Biber e Conrad (2009) explicam que o *tópico* é o aspecto situacional que mais exerce influência sobre as escolhas lexicais de um falante. No caso das TED Talks, é comum que o palestrante simplifique o uso de vocabulário técnico para tornar o conteúdo mais acessível e compreensível a um maior número de pessoas. Mattiello (2019) observou, por exemplo, que as figuras de linguagem como hipérbole, metáforas e perífrases são comumente empregadas para simplificar e substituir termos especializados em TED Talks sobre oncologia.

Em resumo, as variáveis situacionais das TED Talks identificadas no presente estudo podem ser apresentadas da seguinte maneira:

Figura 1 – Principais características situacionais das TED Talks



Fonte: Elaborado pelo autor

As variáveis situacionais apresentadas acima fornecem um quadro abrangente das características das TED Talks e podem servir como base para investigar aspectos linguísticos funcionais dentro desse gênero de comunicação.

Tendo discutido as TED Talks sob a perspectiva do quadro de análise situacional, irei dedicar os últimos parágrafos desta seção a apresentação dos resultados obtidos por Chang e Huang (2015) e Liu e Chen, H. (2020b), que também buscaram descrever as principais

características desse gênero sob a perspectiva de movimentos e passos retóricos e pela comparação com o gênero palestra acadêmica, respectivamente.

No tocante à perspectiva de movimentos e passos retóricos, originalmente proposta por Swales (1990) e aplicada às TED Talks no estudo de Chang e Huang (2015), os autores sugerem que essas palestras são comumente realizadas por meio de sete movimentos retóricos e seus respectivos passos, conforme o Quadro 2.

Quadro 2 – Movimentos/passos retóricos típicos em TED Talks

Movimentos	Passos
1. <i>Orientado ao ouvinte</i>	• Cumprimentar e estabelecer conexão com a plateia
2. <i>Introdução do tópico</i>	• Contextualização • Anunciação do tópico • Apresentação da estrutura da palestra
3. <i>Apresentação do palestrante</i>	• Introdução do palestrante • Estabelecimento de autoridade • Posicionamento acerca do tópico
4. <i>Desenvolvimento do tópico</i>	• Apresentação de um argumento • Explicação • Descrição de um processo/séries de eventos
5. <i>Fechamento</i>	• Chamadas para ação (<i>call for action</i>) • Generalizações e especulações sobre o tema discutido
6. <i>Mensagens finais</i>	
7. <i>Agradecimentos</i>	

*Os movimentos 1 e 3 são não-obrigatórios

Fonte: Traduzido de Chang e Huang (2015, p. 38).

Chang e Huang (2015) ressaltam que os movimentos e passos apresentados no Quadro 2 não são necessariamente lineares e nem obrigatórios à produção das TED Talks, mas sim sensíveis às necessidades de cada palestrante. Os autores argumentam que esses movimentos ocorrem de forma cíclica, e podem ser encontrados mais de uma vez em um mesmo texto.

Resultados semelhantes aos que foram reportados por Chang e Huang (2015) também foram encontrados no estudo de Miranda J. (2018), que também buscou categorizar os movimentos retóricos em TED Talks. Contudo, a classificação proposta por Miranda J. (2018) apresenta apenas os movimentos: *introdução*, *apresentação do palestrante*, *desenvolvimento do tópico*, *mensagens finais* e *agradecimento*. Segundo o autor, tais movimentos foram

identificados em todas as TED Talks investigadas em seu estudo, e, portanto, foram classificados como obrigatórios.

Em relação aos movimentos não obrigatórios: *orientados aos ouvintes e fechamento*; Miranda J. (2018) afirmou não ter encontrado nenhuma ocorrência destes em suas análises. O autor sugere que esta diferença poderia estar associada à temática das TED Talks selecionadas. No estudo de Chang e Huang (2015) foram investigadas TED Talks associadas a cinco diferentes áreas. Miranda J. (2018), por sua vez, focou apenas em palestras retiradas do tópico ciência. No entanto, acredito que as divergências reportadas por Chang e Huang (2015) e Miranda J. (2018) se devam à diferença no tamanho das amostras investigadas, e não às temáticas das apresentações, uma vez que o *corpus* de estudo utilizado por Chang e Huang (2015) foi compilado a partir de 58 palestras, já o *corpus* compilado por Miranda J. (2018) continha apenas 10 vídeos.

Em outro estudo, Liu e Chen, H. (2020b) comparam seis aspectos comunicativos típicos das TED Talks em relação ao gênero palestras acadêmicas. Evidentemente, ambos os gêneros são registros orais da língua e possuem certas características similares, tais como serem predominantemente monólogos, e apresentam conteúdos pré-planejados. Contudo, os autores observaram seis diferenças-chave entre esses gêneros, são elas:

- **Plateia:** Nas TED Talks, o nível de conhecimento e o interesse podem variar muito de um indivíduo para o outro. Em palestras acadêmicas, os ouvintes possuem graus de conhecimentos e interesses similares pelo assunto.
- **Relações de poder:** Segundo Liu e Chen, H. (2020b), as TED Talks seriam mais igualitárias, uma vez que palestrantes e plateia são, em geral, estranhos entre si. Ademais, assim como ocorre em palestras acadêmicas, durante a produção desse gênero, a relação de poder é assimétrica, pois é o palestrante/professor quem possui o poder nas tomadas de decisões.
- **Interatividade:** Durante as TED Talks, é comum haver mínima ou nenhuma interação entre o palestrante e a plateia. Já em palestras acadêmicas, é comum haver um momento dedicado à interação entre professores e alunos, assim como para discussões e solução de dúvidas.
- **Planejamento:** Visto que o tempo médio das TED Talks são 18 minutos, é esperado que elas sejam cuidadosamente planejadas e ensaiadas inúmeras vezes, Liu e Chen, H. (2020b) argumentam que elas podem ser até mesmo memorizadas. Em palestras acadêmicas, é também esperado que o palestrante/professor planeje as aulas com antecedência, no entanto, os autores argumentam que elas seriam mais difíceis de

serem memorizadas completamente dado ao maior tempo de duração, em geral, 50 minutos. Independentemente da possibilidade de serem ou não memorizadas, em ambos os gêneros é comum o uso de recursos que auxiliem no planejamento, tais como: anotações e slides com listas de ponto e marcadores de tempo.

- **Propósito e tópico:** As TED Talks, assim como outras modalidades de palestras voltadas à popularização das ciências, visam compartilhar novas ideias com foco na relevância social. Por outro lado, Liu e Chen, H. (2020b) explicam que palestras acadêmicas possuem natureza instrucional, logo, o foco está no desenvolvimento do conhecimento profissional e acadêmico para um público específico. Em outras palavras, as discussões das TED Talks são mais gerais e acessíveis ao público leigo, ao passo que palestras acadêmicas são mais especializadas para certas disciplinas e profissões.

Pelo exposto na presente seção, a análise dos aspectos situacionais das TED Talks possibilita uma compreensão mais aprofundada não apenas das características discursivas desse gênero, mas também do contexto social em que estão inseridas, do tipo de público que atraem e recriam, e das discussões que elas promovem (*e.g.* ativismo social, ambiental, educação online, empreendedorismo, entre outros). Portanto, as TED Talks emergem como um dos principais expoentes na popularização da ciência no cenário contemporâneo. Nesse sentido, concordo com Hyland (2009) ao afirmar que os gêneros voltados à popularização da ciência impulsionam o desenvolvimento de novas práticas científicas em nossa sociedade, bem como auxiliam na redução das barreiras de acesso ao conhecimento para o público leigo.

2.2 Linguística de *corpus*: concepções centrais, tipos de estudos e estatuto do campo

Em termos mais abrangentes, um *corpus* (ou *corpora*, no plural) consiste em uma coletânea de textos compilados em formato eletrônico, baseando-se em critérios específicos e servindo como objeto, instrumento, ou método para pesquisadores interessados em investigar o funcionamento de uma determinada língua (Berber Sardinha, 2004; McEnery; Wilson, 2001; Sinclair, 2005); Viana (2011, p. 27) sintetiza as múltiplas dimensões de um *corpus* de estudo da seguinte maneira:

- Deve ser compreendido como um conjunto de textos;
- Contempla textos (orais e escritos) que tenham sido efetivamente produzidos por falantes de determinada língua;

- Consiste numa forma de representar empiricamente o uso que se faz de uma língua em seu sentido geral ou específico;
- É uma reprodução da produção linguística de toda a população que se quer investigar ou uma amostra representativa dessa população, com base em princípios claros e bem definidos;
- Assume a forma eletrônica com vistas a ser investigado pelo computador;
- É concebido com o objetivo de possibilitar a realização de uma pesquisa linguística.

Textos são os elementos centrais nos estudos com *corpora*, logo, a importância dos critérios de seleção textual e o processo de documentação dos metadados devem ser sempre enfatizados, pois esses passos asseguram coerência entre os dados coletados e os objetivos do estudo. No entanto, exceto em contextos especializados⁸, *corpora* não são capazes de reproduzir todas as produções linguísticas de uma população. Portanto, qualquer pesquisador interessado em compilar um *corpus* deve adotar um conjunto de critérios para garantir que a amostra investigada seja a mais representativa possível de uma população específica. Na literatura acadêmica, há uma quantidade expressiva de trabalhos descrevendo como devem ser realizadas as compilações de *corpora* de estudo (cf. Berber Sardinha, 2004; Hunston, 2002; McEnery; Hardie, 2012; Sinclair, 2005), dentre os mais recentes, Viana e O'Boyle (2022) apresentam uma lista⁹ com 25 parâmetros a serem considerados como guias na compilação de *corpora* de estudo em contextos de ensino para fins acadêmicos. Entretanto, os autores ressaltam que a aplicação de *corpora* nos estudos linguísticos é extremamente diversa, cabendo, portanto, a cada pesquisador decidir quais critérios serão mais adequados para compilação de um *corpus* de estudo coerente com os objetivos de sua pesquisa.

O termo linguística de *corpus* ganhou popularidade apenas na década de 80 e é comumente atribuído ao linguista Jan M. G. Aarts (Johansson, 2008; Leech, 1992). A LC destaca-se por priorizar a observação sistemática de dados empíricos, variações e padrões linguísticos. Nesse sentido, a premissa básica para os estudos no campo da LC é o conceito de língua como um sistema probabilístico, no qual as escolhas linguísticas não ocorrem de forma

⁸ Casos em que o *corpus* representa a totalidade do uso linguístico (obras de um autor já falecido; coleções de livros, etc.) podem ser vistos como representativos, uma vez que o conjunto de dados é finito, portanto, passível de ser coletado na íntegra.

⁹ Para maiores informações vide Viana e O'Boyle (2022), Capítulo 5 – *Specialized corpus compilation for EAP*.

aleatória, mas são regulares a certos contextos situacionais (Biber; Reppen, 2015; Leech, 1992, 2011; McEnery; Hardie, 2012; Teubert, 2005; Viana, 2011). Conforme Teubert (2005), a LC opera no âmbito social da comunicação, e seu principal foco é o significado daquilo que é verbalmente comunicado entre os membros de uma comunidade linguística. Em outras palavras, a LC possui natureza empírica, e seu objeto de estudo é a língua compreendida como um fenômeno social, observável e acessível por meio de registros textuais armazenados em formato eletrônico.

No que tange ao uso de ferramentas computacionais aplicadas à investigação de *corpora* linguísticos, Leech (2011) argumenta que o surgimento de *corpora* eletrônicos e outras ferramentas computacionais de processamento de texto tiveram um impacto na Linguística análogo aos efeitos que o Telescópio Hubble causou (e ainda causa) na Física e na Astronomia. Segundo o autor:

[*corpora*] eletrônicos e suas ferramentas têm revolucionado a ciência linguística, ao nos capacitar a fazer coisas com dados linguísticos que ninguém poderia ter sonhado cinquenta anos atrás, e abrindo o potencial para um novo empirismo na linguística¹⁰ (Leech, 2011, p. 158).

É inquestionável que o acesso a *corpora* eletrônicos e o desenvolvimento de novas ferramentas computacionais possibilitaram aos pesquisadores investigarem padrões de usos até então ignorados por descrições gramaticais “tradicionais” da língua. Em especial, projeto COBUILD¹¹, iniciado em 1987, foi um dos marcos que possibilitou as investigações de natureza empírica voltarem a ganhar espaço dentro dos estudos linguísticos. Atualmente, a LC é um campo bastante ativos da Linguística, influenciando diversas áreas de estudos, como análise do discurso, lexicografia; ensino e aprendizagem de línguas e linguística computacional, para mencionar algumas.

Ainda no tocante à LC, um tópico que suscita debate entre pesquisadores da área é o status da LC no campo dos estudos linguísticos. De um lado, há aqueles que a concebem como

¹⁰ No original: “*electronic corpus and its associated toolkit have been revolutionizing linguistic science, by empowering us to do things with linguistic data that no one could have dreamed of fifty years ago, opening up the potential for a new empiricism in linguistics.*”

¹¹ Collins Birmingham University International Language Database.

um método de análise; de outro, estão aqueles que argumentam em favor da LC como algo mais complexo, com base teórica e objetos de investigação próprios.

Diversas publicações já abordaram esse tópico, e apresentam visões nem sempre convergentes (McEnery; Hardie, 2012; Shepherd, 2009; Tognini-Bonelli, 2001). Vale destacar aqui a coletânea de entrevistas organizadas por Viana, Zyngier e Barnbrook (2011), na qual foram entrevistados 14 pesquisadores com trabalhos influentes em diversos campos da Linguística. Em especial, as respostas apresentadas à pergunta: “A linguística de *corpus* é uma ciência ou uma metodologia? Onde você situaria a linguística de *corpus*, no panorama científico ou metodológico?¹²”, evidenciam a complexidade dessa discussão.

Em síntese, muitas das respostas apresentadas tendem a convergirem para os seguintes posicionamentos:

- A LC pode ser compreendida tanto como uma ciência, quanto uma metodologia, a depender do modo como ela é aplicada em um dado estudo (cf. Aston (2011); Baker (2011) e Berber Sardinha (2011)).
- A LC é primeiramente uma metodologia que lida com a coleta e organização de dados linguísticos empíricos, que, posteriormente, são investigados sob a perspectiva de outras abordagens teóricas (cf. Davies (2011); Hyland (2011b) e Leech (2011)).
- A LC entendida como uma ferramenta computacional que possibilita análises qualiquantitativas de grandes quantidades de dados linguísticos (cf. Scott (2011) e Swales (2011)).

Observando as caracterizações apresentadas nos excertos acima, percebe-se a dificuldade em delimitar de forma clara como a LC é enquadrada nos estudos linguísticos. Todavia, o posicionamento do pesquisador acerca de como a LC é conceituada em seu estudo é de extrema relevância, uma vez que ela que orientará o modo como serão conduzidas as análises e interpretações dos dados encontrados em um *corpus*.

Na literatura, o estudo da língua por meio da análise de *corpora* é frequentemente descrito, de maneira dicotômica, como estudos dirigidos por *corpus* (*corpus-driven*) e baseados em *corpus* (*corpus-based*). Tognini-Bonelli (2001) foi a primeira a propor o uso desses termos para diferenciar as principais abordagens em que *corpora* podem ser

¹² No original: “Is Corpus Linguistics a science or a methodology? Where would you situate Corpus Linguistics in the scientific or methodological panorama?”.

explorados em estudos com foco na análise e descrição linguística. Nos estudos dirigidos por *corpora*, as evidências encontradas no *corpus* são o ponto inicial de investigação, isto é, os dados se tornam o ponto de partida para teorizações e generalizações acerca do funcionamento da língua. Tognini-Bonelli (2001, p. 84), explica que:

em uma abordagem dirigida por *corpus*, o compromisso do linguista é com a integridade dos dados como um todo, e as descrições visam ser abrangentes no que diz respeito às evidências do *corpus*. O *corpus*, portanto, é visto como mais do que um repositório de exemplos para apoiar teorias pré-existentes, [...] as declarações teóricas são totalmente consistentes e refletem diretamente as evidências fornecidas pelo *corpus*.¹³

Nessa abordagem, o *status teórico* da LC é caracterizado por um ser um estudo de natureza indutiva, no qual o pesquisador é instigando a fazer o *mínimo de suposições* acerca do funcionamento da língua antes do levantamento de dados empíricos. Logo, os padrões linguísticos mais recorrentes em um *corpus* devem ser a base para a elaboração de uma descrição linguística que reflita e explique tais evidências.

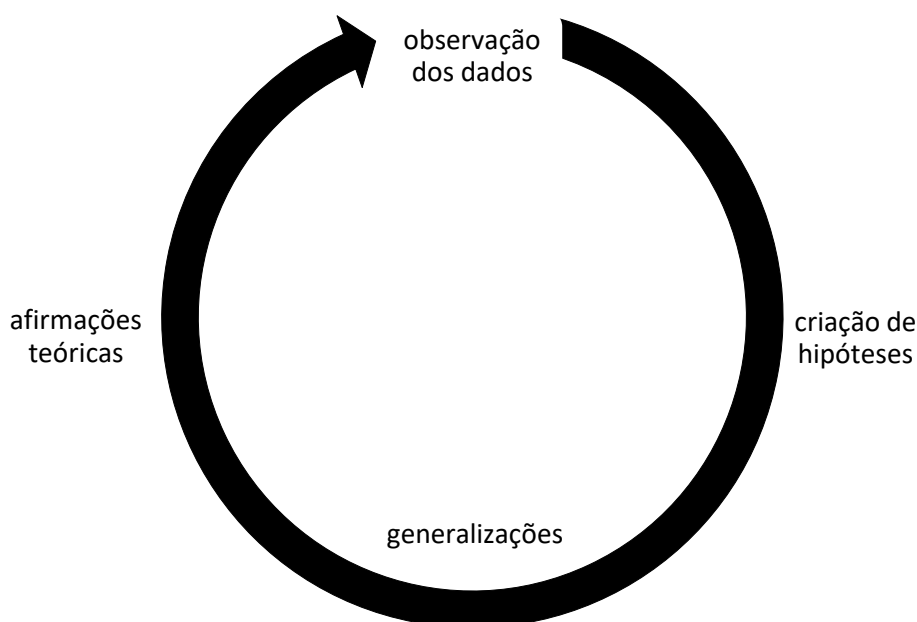
Entretanto, é importante ressaltar que os conceitos de *status teórico* e *mínimo de suposições teóricas* devem ser sempre utilizados com cautela. Primeiramente, “status teórico não é o mesmo que teoria¹⁴” (McEnery; Xiao; Tono, 2006, p. 7). Ademais, conceber um linguista que fosse capaz de explorar um *corpus* sem utilizar algum tipo de conhecimento prévio sobre a língua é algo, no mínimo, ilusório. Tognini-Bonelli (2001, p. 85) reconhece essa contradição ao afirmar que “não existem abordagens puramente indutivas”¹⁵. Contudo, a autora argumenta em favor de estudos nos quais as afirmações teóricas são apresentadas somente após uma análise sistemática das evidências empíricas encontradas no *corpus*. O caminho metodológico a ser seguido em estudos dirigidos por *corpus* está ilustrado na Figura 2.

¹³ No original: “In a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories, [...] the theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus”.

¹⁴ No original: “theoretical status is not theory in itself”.

¹⁵ No original: “[...] there is no such thing as pure induction”.

Figura 2 – Caminho metodológico em estudos dirigidos por *corpus*



Fonte: Traduzido e adaptado de Tognini-Bonelli (2001, p. 85)

Tognini-Bonelli (2001) explica que o caminho apresentado acima não é rígido e nem mecânico, mas sim constantemente mediado pelo pesquisador em todos os estágios do processo. Além disso, é necessário reconhecer que, ao se afastarem (o máximo possível) de teorias já estabelecidas e observarem os dados sob novas perspectivas, as pesquisas dirigidas por *corpora* têm maiores chances de encontrarem padrões linguísticos até então ignorados ou não descritos na literatura. Vale lembrar que os conceitos de colocação, coligação, prosódia semântica, além dos famosos princípios da livre escolha e princípio idiomático propostos por Sinclair (1991) são todos frutos de estudos diretamente ligados à abordagem dirigida por *corpora*.

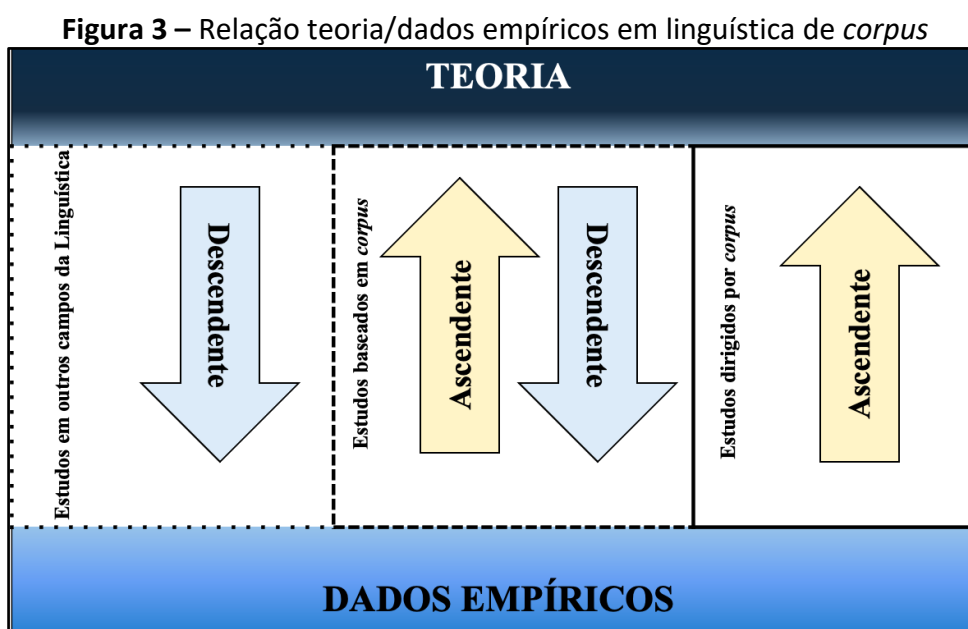
No que se refere aos estudos baseados em *corpus*, eles se caracterizam por serem investigações nas quais o *corpus* é empregado em conjunto com constructos e descrições linguísticas previamente formuladas. Segundo Tognini-Bonelli (2001, p. 65):

[a abordagem baseada em *corpus*] é usada para se referir a uma metodologia que se vale do *corpus* principalmente para expor, testar ou exemplificar teorias e descrições que foram formuladas antes que grandes *corpora* se tornassem disponíveis para informar estudo da linguagem¹⁶.

¹⁶ No original: “*corpus-based [approach] is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study*”.

Observando a definição acima, é perceptível que a autora estabelece um paralelo entre a concepção de LC como um método e a abordagem baseada em *corpus*. Em ambos os casos, é reconhecida a possibilidade de analisar o conteúdo de *corpus* sob a ótica de modelos descritivos preestabelecidos que os pesquisadores acreditem serem adequados para explicarem o funcionamento da língua.

Diante da impossibilidade de separar completamente os fatores “teoria” e “dados empíricos” nas análises de *corpora*, Viana (2012) propõe que a maneira como um *corpus* é aplicado em um estudo pode ser mais claramente evidenciada ao adicionarmos as dimensões de: direcionalidade e composição; conforme é ilustrado na Figura 3.



Fonte: Viana (2012, p. 76)

Viana (2012) esclarece que a direcionalidade (representada por setas) diz respeito ao ponto de origem e à orientação adotada pelo pesquisador: teoria → dados, ou vice-versa, enquanto composição do estudo (representada por retângulos) descreve a movimentação gradual entre os polos teoria e dados empíricos, que pode ocorrer de maneira mais fluída ou rígida a depender de como esses elementos se relacionam. Com a adição desses novos fatores, o autor então propõe uma taxonomia descritiva contendo três possíveis categorias de estudos.

Primeiramente, na extremidade esquerda, estão os estudos conduzidos em outros campos da linguística. Pesquisas dessa natureza priorizam seus constructos teóricos (teoria → dados) e recorrem ao *corpus* apenas na busca por dados que reafirmem suas teorias/hipóteses. Em outras palavras, o *corpus* é somente um repositório de exemplos do uso da língua. Viana (2012) explica que as bordas com linhas tracejadas indicam que o pesquisador consegue se movimentar entre os polos de forma bem fluida.

Entre os extremos, estão os estudos baseados em *corpus* (dados ↔ teoria). Viana (2012) argumenta que estudos baseados em *corpus* possuem natureza híbrida e permitem ao pesquisador iniciar sua investigação tanto a partir de conhecimentos já estabelecidos (natureza descendente) quanto a partir da exploração do *corpus* (natureza ascendente). Para o autor, investigações por este prisma abrem caminho tanto para novas interpretações acerca de modelos teóricos já estabelecidos, quanto para a elaboração de novas descrições.

Por fim, na extremidade à direita, encontram-se os estudos dirigidos por *corpus* (dados → teoria). De acordo com Viana (2012), estudos nessa linha buscam propor explicações com base nos (e apenas para) os dados encontrados em *corpora*. As bordas em linhas contínuas indicam movimentação limitada entre os polos do espectro, representando assim um quadro analítico objetivo, e não enviesado por constructos de outros campos.

Com base nas discussões apresentas acerca do status da LC da taxonomia proposta por Viana (2012), compreendo que a presente pesquisa pode ser caracterizada como baseada em *corpus*. Isso se deve ao fato de os pacotes lexicais serem identificados a partir da exploração dos *corpora* de estudo STEM e SHAPE (dados → teoria) e, posteriormente, analisados com o suporte de explicações descritivas e funcionais já estabelecidas nos estudos linguísticos (teoria → dados). Entretanto, em caso de conflitos, as evidências provenientes dos dados empíricos serão priorizadas em detrimento das prescrições teóricas. Em outras palavras, a observação de pacotes lexicais não previstos/identificados em outros estudos poderá levar a descrições de novos padrões estruturais e/ou funcionais da língua.

Antes de finalizar, é importante ressaltar que o objetivo desta seção não foi atribuir valores ou oferecer respostas definitivas em favor de alguma concepção, mas sim discutir de maneira abrangente algumas das diversas concepções da LC presentes na literatura. Independentemente da forma como a LC é categorizada, é inegável que o acesso a grandes quantidades de dados linguísticos tem permitido que pesquisadores explorem, de maneira mais objetiva, a natureza formulaica da língua.

2.3 Linguagem formulaica: conceitos e contribuições para os estudos linguísticos

O advento de *corpora* eletrônicos e os avanços tecnológicos de ferramentas computacionais possibilitaram a descoberta de evidências robustas de que certas combinações linguísticas são tão recorrentes no uso de uma língua que podem ser classificadas como estruturas pré-fabricadas. A recorrência de tais expressões é elucidada por Sinclair (1991) por meio do princípio idiomático, no qual é proposto que grande parte da produção linguística de um falante provém de um repertório de frases pré-construídas e armazenadas em seu léxico mental. Segundo o autor:

o princípio idiomático consiste em o usuário da língua dispor de um vasto número de frases semiconstruídas que constituem escolhas únicas, ainda que possam ser analisadas em segmentos. De certa forma, isso pode refletir a recorrência de situações similares em assuntos humanos, ou pode ilustrar uma tendência natural para a economia de esforços, ou pode ainda ser motivada em parte pelas exigências da comunicação em tempo real (Sinclair, 1991, p. 110).¹⁷

Em outras palavras, mesmo diante de infinitas possibilidades de combinações lexicais, o falante tende a utilizar rotineiramente certas expressões, em vez de criar novas combinações lexicais. Na presente tese, adoto o termo linguagem formulaica para me referir aos diversos tipos de combinações pré-moldadas da língua¹⁸. Na literatura, este termo é frequentemente atribuído à Wray (2002), que desenvolveu e discutiu critérios necessários para que uma combinação de palavras pudesse ou não ser considerada formulaica, ou seja, uma sequência armazenada em nossa memória e processada pelo cérebro de maneira holística. Wray (2002, p. 9) define linguagem formulaica da seguinte forma:

¹⁷ No original: *“the principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments. To some extent, this may reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort; or it may be motivated in part by the exigencies of real-time conversation”*.

¹⁸ Quando for necessário abordar características particulares, utilizarei termos específicos, tais como: expressões idiomáticas, colocações, pacotes lexicais, *phrasal verbs*, etc.

uma sequência, contínua ou não, de palavras ou outros elementos, que é ou que aparenta ser pré-fabricada: ou seja, armazenada e recobrada da memória em sua totalidade quando utilizada, ao invés de estar sujeita à produção, ou à análise gramatical da língua.¹⁹

A premissa apresentada por Wray (2002) de que o cérebro humano otimiza o processamento de grandes quantidades de dados linguísticos por meio do uso contínuo de estruturas convencionalizadas na língua está fortemente alinhada ao princípio idiomático de Sinclair (1991) e à tendência natural de falantes de uma língua adotarem formas rotineiras (pré-moldadas) de se expressar na mesma, em detrimento de novas combinações lexicogramaticais. Além disso, é possível identificar ao menos dois benefícios inerentes ao uso de formas pré-moldadas na língua: *i)* assegurar fluidez na comunicação e *ii)* evitar que os falantes fiquem sobrecarregados ao decodificar frases nunca ouvidas antes. Devido a essa dupla vantagem, o uso adequado da linguagem formulaica é um dos pontos centrais na investigação de qualquer língua (Schmitt; Carter, 2004; Wray, 2002; Wray; Perkins, 2000).

Apesar da terminologia utilizada para descrever sequências de multipalavras ainda ser uma “zona cinzenta” na literatura, Conrad e Biber (2004) explicam que, independentemente do nome adotado, há algumas características que tendem a ser especialmente recorrentes na identificação da linguagem formulaica, tais como:

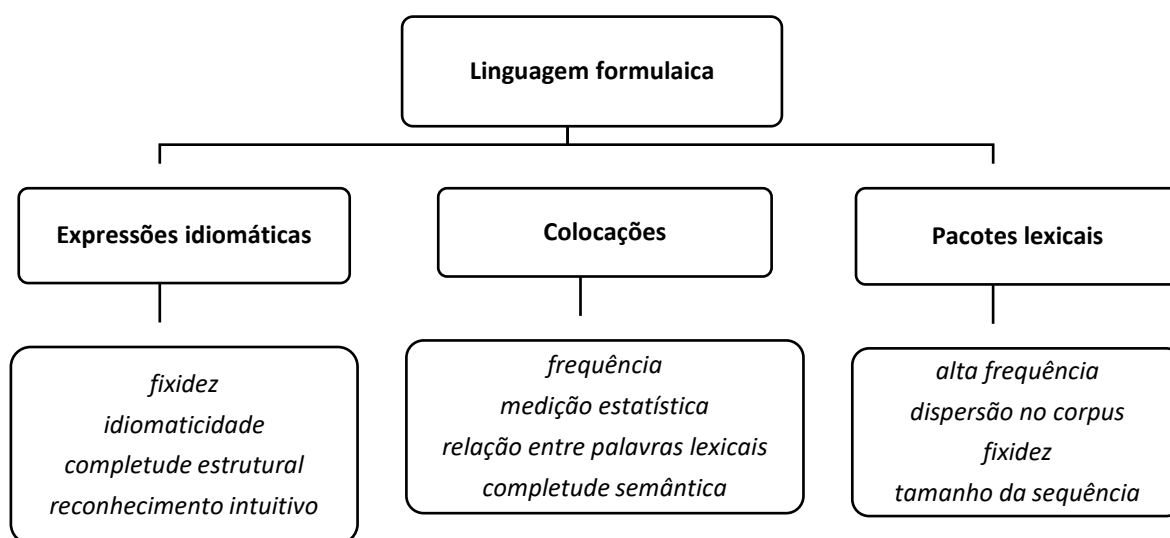
- **Grau de fixidez** – tendência que uma sequência de palavras apresenta em aparecer de forma totalmente fixa (*e.g. ladies and gentlemen, once in a blue moon*) ou com alguma variação interna. A estrutura semifixa *the * of the*, por exemplo, possui um único *slot* variável que possibilita a produção de diversas sequências como: *the end of the*, *the importance of the* e *the beginning of the* (Reppen; Olson, 2020).
- **Idiomaticidade** – o *continuum* de transparência semântica que uma sequência pode apresentar. Em um extremo, estão aquelas completamente transparentes, na qual o significado do todo é idêntico aos componentes individuais (*e.g. I want you to*). No outro extremo, estão sequências semanticamente opacas: provérbios e expressões idiomáticas, por exemplo, são combinações de palavras na qual o significado não está relacionado aos componentes individuais da sequência (*e.g. blood is thicker than water*).

¹⁹ No original: “a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar”.

- **Frequência** – sequências semanticamente transparentes são extremamente frequentes em um *corpus*; em oposição, expressões idiomáticas tendem a serem mais raras.
- **Tamanho da sequência** – o tamanho da sequência que se deseja estudar tem relação direta com a sua frequência em um *corpus*. Quanto maior for a quantidade de elementos em uma sequência, menor será o número de ocorrências que ela apresentará em um *corpus*.
- **Compleitude estrutural e semântica** – colocações e expressões idiomáticas são sequências de palavras que operam como um item lexical completo, expressando um único significado dentro do discurso. Já os pacotes lexicais, por exemplo, representam fragmentos de sintagmas e orações de um texto.
- **Reconhecimento intuitivo** – Conrad e Biber (2004) explicam que expressões idiomáticas como: *birds of a feather flock together* e *slap in the face*, tendem a ser compreendidas sem dificuldade por falantes de inglês como primeira língua, além de serem processadas de forma holística como uma unidade lexical.

Conrad e Biber (2004) também explicam que diferentes tipos de linguagem formulaica são identificados dependendo da prioridade que tais características recebem. Em outras palavras, caso o escopo do estudo seja expressões idiomáticas, é esperado que o pesquisador priorize determinadas características que não seriam interessantes para identificar colocações ou pacotes lexicais. A Figura 4 apresenta algumas das características essenciais na identificação de diferentes tipos de linguagem formulaica de acordo com os autores.

Figura 4 – Diferentes prioridades na identificação de linguagem formulaica



Fonte: Adaptado de Conrad e Biber (2004, p. 57-58).

Além das propriedades a serem consideradas, há, na literatura, três tipos de abordagens comumente utilizadas na identificação de linguagem formulaica: fraseológica, psicolinguística e baseada em frequência. Granger e Paquot (2008) e Durrant e Mathews-Aydinli (2011) descrevem os pontos mais relevantes de tais abordagens da seguinte forma:

- **Abordagem fraseológica:** de natureza descendente (*top-down*) em que o foco está na não composicionalidade de certas expressões, ou seja, o significado do todo não equivale ao significado das partes individuais. Nesse paradigma, as combinações de palavras são situadas em um *continuum* que engloba combinações livres (*e.g. blow a trumpet*), colocações (*e.g. blow a fuse*), linguagem figurada (*e.g. blow your own trumpet*) e expressões idiomáticas (*e.g. blow the gaff*).
- **Abordagem psicolinguística:** foco na eficácia do processamento e no armazenamento do léxico na mente dos falantes. As sequências de palavras são recuperadas da memória e processadas de forma holística, ao invés de serem construídas individualmente em cada momento de uso da língua. O conceito de *priming* lexical proposto por Hoey (2005) argumenta que certas combinações de palavras são ativadas em nosso léxico mental com maior ou menor velocidade a depender do co-texto em que são utilizadas, sendo assim, o autor explica que a combinação *body* e *heart*, por exemplo, seria ativada em nosso léxico mental com maior velocidade do que *trick* e *heart*, dada a relação que temos delas no mundo real.
- **Abordagem baseada em frequência:** foco na tendência estatística que certos padrões linguísticos possuem de ocorrerem em um *corpus*. Essa abordagem busca encontrar diversos tipos de combinações frequentes, tais como: colocações (*e.g. hard work, conduct an experiment*), pacotes lexicais (*e.g. as can be seen, if you think about*) e prosódia semântica (*e.g. cause* + conceito negativo = *cause* + *accident* / *damage* / *disease*).

No tocante à abordagem baseada em frequência, Ädel e Erman (2012) sustentam que o uso adequado das combinações mais recorrentes e típicas de um determinado registro é uma das maneiras pela qual a formulaicidade se manifesta em um texto. De fato, vários estudos argumentam que a natureza empírica da LC tem oferecido ótimos resultados na investigação de linguagem formulaica. Altenberg (1998), por exemplo, revelou que 80% das palavras do *London Lund Corpus* estavam presentes em sequências formulaicas. De maneira similar, Erman e Warren (2000) relataram que as sequências formulaicas ocupavam entre 52 e 57% dos textos compilados em seu *corpus* de estudo.

No que diz respeito à identificação de sequências recorrentes em um *corpus*, alguns autores argumentam em favor da aplicação de medidas de associação como a informação mútua²⁰ (cf. Ellis; Simpson-Vlach; Maynard, 2008; Salazar, 2014; Simpson-Vlach; Ellis, 2010) ou chavicidade²¹ (Bocorny et al., 2021), como forma de distinguir padrões linguísticos com maior relevância pedagógica (e.g. *on the other hand*), daqueles menos “interessantes” ao contexto do ensino de língua inglesa (e.g. *is one of the*).

Especialmente em relação à associação de informação mútua, Biber (2009) apresenta dois potenciais problemas metodológicos ao incorporar tal medida na identificação de pacotes lexicais. Em primeiro lugar, o cálculo de informação mútua foi desenvolvido para medir a atração colocacional entre duas palavras lexicais, independente da direcionalidade da coocorrência, portanto, utilizar esta medida em sequências maiores e fixas de palavras pode gerar resultados passíveis a interpretações conflitantes. Em segundo lugar, essa medida tende a desconsiderar combinações de ordem gramatical em favor de combinações lexicais, ou seja, certas combinações de palavras seriam mais formulaicas, ao passo que outras seriam entendidas como um produto gramatical da língua. Nesse sentido, a medida de informação mútua se torna problemática quando aplicada às sequências maiores e mais frequentes de palavras, pois estas, em geral, incorporam tanto palavras lexicais quanto gramaticais, além disso, Biber (2009, p. 290) explica que “atração colocacional não equivale a status formulaico²²”, logo, a aplicação da medida da informação mútua falharia em reconhecer sequências do tipo *if you look at* e *as can be seen* como sendo formulaicas. Novamente, tal resultado pode ser conflitante uma vez que a natureza formulaica de tais sequências já fora validada em outros estudos (cf. Biber; Conrad; Cortes, 2004; Hyland, 2008a).

Conforme ilustrado na Figura 4, os pacotes lexicais são um tipo de linguagem formulaica evidenciada principalmente através de critérios de frequência e dispersão. Logo, investigações desse objeto de estudo estão alinhadas à abordagem baseada em frequência, visto que ela possibilita a extração semiautomática de grandes volumes de dados linguísticos em um *corpus* e permite ao pesquisador estabelecer critérios quantitativos que melhor se

²⁰ “Medida de significância que mede a força de associação entre as palavras e privilegia palavras menos frequentes” (Tagnin, 2011, p. 360).

²¹ A chavicidade (*keyness*) é uma medida estatística que indica o quanto a ocorrência uma palavra (ou sequência de palavras) é mais significativa em um *corpus* de estudo quando comparada a um *corpus* de referência.

²² No original: “*collocational strength is not equivalent to formulaic status.*”

adéquam às necessidades de cada estudo. Desta forma, busco, na seção subsequente, discutir de maneira mais aprofundada o que são pacotes lexicais e o que investigações anteriores desse tipo de linguagem formulaica revelaram a respeito do discurso em diversos registros acadêmicos.

2.4 Pacotes lexicais

Conforme abordado previamente, certas combinações linguísticas são tão recorrentes na língua que nossos cérebros as processam de forma holística, isto é, como se fossem uma única palavra (Sinclair, 1991; Wray, 2002). Entre os vários tipos de padrões linguísticos passíveis de análise em um *corpus*, a presente tese investiga os pacotes lexicais, uma categoria de sequências formulaicas apresentadas pela primeira vez por Biber et al. (1999, p. 989) como “sequências de palavras mais recorrentes em um registro²³” e evidenciadas por meio de análises distribucionais típicas dos estudos dirigidos por *corpus*. Em especial, os critérios empregados são frequência relativa (por milhão de palavras no *corpus*) e dispersão (quantidade de textos que a sequência ocorre no *corpus*).

Na literatura, diversas pesquisas abordam o uso de pacotes lexicais para contrastar diferentes registros orais e escritos, particularmente aqueles de natureza acadêmica (Ädel; Erman, 2012; Biber, 2006a; Biber; Barbieri, 2007; Conrad; Biber, 2004; Wang, 2017), ou com foco na variação disciplinar (Cortes, 2004; 2006; Durrant, 2015; Hyland, 2008a; Hyland; Jiang, 2018; Reppen; Olson, 2020). No âmbito das TED Talks, exemplos comuns de pacotes lexicais incluem: *thank you very much, for the first time, all over the world* e *I’m going to show you* (Liu; Chen, H. 2020b). Sequências deste tipo são facilmente reconhecidas pelo falante de uma língua graças ao alto grau de recorrência e a transparência semântica de seus componentes.

Nesse contexto, destaca-se que a escolha de investigar pacotes lexicais em TED Talks ocorreu com base no entendimento de que eles não apenas contribuem para a fluência textual e reduzem o tempo de processamento linguístico, mas também podem revelar particularidades no modo como diferentes grupos disciplinares de TED Talks empregam essas estruturas.

²³ No original: “lexical bundles are the sequences of words that most commonly co-occur in a register.”

Também é possível observar que as investigações de pacotes lexicais tendem a convergir para objetivos de natureza bastante similar, conforme evidenciado por Conrad e Biber (2004, p. 58):

Nossas perguntas de pesquisa em relação a esta abordagem são exploratórias. Perguntamos se existem sequências de multipalavras empregadas com alta frequência nos textos, se diferentes registros tendem a usar diferentes tipos de sequências e, se sim, com que extensão os pacotes lexicais desempenham funções discursivas e se apresentam como parte importante do repertório comunicativo de falantes e escritores.²⁴

Além de apresentarem os principais objetivos de estudo com interesses na investigação de pacotes lexicais, Conrad e Biber (2004) destacam o caminho metodológico que deve ser adotado na identificação de tais sequências. Segundo os autores, os pacotes lexicais são investigados por uma abordagem de natureza ascendente, dirigida por *corpus*, e comumente apresentando os seguintes componentes:

- **Identificação por meio de ferramentas de análise de *corpora*.**

Considerando a natureza empírica dos pacotes lexicais, é imprescindível a utilização de ferramentas computacionais que identifiquem as sequências mais recorrentes em um *corpus* de forma automática. Atualmente, há uma variedade considerável de programas e plataformas computacionais para essa função, tais como: #LancsBox® (Brezina; Weill-Tessier; McEnery, 2020); AntConc® (Anthony, 2020a); AntGram® (Anthony, 2020b); Sketch Engine® (Kilgariff *et al.*, 2004) e WordSmith Tools (Scott, 2022), para mencionar alguns.

Após processarem o *corpus*, as ferramentas mencionadas acima geram uma lista de sequências de palavras conforme as configurações de tamanho, frequência e dispersão previamente estabelecidas pelo pesquisador. Na presente tese, a ferramenta de escolha foi o AntConc® (Anthony, 2020a), que será descrita com maiores detalhes no Capítulo 3, Seção 3.4.

²⁴ No original: “Our research questions in this approach are exploratory. We ask whether there are multi-word sequences that are used with high frequency in texts, whether different registers tend to use different sets of these sequences, and, if so, to what extent the bundles fulfill discourse functions and thus play an important part in the communicative repertoire of speakers and writers.”

- **Estabelecimento de critérios quantitativos (frequência e dispersão).**

Para que uma sequência de palavras seja considerada como pacote lexical, é necessário que ela atenda ao número mínimo de ocorrências e de dispersão estabelecido pelo pesquisador. Entretanto, critérios de frequência mínima tendem a serem arbitrários e decididos de acordo com o escopo de cada pesquisa. Biber *et al.* (1999), por exemplo, foram mais inclusivos e optaram por adotar um corte de dez ocorrências por um milhão de palavras; Hyland (2008a) e Biber, Conrad e Cortes (2004), por sua vez, foram mais conservadores e estabeleceram cortes mais altos de 20 e 40 ocorrências por um milhão de palavras, respectivamente.

Ainda em relação à frequência, outro aspecto comum em pesquisas com pacotes lexicais é a preferência por *corpora* de estudo com mais de 1 milhão de palavras. Diversos autores concordam que essa convenção atribui maior confiabilidade aos dados, uma vez que a normalização de frequência em *corpora* pequenos pode inflar os números, fazendo com que sequências pouco recorrentes sejam erroneamente reconhecidas como pacotes lexicais pelos programas computacionais (Bestgen, 2020; Biber; Barbieri, 2007; Cortes, 2015). Contudo, caso não seja possível a compilação de grandes *corpora* de estudo, uma solução proposta por Bestgen (2020) seria o aumento no corte mínimo de ocorrências, de modo a assegurar que somente as sequências mais frequentes do *corpus* seriam selecionadas.

Paralelamente à frequência, é necessário também que uma mesma sequência esteja distribuída em uma quantidade mínima de textos a fim de que seja evitada idiossincrasia. A dispersão mínima que um pacote lexical precisa apresentar em um *corpus* também parece ser arbitrária. Em geral, são utilizados números absolutos para determinar o corte mínimo de dispersão. Por exemplo, Biber *et al.* (1999) e Biber, Conrad e Cortes (2004) optaram por um corte mínimo de cinco textos, independentemente do tamanho do *corpus*. No estudo de Biber, Conrad e Cortes (2004), mesmo os pacotes lexicais menos frequentes em seu *corpus* de estudo apareceram em ao menos 20 textos distintos. Logo, um número arbitrário como cinco textos teve pouca influência na extração dos dados. De modo a contornar esse problema, alguns autores optaram por utilizar uma taxa de dispersão relativa. Hyland (2008a) e Hyland e Jiang (2018), por exemplo, adotaram um ponto de corte de dispersão de 10% dos textos no *corpus*, e Liu e Chen, H. (2020b) consideraram que um pacote lexical deveria ocorrer em um mínimo de 5% dos textos no *corpus*.

Os critérios de frequência e dispersão adotados para a investigação de pacotes lexicais no presente estudo estão descritos no Capítulo 3, Seção 3.5.

- **Descrição de seus componentes internos.**

A composição interna dos pacotes lexicais é bastante distinta de outros tipos de sequências formulaicas, como colocações e expressões idiomáticas. Em primeiro lugar, seus componentes são semanticamente transparentes, ou seja, o significado da sequência está vinculado ao significado de seus componentes individuais. Do ponto de vista semântico, sequências de palavras como *do you want me to*, *going to be a* e *I don't know why*, são completamente transparentes, já que seus respectivos significados são facilmente obtidos por meio de suas partes constituintes e da maneira como estão organizados dentro da sequência. Ao contrário do que acontece com expressões idiomáticas, como *slap in the face* e *spill the beans*, a transparência dos pacotes lexicais permite que eles sejam processados de forma rápida, e facilmente compreendidos por um falante de inglês como primeira língua ou língua estrangeira (Biber; Conrad; Cortes, 2004).

Em segundo lugar, os pacotes lexicais são, na maioria das vezes, estruturas linguísticas incompletas, usadas como pontes entre duas frases (*e.g. in the case of*) ou duas orações (*e.g. I want to know*). Mesmo sendo estruturas linguísticas incompletas, diversos estudos apontam que a categorização dos elementos gramaticais que compõem os pacotes lexicais é um passo importante na investigação de como essas sequências são utilizadas em determinados tipos de registros. Na presente tese, foi adotada a taxonomia proposta por Biber, Conrad e Cortes (2004), o motivo de me basear em tal taxonomia se deu por ela ser comumente utilizada na literatura acadêmica e por oferecer uma categorização capaz de ser adaptada para pacotes lexicais em diversos registros linguísticos. Dessa forma, a taxonomia proposta para categorizar a estrutura interna dos pacotes lexicais em minha tese é apresentada da seguinte maneira:

Quadro 3 – Taxonomia estrutural proposta por Biber, Conrad e Cortes (2004)

Estrutura	Descrição	Exemplo
Tipo 1 - Fragmentos de sintagmas verbais (SV)	sequências formadas por: sujeito + verbo principal, fragmentos SV ou fragmentos de perguntas.	<i>and I'm going to; what do you think</i>
Tipo 2 - Fragmentos de orações subordinadas	sequências que além de conterem fragmentos SV, também apresentam indícios de uma oração subordinada.	<i>if you look at; I don't know if</i>
Tipo 3 - Fragmentos de sintagmas nominais (SN) e/ou preposicionais (SP)	caracterizada pelo núcleo do SN ou SP + pós-modificador.	<i>the end of the; one of the most;</i>

Fonte: Biber, Conrad e Cortes (2004, p. 381)

Em textos do registro acadêmico escrito (livros didáticos e prosa acadêmica), Biber *et al.* (1999) observaram que a maioria dos pacotes lexicais extraídos eram formados por estruturas do Tipo 3 (*e.g. as a result of, at the end of, one of the most*). Por outro lado, em textos do registro acadêmico oral (aulas expositivas, palestras acadêmicas, apresentações e seminários), Biber, Conrad e Cortes (2004) e Nesi e Basturkmen (2006) relataram que os três tipos de pacotes lexicais tendem a ocorrerem de maneira similar. Segundo os autores, isso se deve ao fato de esses registros produzirem textos que, embora sejam pré-planejados, também sofrem influência de fatores típicos da produção oral, tais como: produção em tempo real e relações entre os participantes. Logo, textos do discurso oral acadêmico tendem a incorporar pacotes lexicais típicos de registros orais e escritos.

• Atribuição de funções discursivas

Visto que “o *corpus* pode oferecer evidências, mas não informações²⁵” (Hunston, 2002, p. 23), a investigação das funções discursivas desempenhadas em um determinado registro é, na maioria dos casos, o objetivo final dos estudos envolvendo pacotes lexicais. A primeira classificação funcional para pacotes lexicais foi proposta no estudo de Biber, Conrad e Cortes (2004), na qual é apresentada três grandes categorias de pacotes lexicais: *expressões de posicionamento*; *organizadores discursivos* e *expressões referenciais*. Por ter sido desenvolvida a partir de *corpus* de estudo bastante heterogêneo, essa taxonomia é considerada suficientemente flexível para categorizar pacotes lexicais em diferentes registros.

²⁵ No original: “a corpus can offer evidence but cannot give information.”

Quadro 4 – Taxonomia funcional proposta por Biber, Conrad e Cortes (2004)

Subdivisões		Exemplos
Expressões de posicionamento		
A. Epistêmico	Pessoal	<i>I don't know if</i>
	Impessoal	<i>are more likely to</i>
B. Atitudinal/modal		
Desejo	Pessoal	<i>I don't want to</i>
Obrigações/Direcionamentos	Pessoal	<i>I want you to</i>
	Impessoal	<i>it is necessary to</i>
Intenções/Predição	Pessoal	<i>I'm not going to</i>
	Impessoal	<i>it's going to be</i>
Habilidades	Pessoal	<i>to be able to</i>
	Impessoal	<i>can be used to</i>
Organizadores discursivos		
A. Introdução de tópico		<i>if you look at</i>
B. Elaboração/clarificação		<i>on the other hand</i>
Expressão de referência		
A. Identificação/foco		<i>one of the things</i>
B. Imprecisão		<i>or something like that</i>
C. Especificação de atributos		
Quantidades		<i>there's a lot of</i>
Atributos tangíveis		<i>the size of the</i>
Atributos intangíveis		<i>the nature of the</i>
D. Tempo/lugar/texto		
Referência espacial		<i>in the United States</i>
Referência temporal		<i>at the time of</i>
Referência dêitica		<i>as shown in figure</i>
Referência multifuncional		<i>the end of the</i>

Fonte: Adaptado de Biber, Conrad e Cortes (2004)

Na categoria *expressão de posicionamento*, encontram-se os pacotes lexicais que indicam o posicionamento do falante em relação às suas proposições. Essa categoria está subdividida em dois grupos: epistêmico e atitudinal/modal. No primeiro grupo, estão os pacotes lexicais que expressam o grau de assertividade do falante (certeza, incerteza, possibilidade); já o segundo grupo sinaliza a postura do falante em relação a seus desejos, obrigações, intenções e habilidades. Ambas as subcategorias podem conter expressões de cunho pessoal (*e.g. and I think that*), ligadas diretamente ao falante, ou impessoais (*e.g. the fact that the*), que expressam as mesmas ideias sem estarem atribuídas explicitamente ao mesmo.

Na categoria de *organizadores discursivos*, encontram-se os pacotes lexicais utilizados para estruturar o texto e indicar o modo como o falante organiza suas ideias e a coerência de

seu discurso. Essa categoria também está dividida em dois grandes grupos: introdução de tópico e elaboração/clarificação de tópico.

No que diz respeito à categoria de *expressões de referência*, estão incluídos os pacotes lexicais de identificação e especificação de atributos para elementos do discurso. Os autores dividem essa categoria em quatro grandes grupos: identificação/foco, imprecisão, especificação de atributos e referências de tempo, lugar e texto.

Uma variação da taxonomia proposta por Biber, Conrad e Cortes (2004) foi apresentada por Hyland (2008a) e está descrita no Quadro 5. O autor justifica a necessidade de adaptação pelo fato de que, quanto mais específico for o registro analisado, mais particular será a sua função discursiva dos pacotes lexicais encontrados. Especificamente, o autor buscou representar melhor as funções discursivas de pacotes lexicais em registros especializados de escrita acadêmica, tais como: artigos científicos; dissertações e teses.

Quadro 5 – Taxonomia funcional proposta por Hyland (2008a)

Funções	Descrições	Exemplos
Direcionados à pesquisa	Descrevem o objeto e o contexto do estudo, especificam aspectos acerca de modelos, de equipamentos e de materiais do meio científico.	<i>at the same time the structure of the the purpose of the</i>
Direcionados ao texto	Apresentam a organização, a mensagem e os argumentos utilizados pelos autores no texto.	<i>on the other hand these results suggest that it was found that</i>
Direcionados ao participante	Pacotes lexicais orientados aos escritores e aos leitores do texto. Podem transmitir opinião e avaliação pessoal do autor, assim como se referir diretamente ao leitor.	<i>as can be seen it is possible that it should be noted that</i>

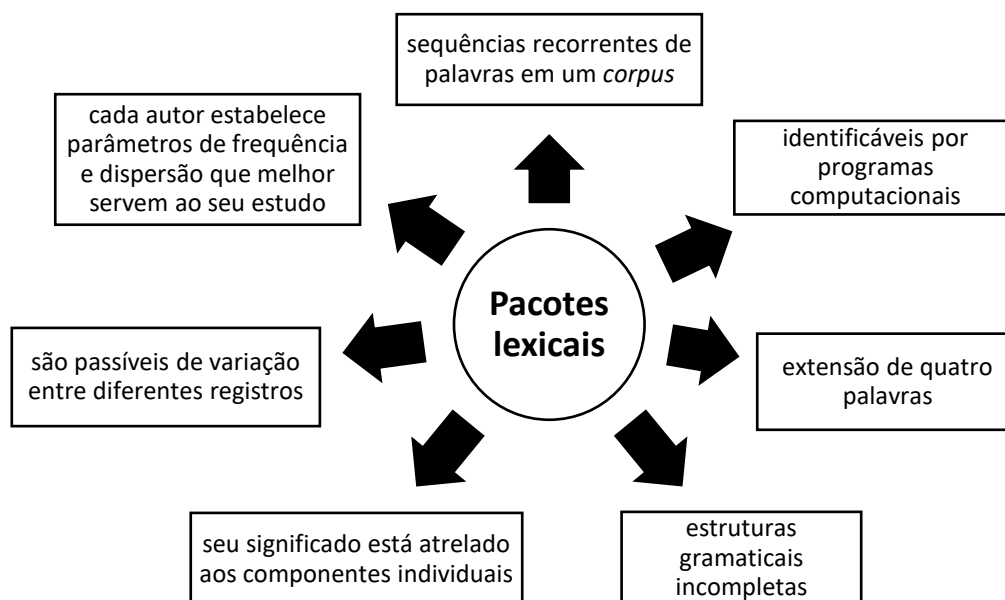
Fonte: Adaptado de Hyland (2008a).

Os estudos de Durrant (2015), Hyland e Jiang (2018), Salazar (2014) e Shirazizadeh e Amirfazlian (2021) são alguns dos que aplicaram a taxonomia de Hyland (2008a) para classificarem os pacotes lexicais mais frequentes em produções escritas de gêneros acadêmicos.

No entanto, as categorias funcionais de Biber, Conrad e Cortes (2004) me parecem mais alinhadas ao tipo de texto em análise na presente tese pelos seguintes motivos: (i) ter sido proposta a partir de um *corpus* contendo diferentes registros acadêmicos (orais e escritos) e (ii) por ser amplamente adotada na literatura e adaptável a diversos registros, inclusive TED Talks (*cf.* Liu; Chen, H. 2020b).

Dessa forma, o conceito de pacotes lexicais utilizado na presente tese pode ser sintetizado da seguinte maneira:

Figura 5 – Principais características relacionadas aos pacotes lexicais



Fonte: Elaborado pelo autor

Antes de concluir a presente seção, na Tabela 1, apresento uma breve compilação de trabalhos relacionados a investigação de pacotes lexicais em diferentes contextos acadêmicos e, posteriormente, discuto o que alguns deles revelaram sobre o uso da língua em diferentes registros e domínios disciplinares. Todavia, decidi por priorizar trabalhos nos quais os pacotes lexicais foram identificados por meio dos parâmetros de frequência, dispersão e função discursiva, uma que vez que estudos nos quais foram utilizados parâmetros adicionais (*e.g. MI score*, *chavicidade*) não estariam investigando pacotes lexicais propriamente ditos, mas sim algum outro tipo de sequência formulaica (Biber, 2009; Cortes, 2015; Wood, 2015), tais como: fórmulas acadêmicas (Simpson-Vlach; Ellis, 2010) e *key lexical bundles* (Bocorny *et al.*, 2021).

Tabela 1 – Coletânea de trabalhos relacionados a pacotes lexicais na literatura acadêmica

(continua)

Autor(es)	Tipologia do <i>corpus</i> de estudo	Parâmetros para identificação dos pacotes lexicais		
		Extensão	FR (por milhão de palavras)	Dispersão
Bértoli e Shepherd (2015)	escrito / ensaio acadêmico	4 palavras	5 ocorrências (absolutas)	não especificado
Biber (2006a)	diversos registros escritos e orais	4 palavras	40 ocorrências	5 textos
Biber, Conrad e Cortes (2004)	diversos registros escritos e orais	4 palavras	40 ocorrências	5 textos
Biber e Barbieri (2007)	diversos registros escritos e orais	4 palavras	40 ocorrências	3 textos
Byrd e Coxhead (2010)	escrito / artigo acadêmico	4 palavras	20 ocorrências	não especificado
Cao (2021)	escrito / artigo acadêmico	4 palavras	20 ocorrências	10% dos textos
Chen, L. (2010)	escrito / livro didático	4 palavras	20 ocorrências	5 textos
Chen, Y. e Baker (2010)	escrito / artigo acadêmico	4 palavras	25 ocorrências	3 textos
Conrad e Biber (2004)	oral / palestra acadêmica e conversa face-a-face	3 e 4 palavras	10 ocorrências	5 textos
Cortes (2004)	escrito / artigo científico	4 palavras	20 ocorrências	5 textos
Cortes (2013)	escrito / artigo acadêmico (seção introdução)	4(+) palavras	variada	5 textos
Cortes e Csomay (2007)	oral / palestra acadêmica	4 palavras	40 ocorrências	5 textos
Csomay (2013)	oral / aula expositiva	4 palavras	40 ocorrências	não especificado
Durrant (2015)	escrito / artigo científico	4 palavras	variada	variada
Dutra e Berber Sardinha (2014)	escrito / redações argumentativas	3 palavras	variada	variada
Dutra, Orfano e Berber Sardinha (2014)	escrito / ensaios acadêmicos	4 palavras	20 ocorrências	5 textos
Freitas (2016)	escrito / <i>abstract</i>	4 palavras	5 ocorrências (absolutas)	5 textos

(conclusão)

Autor(es)	Tipologia do <i>corpus</i> de estudo	Parâmetros para identificação dos pacotes lexicais		
		Extensão	FR (por milhão de palavras)	Dispersão
Gray (2016)	escrito / fórum de perguntas e respostas	4 palavras	10 ocorrências (absolutas)	5 textos
Grondona (2015)	escrito / ensaio acadêmico	seleção prévia	variada	variada
Hyland (2008a; 2008b; 2012)	escrito / artigo acadêmico	4 palavras	20 ocorrências	10% dos textos
Hyland e Jiang (2018)	escrito / artigo acadêmico <i>corpus</i> diacrônico	4 palavras	20 ocorrências	10% dos textos
Kashiha e Heng (2014)	oral / palestra acadêmica	4 palavras	10 ocorrências (por 100 mil palavras)	3 textos
Liu e Chen, H. (2020a)	oral / palestra acadêmica	4 palavras	40 ocorrências	10% dos textos
Liu e Chen, H. (2020b)	oral / palestra acadêmica e TED Talks	4 palavras	30 ocorrências	5% dos textos
Matte e Goulart (2020)	escrito / redações de alunos idioma: português	3 palavras	12 ocorrências (absolutas)	5% dos textos
Nesi e Basturkmen (2006)	oral / palestra acadêmica	4 palavras	10 ocorrências (absolutas)	não especificado
Reppen e Olson (2020)	escrito / livro didático	4 palavras	20 ocorrências	2% dos textos
Salazar (2014)	escrito / artigo acadêmico	4-6 palavras	10 ocorrências + <i>MI score</i>	5 textos
Shepherd, Zyngier e Viana (2006)	escrito / redações de alunos idioma: português	4 palavras	3 ocorrências (absolutas)	3 textos
Simpson-Vlach e Ellis (2010)	escrito e oral / artigos e palestras acadêmicas	3-5 palavras	10 ocorrências + <i>MI score</i>	variada
Viana, Fausto e Zyngier (2007)	escrito / obras literárias língua: português	4 palavras	4 ocorrências (por 100 mil palavras)	não aplicado
Wang (2017)	oral / palestra acadêmica	4 palavras	25-26 ocorrências	2 textos
Wright (2019)	escrito / revisão de literatura	4 palavras	40 ocorrências	5 vezes em cada disciplina do <i>corpus</i>

Fonte: Elaborado pelo autor

No tocante aos resultados apresentados, o estudo conduzido por Cortes (2004), por exemplo, comparou o uso de pacotes lexicais em artigos acadêmicos nas disciplinas de história e de biologia, e descobriu que ambas as disciplinas apresentam diversos pacotes lexicais formados por sintagmas nominais e preposicionais (algo já esperado em registros escritos). No entanto, a disciplina de biologia também utiliza um número relevante de pacotes lexicais formados por outras estruturas, tais como: *it + be + adjetivo* (e.g. *it is important to*) e voz passiva + *that* (e.g. *has been shown that*).

A autora também observou variações nas ocorrências de pacotes lexicais com função anguladora (*hedge*) na produção escrita de textos da biologia (e.g. *are likely to be, it is possible that*). Tais dados sugerem que autores dessa disciplina tendem a utilizar sequências mais extensas para realizar tal função, enquanto na disciplina de história a assertividade do autor tende a ser expressa ou por palavras individuais, ou sequências menores que quatro palavras, tais como: *possibly, probably* e *in most cases*.

Hyland (2008a; 2008b), por sua vez, apresentou evidências de uma variação sistemática na forma como as disciplinas de: biologia, engenharia elétrica, linguística aplicada e economia empregam os pacotes lexicais para desenvolver seus argumentos, estabelecer credibilidade e persuadir seus interlocutores. Em resumo, os estudos Hyland (2008a; 2008b) reportaram que autores da disciplina de engenharia elétrica tendem a utilizar um conjunto maior de sequências pré-fabricadas em suas produções linguísticas (n = 213), ao passo que as outras disciplinas apresentam um leque mais restrito de opções. A disciplina de biologia, por exemplo, exibiu o menor número de pacotes lexicais distintos (n = 131). O autor sugere que essa diferença pode estar relacionada ao fato de a disciplina de engenharia elétrica apresentar dados com o suporte de representações gráficas e numéricas, levando os autores dessa disciplina a utilizarem mais expressões rotineiras para estabelecerem a relação entre dados e argumentos. A biologia, por outro lado, tende a ser mais discursiva e descritiva, o que leva os autores dessa disciplina a empregarem um conjunto menor de formas rotineiras em seus discursos.

A estrutura e a função discursiva dos pacotes lexicais também mostraram evidências de natureza disciplinar. Hyland (2008a) identificou que as disciplinas de linguística aplicada e de economia empregam mais pacotes lexicais iniciados por sintagma preposicional + *of* (e.g. *on the basis of, in the case of*), comumente utilizados para expressar relações lógicas no texto. Por outro lado, as disciplinas de biologia e de engenharia elétrica tendem ao uso maior de

pacotes lexicais na voz passiva, seguidos de um sintagma preposicional (*e.g. can be found in, is based on the*). Hyland (2008a) sugere que esta diferença pode estar relacionada à preocupação natural de campos mais técnicos em destacar os dados científicos e reduzir a interpretação pessoal do pesquisador, sugerindo que os resultados obtidos seriam iguais, independentemente de quem conduzisse o experimento. Por outro lado, autores das disciplinas de linguística aplicada e economia se preocupam em apresentarem argumentos que se conectem às condições delimitadas e particulares de suas pesquisas.

Conclusões semelhantes também foram apresentados por Durrant (2015), ao examinar a ocorrência de pacotes lexicais em diferentes disciplinas do *British Academic Written English Corpus* (BAWE). O autor argumenta que alunos das disciplinas de humanidades e ciências sociais (*soft disciplines*) tendem a posicionamentos menos assertivos, com conclusões expostas de maneira contextualizada, interpretativa e explanatória. Por outro lado, alunos de ciências e tecnologia (*hard disciplines*) tendem a relatar suas descobertas de forma mais direta, embasando seus argumentos em relações lógicas de causa e efeito já aceitas como fatos dentro da disciplina e, portanto, limitadas a novas interpretações.

Em um estudo mais recente, Reppen e Olson (2020) investigaram pacotes lexicais em um *corpus* de estudo composto por 25 milhões de palavras dividido em nove disciplinas. Após o processo de extração dos dados, mais de 700 pacotes lexicais foram encontrados e classificados em dois grupos: os transdisciplinares (*cross-disciplinary*) e os disciplinares (*discipline-specific*). No primeiro grupo, estão os pacotes lexicais que ocorreram em pelo menos quatro disciplinas, enquanto no segundo grupo, a sequência ocorre em, no máximo, duas disciplinas e, comumente, expressam conteúdos de caráter especializado (*e.g. of the interior design, role of hotel owners*). Devido ao interesse no aspecto da variação disciplinar, as autoras optaram por investigarem somente os pacotes lexicais do primeiro grupo.

Um aspecto interessante do estudo de Reppen e Olson (2020) foi a identificação de uma variável produtiva interna nos pacotes lexicais. As autoras constataram que, em geral, o segundo componente em pacotes lexicais de quatro palavras [1*34] tende a ser o mais produtivo em termos de variação. No caso da estrutura *the * of the*, por exemplo, as autoras identificaram 23 possíveis "encaixes" recorrentes para o segundo componente, tais como: *end, size, nature, top, center, importance*, entre outras; revelando que essa estrutura possui grande variação interna e, conseqüentemente, maior produtividade linguística.

Em relação à variação disciplinar em textos do registro oral, o estudo de Wang (2017) investigou o uso de pacotes lexicais em palestras acadêmicas de um *corpus* de inglês como *lingua franca*, organizado em três grupos disciplinares: medicina, ciências sociais e ciências naturais. O autor apresenta evidências de convencionalidade e particularidades entre esses grupos. No que tange à convencionalidade, o pacote lexical *I would like to* foi o único em comum entre as três disciplinas. Segundo o autor, isso poderia indicar uma associação entre o pacote lexical *I would like to* e falantes de inglês como *lingua franca*. Todavia, o *corpus* de estudo utilizado por Wang (2017) pode ser considerado como bem pequeno (500 mil palavras), portanto, generalizações deste tipo devem ser vistas com cautela, já que, curiosamente, pacote lexical *I would like to* não está entre os mais recorrentes em um estudo com o mesmo registro feito por Nesi e Basturkmen (2006), por exemplo.

Outro dado interessante está no maior uso de recursos visuais em palestras de Medicina. Muitos dos pacotes lexicais recorrentes nessa disciplina foram formados pelo pronome pessoal *you* + o verbo lexical *see* (e.g. *as you can see, you can see that*). Já os palestrantes da disciplina de ciências naturais tendem ao uso de pacotes lexicais formados pelo pronome pessoal *we* (e.g. *we are interested in*), posicionando o falante como membro de um grupo/comunidade que suporta suas afirmações e argumentos.

2.5 Fundamentação teórica: Considerações finais

Neste capítulo, busquei conduzir uma discussão acerca dos principais aspectos teóricos que norteiam a presente tese. Inicialmente, apresentei uma caracterização do gênero TED Talks pelo prisma do quadro de análises situacionais de Biber e Conrad (2009) para mapear os principais aspectos discursivos das TED Talks por meio de sete características principais, sendo elas: participantes, relacionamento entre os participantes, canal, circunstâncias de produção, cenário, propósitos comunicativos e tópico. Acredito que tal perspectiva contribuiu para um melhor entendimento da natureza deste registro e de como seu processo de produção ocorre em nossa sociedade.

Em seguida, abordei, de modo abrangente, o histórico da LC e as principais implicações acerca da utilização de *corpora* em investigações empíricas da língua em uso. Especificamente, direcionei a discussão aos paradigmas dos estudos dirigidos e baseados por *corpora* e ponderei que, embora já tenham sido apresentados como abordagens distintas, na verdade,

se complementam e são igualmente importantes na descrição de diversos tipos de linguagem formulaica.

Por fim, apresentei o conceito de pacotes lexicais, um tipo de sequência formulaica identificada mediante critérios quantitativos (frequência e dispersão), e discuti alguns estudos em que eles foram utilizados para investigar variações disciplinares em diferentes registros acadêmicos.

Com a fundamentação teórica estabelecida, a discussão agora se volta para a descrição detalhada dos materiais e métodos empregados nesta tese. No próximo capítulo, descrevo como foi realizado o processo de organização, coleta e limpeza dos textos compilados para os *corpora* de estudo TED-STEM e TED-SHAPE, assim como as ferramentas e técnicas e critérios adotados para identificação e análise dos pacotes lexicais. Sendo assim, o objeto do capítulo seguinte é fornecer uma visão abrangente do processo de pesquisa, permitindo a replicação e compreensão dos resultados obtidos, além de discutir como os métodos empregados se alinham com o referencial teórico apresentado neste capítulo.

3 MATERIAL E MÉTODOS

Neste capítulo, apresento o material e os métodos de pesquisa adotados para a condução deste estudo. Na Seção 3.1, descrevo o processo de coleta dos dados, incluindo os motivos para a compilação dos *corpora* de estudo TED-STEM e TED-SHAPE ao invés de utilizar *corpora* de TED Talks disponíveis publicamente e os princípios referentes ao *design* inicial de *corpora* de estudo propostos por Sinclair (2005). Em seguida, na Seção 3.2, discorro acerca da coleta dos textos, organização dos metadados e tipologia dos *corpora* de estudo. Na Seção 3.3, apresento a descrição dos *corpora* de estudo TED-STEM e TED-SHAPE. Já na Seção 3.4, abordo as funcionalidades da ferramenta AntConc®, utilizada para a análise dos *corpora*. Na Seção 3.5, descrevo os critérios para extração dos pacotes lexicais, elencando as principais decisões tomadas acerca da categorização estrutural e funcional deste objeto de estudo. Na Seção 3.6, discuto o processo de refinamento de pacotes lexicais em contexto de sobreposição e dependentes de contexto (Chen, Y. 2009; Grondona, 2015), a fim de evitar inflação e redundância durante a análise dos dados extraídos. Por último, na Seção 3.7, apresento as considerações finais do presente capítulo.

3.1 Princípios envolvidos na coleta de textos e compilação de *corpora* linguísticos

Ao realizar buscas por conteúdos e materiais disponibilizados para acesso público, encontrei o *corpus* especializado em TED Talks: *TED Talks Transcript*²⁶, o banco de dados online *TED Corpus Search Engine*, TCSE²⁷, (Hasebe, 2015) e o *Web Inventory of Transcribed and Translated Talks*²⁸, WIT³, um repositório de transcrições e traduções de TED Talks desenvolvido por Cettolo, Girardi e Federico (2012).

Entretanto, após uma análise cuidadosa desses recursos, observei que nenhum deles apresenta informações ou filtros que possibilitem uma investigação focada na análise de variação disciplinar em TED Talks. O WIT³, por exemplo, foi desenvolvido para pesquisas no

²⁶ *Corpus* integrado à plataforma *Sketch Engine*®, e compilado a partir de 1.361 TED Talks apresentadas entre os anos de 2006 e 2013. Contém aproximadamente 2.8 milhões de palavras. Disponível em: <https://www.sketchengine.eu/>. Acesso em: 01 maio 2021.

²⁷ Disponível em: <https://yohasebe.com/tcse/>. Acesso em: 01 maio 2021.

²⁸ Disponível em: <https://wit3.fbk.eu>. Acesso em: 01 maio 2021.

campo da linguística computacional e da tradução automática. Por sua vez, o TCSE e o *TED Talk Transcripts*, apesar de oferecerem funcionalidades como linhas de concordância, listas de palavras e até mesmo listas de pacotes lexicais, não apresentam nenhum tipo de organização disciplinar. Reconheço que os três recursos apresentados acima possam ser descritos como materiais especializados no gênero textual TED Talks, contudo não encontrei indícios de que foram aplicados critérios de seleção ou restrição quanto ao conteúdo das apresentações incluídas nesses bancos de dados. Por esse motivo, eles foram julgados como inadequados dentro dos objetivos propostos para este estudo.

Diante da possível inexistência de um material de acesso público alinhado com o escopo desta tese, optei por compilar meus próprios *corpora* de estudo. Para tanto, foi necessário primeiro estabelecer critérios que assegurassem uma coleta de dados com delimitações consistentes e objetivas, possibilitando análises linguísticas mais finas e granuladas dos pacotes lexicais em TED Talks.

Embora os processos de compilação de *corpora* estejam sujeitos às particularidades de cada pesquisa (Flowerdew, 2004; Viana; O'Boyle, 2022), Sinclair (2005) propõe um conjunto de princípios que operam como diretrizes básicas tanto para compilação de *corpora* escritos quanto orais. Nos parágrafos subsequentes, discutirei como tais princípios influenciaram as tomadas de decisões envolvidas na compilação dos *corpora* de estudo utilizados nesta tese.

Princípio 1 – Os conteúdos de um *corpus* não devem ser selecionados pela linguagem que contêm, mas sim de acordo com a função comunicativa que desempenham²⁹ (Sinclair, 2005, n.p.).

É possível afirmar que este princípio foi respeitado uma vez que não houve levantamento acerca das características linguísticas (e.g. uso de substantivos, frases verbais e nominais) das TED Talks. Neste estudo, não utilizei nenhuma *seed word*³⁰ como gancho para a coleta de dados. Logo, os textos foram selecionados com base em critérios externos à linguagem neles contida.

²⁹ No original: “The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise”.

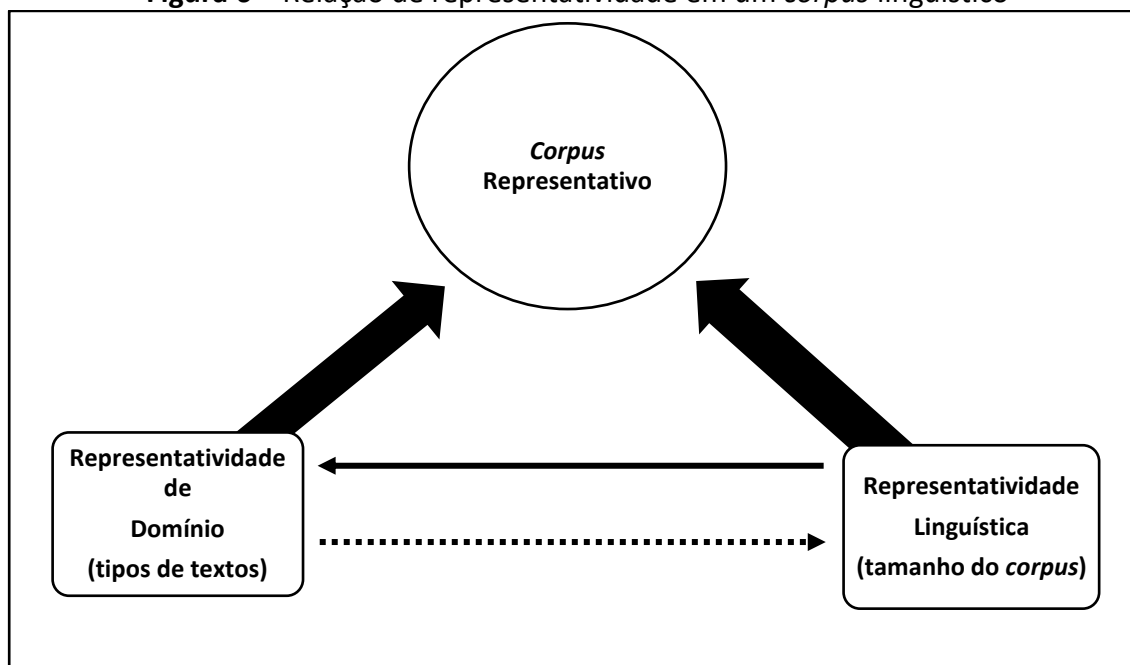
³⁰ *Seed words* são palavras utilizadas por *softwares* para encontrar textos relevantes à temática de um *corpus*.

Princípio 2 – Os compiladores de um *corpus* devem se esforçar para torná-lo o mais representativo possível da língua a ser analisada³¹ (Sinclair, 2005, n.p.).

O adjetivo “representativo” deve ser utilizado com cautela nos estudos de LC, visto que, em geral, um *corpus* é descrito como uma amostra de um conjunto de dados linguísticos cuja real dimensão é desconhecida. A questão do que seria ou não um *corpus* representativo já foi abordada em diversos trabalhos na literatura acadêmica (Atkins; Clear; Ostler, 1992; Biber, 1993; Egbert, 2019; Kennedy, 1998; Leech, 2007; Viana; O’Boyle, 2022), e, embora diversos autores já tenham se debruçado sobre este tema ainda não há um consenso claro e objetivo que garanta um *corpus* verdadeiramente representativo.

Para Biber (1993, p. 243), a representatividade estaria associada à “extensão na qual uma amostra inclui todo o conjunto de variabilidade de uma população”³². Ademais, tanto Biber (1993) quanto Egbert (2019) argumentam que a representatividade de um *corpus* pode ser avaliada por meio de dois prismas: domínio e linguístico; conforme ilustra a Figura 6.

Figura 6 – Relação de representatividade em um *corpus* linguístico



Fonte: Elaborado pelo autor com base nas leituras de Biber (1993) e Egbert (2019).

³¹ No original: “Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.”

³² No original: “representativeness refers to the extent to which a sample includes the full range of variability in a population.”

De acordo com os autores, um *corpus* é representativo de um domínio no sentido de quão generalizável é a sua variedade textual em relação à população estudada. Já a representatividade linguística determina o quão apropriado é o *corpus* para investigações acerca do uso da língua. As setas horizontais indicam dois tipos de relações entre representatividade linguística e domínio. A seta contínua indica que representatividade linguística pressupõe a representatividade de domínio, contudo o oposto não é necessariamente verdadeiro. Em outras palavras, é possível obter um *corpus* com uma quantidade considerável de textos de um determinado domínio, mas que ainda assim não seja grande o suficiente para fazer generalizações acerca do uso linguístico de uma população.

Leech (2007; 2011), por sua vez, faz uma distinção entre os termos: *representativeness* e *representativity*. O primeiro refere-se a um conceito de *corpus* polarizado (representativo X não representativo), enquanto o segundo descreve o *corpus* em um espectro gradativo, que varia entre o mais ou o menos representativo de uma língua. Na presente tese, a linha de pensamento adotada se alinha com o conceito de *representativity* de Leech (2007), no sentido de terem sido adotados procedimentos que garantissem uma escala de representatividade. Vale ressaltar que o conceito de representatividade almejado na presente tese não diz respeito a todas as TED Talks produzidas, mas sim às aquelas associadas a conteúdos disciplinares.

A fim de obter uma amostra que fosse a mais representativa possível, o primeiro passo foi determinar a população total de TED Talks disponíveis, que, conforme o site oficial³³, era em torno de 3.800. Em seguida, levantei quais eram os tópicos disciplinares utilizados pelo site, assim como a quantidade palestras associadas a eles (cf. Tabela 2, Seção 3.2). Somente após a obtenção destes dados é que se deu a coleta dos textos.

Sendo assim, acredito que os *corpora* de estudo compilados neste estudo apresentam um alto grau de representatividades de TED Talks disciplinares, no que diz respeito a: i) *tamanho do corpus* – ambos possuem aproximadamente 1.5 milhão de palavras; ii) *número de textos* – os *corpora* de estudo possuem mais de 700 textos cada e iii) *disciplinas* – o *corpus* TED-STEM contém textos associados a 16 disciplinas distintas, ao passo que o TED-SHAPE possui 18 disciplinas. Tais parâmetros podem ser utilizados como indicadores do quanto os

³³ Segundo os dados coletados em junho de 2021.

corpora de estudo compilados na presente tese são representativos no contexto das TED Talks disciplinares.

Princípio 3 – Apenas os componentes dos *corpora* que foram projetados para análises comparativas devem ser comparados³⁴ (Sinclair, 2005, n.p.).

Embora os *corpora* de estudo tenham sido compostos por disciplinas individuais, é importante ressaltar que o objetivo primário deste estudo não é apresentar generalizações acerca das características linguísticas específicas dessas disciplinas, mas sim comparar o uso de pacotes lexicais em dois grupos disciplinares distintos: STEM e SHAPE.

Ademais, os *corpora* de estudo foram compilados de maneira não-probabilística (por conveniência), restringindo, assim, a representatividade dos *corpora* à própria amostra. Em outras palavras, a análise dos resultados diz respeito somente aos dados encontrados nos *corpora* TED-STEM e TED-SHAPE e, portanto, não podem ser utilizados como base para generalizações sobre o uso da língua como um todo.

Contudo, reconheço que alguns componentes dos *corpora* são substanciais o suficiente para serem comparados entre si. As disciplinas de: arquitetura, biologia, direito, economia, medicina e psicologia, por exemplo, possuem dimensões (de domínio e linguísticas) que permitem generalizações acerca das características linguísticas de TED Talks relacionadas a tais disciplinas, e, caso seja oportuno, poderão ser discutidas durante a análise de dados.

Princípio 4 – Os critérios para determinar a estrutura de um *corpus* não devem ser numerosos e devem ser claramente separados e sinérgicos no delineamento de um *corpus* que seja representativo da língua ou variedade linguística estudada³⁵ (Sinclair, 2005, n.p.).

Sinclair (2005) explica que alguns critérios já são pré-determinados pelas circunstâncias associadas ao *design* do *corpus*. No presente estudo, por exemplo, a compilação de *corpora* de TED Talks já estabelece um conjunto de critérios que são inerentes aos mesmos, tais como:

³⁴ No original: “Only those components of corpora which have been designed to be independently contrastive should be contrasted”.

³⁵ No original: “Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination”.

- *Modo do texto*: Originado da língua falada;
- *Tipo de texto*: Palestras de popularização do conhecimento;
- *Domínio dos textos*: Híbrido (acadêmico + popular);
- *Origem dos textos*: Web

No entanto, o escopo desta tese são as TED Talks de natureza disciplinar, portanto, foi necessário também adotar critérios internos ao contexto deste estudo, tais como:

- *Disciplina*: Apenas TED Talks associadas a algum domínio disciplinar foram selecionadas;
- *Tipo de texto (2)*: TED Talks produzidas em forma de entrevistas (perguntas e respostas) e animações visuais (TED-Ed) não foram selecionadas, mesmo possuindo natureza disciplinar;
- *Origem dos textos (2)*: Apenas as transcrições disponibilizadas pelo site oficial foram aceitas. Transcrições produzidas por algoritmos de reconhecimento automático (Youtube®) ou disponibilizados por terceiros não foram considerados.

Vale lembrar que os critérios apresentados acima são pertinentes ao escopo desta tese e não necessariamente seriam interessantes para outros tipos de investigações linguísticas envolvendo as TED Talks.

Princípio 5 – Qualquer informação sobre um texto que não seja a sequência alfanumérica de suas palavras e pontuação deve ser armazenada separadamente do texto simples e recuperada apenas quando necessário³⁶ (Sinclair, 2005, n.p.).

Este princípio mostrou-se fundamental durante a organização dos metadados dos *corpora* de estudo. Diversas informações extralinguísticas dos textos foram armazenadas em planilhas do Microsoft Excel®, permitindo fácil acesso a essas informações sempre que necessário. Entre os metadados coletados, vale destacar aqui o nome do palestrante, ano de produção, disciplina na qual o texto foi associado e notas do pesquisador justificando a sua escolha, quando necessário.

³⁶ No original: “Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications”.

A organização dos metadados aplicados nos *corpora* de estudo está descrita com maiores detalhes na Seção 3.2.

Princípio 6 – O *corpus* deve, sempre que possível, incluir documentos inteiros ou transcrições de eventos de fala completos, ou devem chegar o mais próximo possível desse objetivo. Isso significa que as amostras serão substancialmente diferentes em tamanho³⁷ (Sinclair, 2005, n.p.).

As TED Talks são disponibilizadas publicamente sob a licença *Creative Commons* (BY-NC). Tal licença, permitiu a adição de textos na íntegra aos *corpora* de estudo sem que houvesse conflitos acerca dos direitos autorais dos palestrantes, uma vez que as TED Talks são utilizadas aqui para fins não comerciais.

Conforme esperado, o tamanho das amostras em cada *corpora* de estudo apresentou diferenças, com o *corpus* TED-STEM apresentando aproximadamente 180 mil palavras a mais em comparação com TED-SHAPE. Entretanto, essa disparidade foi abordada por meio do cálculo de frequências relativas, garantindo análises consistentes e comparáveis entre os *corpora* de estudo.

Princípio 7 – O *design* e a composição de um *corpus* devem ser completamente documentados com informações sobre os conteúdos e argumentos que justificam as decisões tomadas³⁸ (Sinclair, 2005, n.p.).

Em consonância com o quinto princípio, os metadados de ambos os *corpora* de estudo foram meticulosamente organizados em uma planilha do Microsoft Excel®. Isso permitiu o fácil acesso a informações relevantes, como: nome do palestrante, título da TED Talk, duração do vídeo, dentre outros. Essa documentação detalhada das decisões tomadas e dos conteúdos incluídos nos *corpora* facilita a compreensão do processo de compilação e possibilita a replicabilidade do estudo por outros pesquisadores interessados em abordagens similares.

³⁷ No original: “Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events or should get as close to this target as possible. This means that samples will differ substantially in size”.

³⁸ No original: “The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken”.

Princípio 8 – O compilador do *corpus* deve manter como noções-alvo a representatividade e equilíbrio. Embora tais objetivos não sejam precisamente definíveis ou alcançáveis, eles devem orientar o desenho do *corpus* e a seleção de seus componentes³⁹ (Sinclair, 2005, n.p.).

Conforme discutido no Princípio 2, busquei compilar *corpora* de estudo com o máximo de representatividade de TED Talks disciplinares possível. Como ilustrado na Tabela 2, identifiquei 57 etiquetas disciplinares no site oficial das TED Talks, que foram posteriormente organizadas em 34 disciplinas divididas entre os domínios disciplinares STEM e SHAPE.

No que se refere à noção de equilíbrio, Sinclair (2005, n.p.) esclarece que “em linhas gerais, para que um *corpus* seja considerado equilibrado, as proporções dos diferentes tipos de textos que ele contém devem corresponder a julgamentos de natureza objetiva quanto intuitiva”⁴⁰. Assim, após determinar a população total de textos em cada disciplina, busquei deixar aquelas com mais vídeos (biologia, ciências políticas, economia e medicina, por exemplo) com as maiores proporções nos *corpora* de estudo. No entanto, reconheço aqui duas questões que dificultaram a orientação de equilíbrio em meus *corpora* de estudo. Em primeiro lugar, a atribuição de uma única disciplina para cada texto é discutível e subjetiva às experiências de cada pesquisador. Em segundo lugar, devido aos critérios estabelecidos antes da compilação dos textos, nem todos puderam ser selecionados, alterando assim a noção de equilíbrio dentre as disciplinas. Um exemplo disso ocorreu com a disciplina de história, que inicialmente era a quarta etiqueta disciplinar com maior quantidade de vídeos (n= 274), contudo, durante a compilação dos textos, percebi que mais de 80% deles eram em formatos de animações gráficas (TED-Ed), o que reduziu drasticamente o número de textos selecionados para a amostra em comparação com a população total da etiqueta disciplinar.

Princípio 9 – Qualquer controle sobre o conteúdo em um *corpus* deve ser imposto pelo uso de critérios externos e não internos⁴¹ (Sinclair, 2005, n.p.).

³⁹ No original: “The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components”.

⁴⁰ No original: “Roughly, for a corpus to be pronounced balanced, the proportions of different kinds of text it contains should correspond with informed and intuitive judgements”.

⁴¹ No original: “Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria”.

Conforme Sinclair (2005), pesquisadores podem sentir-se tentados a controlar o conteúdo de seus *corpora*, especialmente em contextos de linguagem altamente especializada. Assim, a seleção de TED Talks com base em tópicos disciplinares pode ser entendida como um controle semântico de conteúdo e, portanto, interno. No entanto, reitero que o uso de etiquetas disciplinares neste estudo não teve o objetivo de restringir o conteúdo semântico dos *corpora* de estudo, mas sim servir como parâmetros de natureza sociocultural.

Embora a decisão de compilar apenas TED Talks indexadas com etiquetas disciplinares possa influenciar indiretamente o conteúdo dos *corpora* STEM e SHAPE, Sinclair (2005, n.p.) esclarece que "a classificação de um tópico é muito mais sociolinguística do que semântica, e, portanto, é associada à cultura [de uma comunidade] e não aos significados das palavras"⁴². Ademais, como não foram realizadas análises semânticas prévias à compilação dos textos, pode-se argumentar que o principal critério de seleção de TED Talks apresentado neste estudo, etiquetas disciplinares; é externo ao vocabulário contido nos textos compilados.

Princípio 10 – Um *corpus* deve ter como objetivo a homogeneidade em seus componentes e ao mesmo tempo possuir uma abrangência adequada, além disso, textos intrusos devem ser evitados⁴³ (Sinclair, 2005, n.p.).

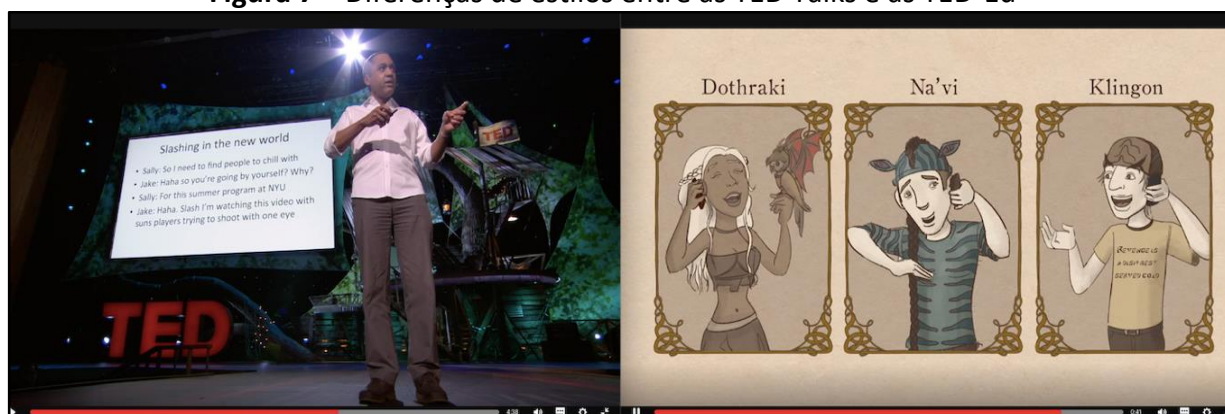
Com o objetivo de compilar *corpora* de estudo que fossem os mais abrangentes possíveis, coletei textos tanto de palestras oficiais (TED Talks), quanto de eventos independentes organizados sob o termo guarda-chuva TED, denominados TEDx. O diferencial desses eventos está no foco na divulgação de projetos com relevância regional onde são organizados.

Já os vídeos animados da série TED-Ed não foram selecionados para compor os *corpora* de estudo. Embora muitos estejam associados às etiquetas disciplinares, é possível observar na Figura 7 que eles apresentam divergências significativas em relação ao estilo e modo de apresentação das TED Talks "tradicionais".

⁴² No original: "topic classification turns out to be much more sociolinguistic than semantic, a therefore, dependent on the culture and not on the meaning of the words."

⁴³ No original: "A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided".

Figura 7 – Diferenças de estilos entre as TED Talks e as TED-Ed



Fonte: Site TED Talks

A figura acima ilustra duas apresentações produzidas pelo mesmo autor, John McWhorter, a imagem da esquerda ilustra o modo e o cenário de produção tradicionais ao gênero TED Talk. Já na imagem à direita, é possível observar de imediato uma grande mudança visual associada à produção do texto, além disso, o grande diferencial das TED-Ed está no fato delas não serem produzidas em tempo real pelo palestrante. Na verdade, elas são textos escritos, lidos em voz alta por um narrador. Logo, aspectos típicos da língua oral, tais como: repetição e reformulação não são marcas discursivas destes textos.

Outra grande diferença das TED-Ed é o grau de edição e revisão textual⁴⁴ que elas possuem quando comparadas às TED Talks tradicionais. Na Figura 7, a TED-Ed apresentada, além do narrador, também possui um diretor, e um designer audiovisual. Por estes motivos, a categoria de vídeos TED-Ed foi considerada aqui um tipo de gênero textual independente em relação às TED Talks.

Os princípios de Sinclair (2005) descritos acima são extremamente relevantes para qualquer pesquisador que precise compilar seu próprio *corpus* de estudo, pois asseguram consistência, representatividade e adequação dos dados aos objetivos de qualquer estudo com a LC.

Tendo conduzido uma breve discussão acerca de como tais critérios foram aplicados neste estudo, a próxima seção detalhará os processos envolvidos e as decisões tomadas no que diz respeito à coleta dos textos e à organização dos metadados que compõem os *corpora* de estudo.

⁴⁴ Em alguns casos, encontrei TED-Eds com até nove criadores, divididos entre os papéis de: educador, editor de roteiro, diretor, produtor, designer, animador, compositor e narrador.

3.2 Coleta dos textos, organização dos metadados e tipologia dos *corpora* de estudo

Considerando a vasta quantidade de material disponível no website das TED Talks, tornou-se imprescindível estabelecer parâmetros que assegurassem uma seleção objetiva e coerente dos textos a serem analisados. Tendo isso em mente, a etapa inicial consistiu no levantamento de todas as etiquetas indexadoras (*tags*) disponibilizadas no site. Essas etiquetas desempenham a função de categorizar o conteúdo interno do site, possibilitando o rápido filtro de vídeos associados a uma determinada etiqueta e, conseqüentemente, otimizando o processo de busca. Ademais, é possível combinar múltiplas etiquetas para buscas mais refinadas. A Figura 8 ilustra como os conjuntos de etiquetas iniciados com letra “A” são disponibilizados aos usuários do website.

Figura 8 – Etiquetas com a letra A na plataforma *TED Talks*

A			
Activism	Algorithm	Anthropocene	Astrobiology
Addiction	Aliens	Anthropology	Astronomy
Africa	Alzheimer's	Archaeology	Atheism
Aging	Ancient world	Architecture	Audacious Project
Agriculture	Animals	Art	Augmented reality
AI	Animation	Asia	Autism spectrum disorder
AIDS	Antarctica	Asteroid	

Fonte: Site TED Talks

Até onde foi possível averiguar, o website TED Talks não dispõe de diretrizes específicas para etiquetagem de seus vídeos⁴⁵, entretanto, essa funcionalidade provou-se extremamente útil para filtrar vídeos dentro da plataforma. Ao explorar tal recurso, identifiquei 349 etiquetas distintas⁴⁶, das quais, 57 foram consideradas como disciplinares. Essa relação é apresentada na Tabela 2, juntamente com as respectivas quantidades de

⁴⁵ Em uma busca pelas diretrizes de escolhas e publicações das *TED Talks*, não foi possível localizar nenhuma referência acerca dos padrões de etiquetas. Uma hipótese é que tal processo possa ser feito pelo próprio palestrante quando ele propõe o tema de sua comunicação, ficando a sua escolha a ordem e quantidade de etiquetas.

⁴⁶ É importante reforçar que o site recebe atualizações periódicas, logo a quantidade de etiquetas e vídeos pode variar de estudo para estudo.

vídeos. Convém mencionar que a quantidade de vídeos refere-se a maio de 2021, quando a coleta de dados foi realizada.

Tabela 2 – Etiquetas disciplinares no site TED Talks

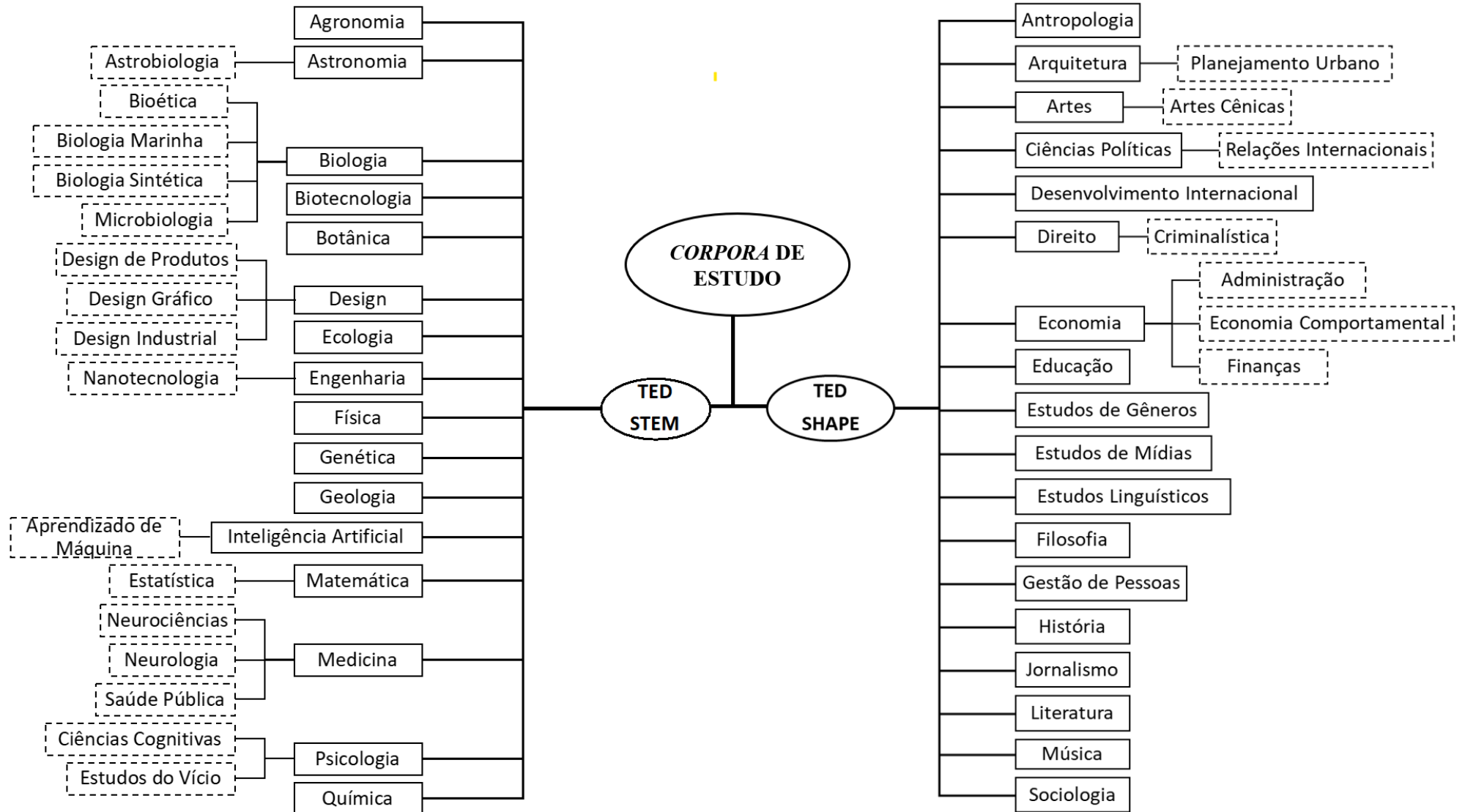
Etiquetas disciplinares	Vídeos	Etiquetas disciplinares	Vídeos	Etiquetas disciplinares	Vídeos
Administração	473	Design de produtos	60	Inteligência artificial	83
Agronomia	52	Design gráfico	10	Jornalismo	89
Antropologia	28	Direito	86	Literatura	27
Aprendizado de máquina	38	Ecologia	60	Matemática	65
Arquitetura	105	Economia	239	Medicina	238
Artes	326	Economia comportamental	27	Microbiologia	45
Artes cênicas	22	Educação	231	Música	186
Astrobiologia	14	Engenharia	170	Nanotecnologia	24
Astronomia	66	Estatísticas	33	Neurociência	114
Bioética	14	Estudos de gênero	105	Neurologia	2
Biologia	282	Estudos de mídia	97	Paleontologia	14
Biologia marinha	38	Estudos do vício	10	Planejamento urbano	46
Biologia sintética	21	Estudos linguísticos	78	Psicologia	169
Biotecnologia	85	Filosofia	74	Química	47
Botânica	18	Finanças	60	Relações internacionais	9
Ciência cognitiva	71	Física	95	Saúde pública	109
Ciências políticas	215	Genética	73	Sociologia	24
Criminalística	5	Geologia	8		
Desenho industrial	32	Gestão de pessoas	146		
Desenvolvimento internacional	84	História	274		

Fonte: Site TED Talks (maio, 2021)

Em maio de 2021, as TED Talks totalizavam aproximadamente 4 mil vídeos distintos. Nesse sentido, é importante esclarecer que os números apresentados acima indicam uma quantidade inflada de vídeos na plataforma. Isso acontece porque uma TED Talk que apresente, por exemplo, as etiquetas: Biologia, Medicina, Saúde Pública e Genética, estaria contabilizada quatro vezes na Tabela 2, visto que ela seria encontrada em cada etiqueta marcada.

Durante a organização dos metadados, também observei que algumas disciplinas se sobrepõem por possuírem objetos de estudo muito semelhantes, tais como Biologia e suas ramificações (bioética, marinha, microbiologia, sintética). A fim de evitar o uso excessivo de nomenclaturas, optei por unificar disciplinas similares àquela que possuísse o maior número de vídeos. Desta maneira, a Figura 9 seguir apresenta como foram organizados os principais grupos (multi)disciplinares em ambos os *corpora* de estudo.

Figura 9 – Organização dos agrupamentos multidisciplinares



Fonte: Elaborado pelo autor

As 57 etiquetas disciplinares inicialmente identificadas (*cf.* Tabela 2) foram organizadas em 34 grupos (multi)disciplinares. Em seguida, iniciei a organização dos *corpora* de estudo que englobassem tais disciplinas. Para denominar tais *corpora*, adotei as respectivas siglas: TED-STEM (*Science, Technology, Engineering, and Mathematics*) e TED-SHAPE (*Social Sciences, Humanities and the Arts for People and the Economy*). Observando o organograma ilustrado na Figura 10, percebe-se que o *corpus* TED-STEM foi desenhado pensando em agrupar palestras que discutissem conhecimentos produzidos nos campos das ciências exatas e biológicas. Da mesma forma, as TED Talks com temas associados às áreas de humanidades e ciências sociais foram agrupadas no *corpus* TED-SHAPE. A taxonomia STEM já é amplamente empregada para designar o ensino de conteúdos relacionados às ciências exatas, tecnológicas e biológicas (Bybee, 2013). Por outro lado, a proposta SHAPE, foi adotada recentemente por organizações como a Academia Britânica e pela Academia de Ciências Sociais do Reino Unido. Ambas as taxonomias têm como objetivo promover o ensino, aprendizagem e a realização de pesquisas que apliquem conhecimentos de diferentes campos do saber na busca por soluções para problemas e desafios de nossa sociedade contemporânea. A meu ver, esses termos também se adequam aos *corpora* de TED Talks, visto que essas palestras apresentam um alto nível de dinamismo e interação entre os mais variados campos do conhecimento.

Após a identificação das etiquetas de natureza disciplinar, o passo subsequente envolveu a criação de um banco de dados visando evitar redundâncias no fluxo de informações e assegurar que cada TED Talk fosse coletada uma única vez e atribuída a somente uma disciplina. Nesse sentido, utilizei o Microsoft Excel® para compilar os principais metadados que, posteriormente, nortearam as minhas tomadas de decisões. O respectivo banco de dados foi alimentado com as seguintes informações:

Figura 10 – Modelo de arquivamento dos metadados

A	B	C	D	E	F	G	H	I
Autor	TED	Ano	Duração	Etiquetas	Etiquetas disciplinares	Disciplina atribuída	Corpus	Profissão
Yoav Medan	Ultrasound surgery — healing without cuts	2011	0:15:45	Health Care Medical Research Medicine Science	Medicine Biotech Engineering Neuroscience Physics	Medicina	TED-STEM	Medical Inventor
J	K		L			M		
Link	Resumo do vídeo		Justificativa			Perguntas e discussões		
https://www.ted.com/talks/yoav_medan_ultrasound_surgery_healing_without_cuts	Imagine having a surgery with no knives involved. At TEDMED, Yoav Medan shares a technique that uses MRI to find trouble spots and focused ultrasound to		Descrição do vídeo: Foco no uso de ultrasonografia para aplicações cirúrgicas Palavras-chave: medical inventor, medical research, medical imaging e TEDMED					

Fonte: Elaborado pelo autor

O Quadro 6 a seguir apresenta maiores detalhes acerca dos metadados coletados.

Quadro 6 – Descrição dos metadados

Coluna	Tipo de informação
A – Autor	Nome do(s) palestrante(s).
B – TED	Título da palestra.
C – Ano	Ano em que a palestra foi apresentada ⁴⁷ .
D – Duração	Tempo de duração em [h]: [mm]: [ss].
E – Etiquetas	Relação de todas as etiquetas atribuídas ao vídeo, segundo a organização do próprio site.
F – Etiquetas disciplinares	Relação de etiquetas com caráter disciplinar.
G – Disciplina atribuída	Decisão final acerca de qual disciplina o texto foi designado.
H – <i>Corpus</i>	Indicação do <i>corpus</i> de estudo em que a TED Talk foi incluída.
I – Profissão	Informações relevantes acerca das experiências profissionais do palestrante.
J – Link	URL para extração das transcrições dos vídeos.
K – Resumo do vídeo	Breve descrição oferecida pelo site acerca do tema abordado durante a palestra. Esta coluna também serviu para informar as decisões finais sobre a organização disciplinar dos textos.
L – Justificativa	Em casos em que houve conflitos de disciplinas, é apresentado um pequeno texto explicando o porquê de tal decisão.
M – Perguntas e discussões	Indicação se ao final da palestra houve momento de discussão entre o palestrante e o mediador do evento. Em caso afirmativo, a interação foi ocultada usando as tags <discussion> </discussion>.

Fonte: Elaborado pelo autor

⁴⁷ Visto que há um intervalo de tempo entre a realização do evento e a postagem dos vídeos, decidi usar apenas a informação que é referente ao ano em que o evento foi realizado. Ambas as informações podem ser encontradas dentro site.

Em relação à Coluna F, diversas TED Talks apresentavam um conjunto de etiquetas disciplinar bem heterogêneo, conforme mostra a Figura 10. Neste caso específico, o mesmo vídeo foi associado a cinco etiquetas disciplinares⁴⁸: Medicina, Biotecnologia, Engenharia, Neurociência e Física. Em situações como essa, foi necessária uma avaliação manual para determinar qual disciplina melhor representaria o conteúdo da palestra. Para tanto, utilizei evidências como nome da TED Talk, descrição do vídeo, evento em que o vídeo foi apresentado, profissão do palestrante e outras etiquetas atribuídas ao vídeo. Dessa forma, as palavras: *medical inventor*, *medical research*, *medical imaging* e *TEDMED* foram interpretadas como indícios de uma associação mais forte à disciplina de medicina, em oposição à engenharia ou física, já a disciplina de neurociência foi integrada à medicina, conforme fora explicado nos parágrafos acima.

No tocante à organização dos *corpora* de estudo, reconheço a dificuldade e, até mesmo, a impossibilidade de rotular as TED Talks em uma disciplina específica. Portanto, ressalto que a organização apresentada na Figura 10, assim como a atribuição disciplinar dos textos, são relativas ao contexto desta tese e estão sujeitas a interpretações subjetivas.

Após a definição da estrutura dos *corpora* de estudo, o processo subsequente foi a coleta dos textos. Para esse fim, utilizei a ferramenta *WebBootCaT* da plataforma Sketch Engine® para extração automática dos textos a partir de listas de URLs. A escolha por essa ferramenta deve-se a duas vantagens metodológicas: (i) otimização do tempo – em questão de minutos, foi possível extrair o *script* completo de dezenas de TED Talks, automatizando assim um trabalho que levaria dias se realizado manualmente; (ii) os textos extraídos pelas URLs são anotados automaticamente com etiquetas codificadas em SGML⁴⁹. Os códigos *<doc>* e *</doc>* são especialmente importantes aqui, pois indicam, respectivamente, o início e fim de um documento de texto, possibilitando a segmentação do *corpus* em arquivos *.txt* individuais.

Para que a extração automática ocorresse sem problemas, era necessário que o final das URLs apresentasse o padrão: *“/transcript”*; isso assegurou que outras informações no link

⁴⁸ No presente estudo, decidi por não incluir a etiqueta *“science”* como disciplinar por entender que ela carrega um significado que engloba qualquer disciplina acadêmica.

⁴⁹ O SGML (*Standard Generalized Markup Language*) é uma linguagem internacional para formatação de arquivos. Documentos codificados em linguagem SGML podem ser transportados de um ambiente virtual para outro sem que ocorra perda de alguma informação.

da página não fossem coletadas com os textos do *corpus*. A figura abaixo ilustra como os textos foram extraídos.

Figura 11 – Modelo de extração das URLs

← TEXTS FROM WEB

Input type

- ☐ Web search ?
- ☒ URLs ?
- ☐ Website ?

https://www.ted.com/talks/amber_mcreynolds_an_election_system_that_puts_voters_not_politicians_first/transcript

https://www.ted.com/talks/katherine_m_gehl_us_politics_isn_t_broken_it_s_fixed/transcript

https://www.ted.com/talks/abigail_spanberger_how_to_connect_with_people_who_are_different_than_you/transcript

https://www.ted.com/talks/ameenah_gurib_fakim_and_stephanie_busari_an_interview_with_mauritius_s_first_female_president/transcript

Fonte: Sketch Engine®. Acesso em 08 de maio, 2021.

Ao término do processo, os textos de cada *corpora* foram baixados em arquivos no formato *.txt*. Após isso, as informações não linguísticas contidas nos documentos foram ocultadas utilizando a codificação SGML, conforme a descrição abaixo.

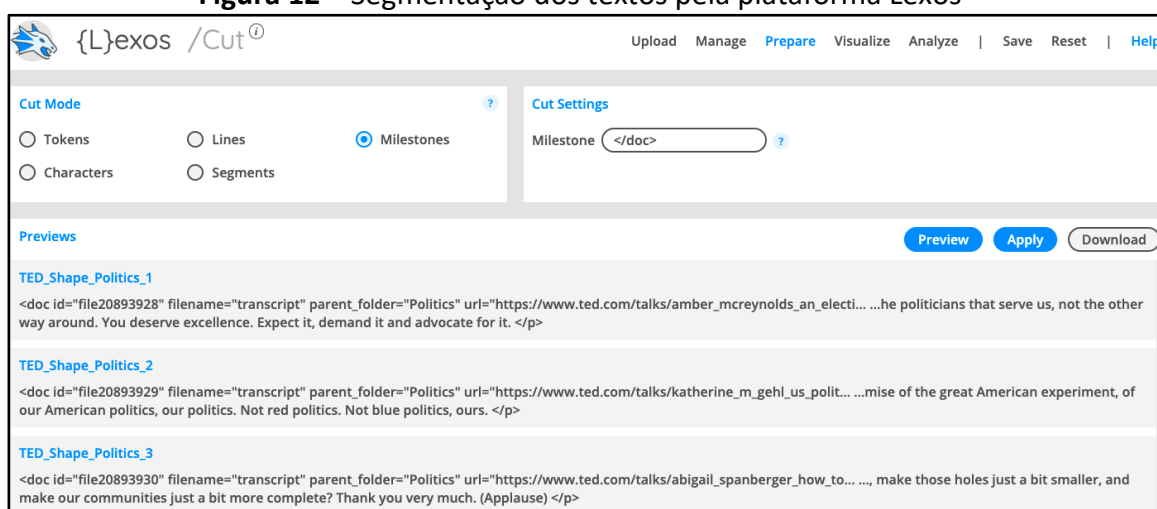
Quadro 7 – Descrição da codificação SGML utilizada

Etiqueta	Função
<header>	Informações contextuais fornecidas pelo <i>Sketch Engine</i> ® indicando o código de identificação do arquivo, diretório e URL.
<speaker>	Caso o nome do palestrante esteja indicado no início de um parágrafo da transcrição, ele é substituído por esta etiqueta e ignorado pelo software de análise lexical.
<laughter>	A indicação de risos durante a palestra foi codificada para não ser reconhecida durante a extração dos dados.
<video> </video>	Trechos da transcrição que são conteúdo de algum vídeo ou recurso audiovisual trazido pelo palestrante não foram deletados, contudo, eles foram codificados para serem ignorados pelo software de análise lexical.
<interviewer> </interviewer>	Em caso de haver um mediador na palestra, o seu nome e o trecho da transcrição produzida por ele são substituídos por esta etiqueta e codificados para serem ignorados pelo software de análise lexical.
<discussion> </discussion>	As interações entre o palestrante e o mediador, ao final da TED Talk, foram codificadas para serem ignorados pelo software de análise lexical.
<doc> </doc>	Etiqueta indicando o início e fim da transcrição.

Fonte: Elaborado pelo autor.

Para segmentar os textos em arquivos individuais, utilizei a plataforma Lexos⁵⁰, que, dentre diversas funcionalidades, permite separar o conteúdo de um documento de texto em arquivos individuais a partir de um marco de referência (*milestone*) selecionado pelo usuário. A etiqueta `</doc>` foi estabelecida como o ponto de criação de um novo arquivo de texto. A Figura 12, a seguir, ilustra como a plataforma Lexos[®] segmenta os arquivos de textos.

Figura 12 – Segmentação dos textos pela plataforma Lexos[®]



Fonte: Lexos[®]. Acesso em 08 de maio, 2021.

Cada arquivo foi extraído contendo algumas informações contextuais do Sketch Engine[®] no cabeçalho (*header*) de cada texto, conforme é apresentado na Figura 13.

Figura 13 – Modelo de cabeçalho do Sketch Engine[®] em SGML

```
<docid="file20893928" filename="transcript" parent_folder="Politics"
url="https://www.ted.com/talks/amber_mcreynolds_an_election_system_th
at_puts_voters_not_politicians_first/transcript">
```

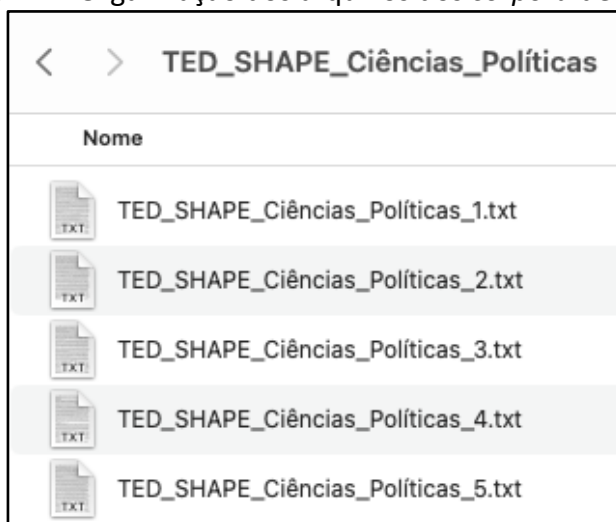
Fonte: Elaborado pelo autor.

Tais informações contextuais foram codificadas em marcação SGML para serem ignoradas pelo *software* AntConc[®], prevenindo assim possíveis equívocos na quantificação dos dados. Finalmente, os arquivos segmentados foram armazenados em uma pasta designada para cada disciplina e tiveram seus nomes padronizados para indicar três tipos de

⁵⁰ Disponível em <http://lexos.wheatoncollege.edu/upload>. Acesso em 08 de maio de 2021.

informações: nome do *corpus*, disciplina e número identificador do arquivo dentro da pasta, conforme é apresentado na Figura 14.

Figura 14 – Organização dos arquivos dos *corpora* de estudo



TED_SHAPE_Ciências_Políticas	
Nome	
	TED_SHAPE_Ciências_Políticas_1.txt
	TED_SHAPE_Ciências_Políticas_2.txt
	TED_SHAPE_Ciências_Políticas_3.txt
	TED_SHAPE_Ciências_Políticas_4.txt
	TED_SHAPE_Ciências_Políticas_5.txt

Fonte: Elaborado pelo autor

Com o *design*, compilação e organização dos *corpora* de estudo TED-STEM e TED-SHAPE concluídos, apresento, na seção seguinte, os detalhes relacionados à composição de tais *corpora*, a fim de proporcionar uma compreensão mais profunda dos aspectos metodológicos e teóricos envolvidos no processo de coleta e análise desses dados.

3.3 Descrição dos *corpora* TED-STEM e TED-SHAPE

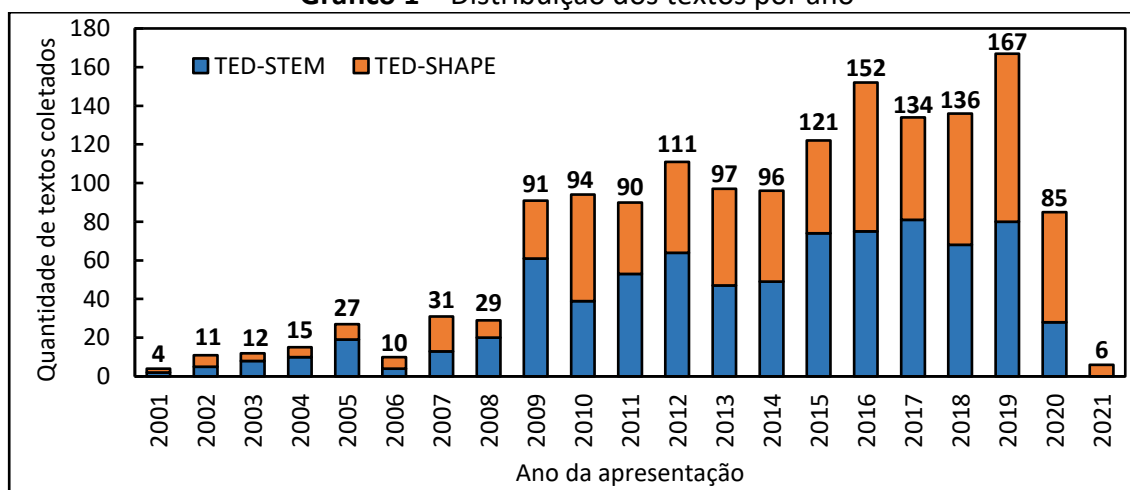
Para caracterizar os *corpora* de estudo compilados nesta investigação, decidi adotar a taxonomia descritiva proposta por Viana (2011), que resume bem os principais critérios a serem considerados no *design* inicial de qualquer *corpus* de estudo. Segundo o autor, *corpora* linguísticos podem ser descritos da seguinte forma:

Quadro 8 – Taxonomia de *corpora* linguísticos

Critérios		Corpus	
Abrangência		Geral Especializado	
Meio		Oral Escrito	
Tempo	Número de períodos históricos	Sincrônico Diacrônico	
	Relação com a atualidade	Contemporâneo Histórico	
Renovação		Dinâmico Estático	
Línguas	Número	Monolíngue Multilíngue	Paralelos Alinhados Não alinhados Não paralelos
	Produtores	De primeira língua (L1) De segunda língua (L2) De língua estrangeira (LE)	
Emprego		Estudo Referência	

Fonte: Viana (2011, p. 30)

No que se refere à abrangência e ao meio, os *corpora* de estudo desta tese são considerados de caráter especializados e orais, na medida em que representam apresentações orais em LI. No tocante ao aspecto temporal, o objetivo aqui não é examinar a língua em diferentes períodos (diacrônico), mas sim analisá-la em um espaço de tempo específico, ou seja, sincrônico. Desse modo, as TED Talks investigadas neste estudo ocorreram entre os anos de 2001 e 2021, caracterizando ambos os *corpora* de estudo como sincrônicos e contemporâneos. O Gráfico 1 abaixo apresenta a distribuição temporal dos textos coletados.

Gráfico 1 – Distribuição dos textos por ano

Fonte: Elaborado pelo autor

No tocante à renovação, os *corpora* de estudo foram compilados entre os meses de junho de 2020 e maio de 2021. Em virtude das atualizações periódicas de vídeos que o website recebe, em maio de 2021, efetuei uma verificação final para assegurar que os dados coletados corresponderiam às informações apresentadas no site⁵¹. Uma vez que não foram realizadas modificações nos *corpora* de estudo após maio de 2021, eles são classificados neste estudo como estáticos. Embora os textos sejam monolíngues (inglês), a lista de palestrantes de TED Talks é extremamente diversificada, e não é possível determinar com precisão se são falantes de inglês como L1, L2 ou LE. Contudo, essa distinção não é relevante no presente estudo, uma vez que o foco recai sobre o uso de pacotes lexicais em diferentes domínios disciplinares, independentemente se o palestrante possui o inglês como língua nativa ou adicional.

Por último, apresento os dados quantitativos das partes integrantes dos *corpora* TED-STEM e TED-SHAPE. A Tabela 3 exibe a quantidade de textos, juntamente com o número de formas (*types*) e itens (*tokens*) de cada disciplina do *corpus* TED-STEM.

Tabela 3 – Distribuição do *corpus* TED-STEM

Disciplinas	Textos	Formas (<i>Types</i>)	Itens (<i>Tokens</i>)
Agronomia	19	4.333	32.176
Astronomia	46	6.882	90.905
Biologia	142	15.283	290.338
Biotecnologia	31	6.130	62.605
Botânica	8	2.460	14.495
Design	28	6.261	56.041
Ecologia	27	6.322	58.535
Engenharia	78	9.883	149.645
Física	46	6.621	88.682
Genética	25	5.587	57.866
Geologia	5	2.117	10.454
Inteligência Artificial	15	3.957	31.414
Matemática	36	6.576	73.528
Medicina	124	12.520	223.228
Psicologia	146	15.694	332.730
Química	20	4.590	37.832
Total:	796	34.108 ⁵²	1.610.474

Fonte: Elaborado pelo autor.

⁵¹ Tal processo foi necessário ao perceber que algumas etiquetas haviam sido retiradas ou incluídas entre o intervalo de coleta dos textos.

⁵² Número de *types* refere-se ao *corpus* TED-STEM como um todo e não a somatória de *types* nas disciplinas individuais.

A Tabela 4, por sua vez, contém a quantidade de textos, juntamente com o número de itens e formas de cada disciplina do *corpus* TED-SHAPE.

Tabela 4 – Distribuição do *corpus* TED-SHAPE

Disciplinas	Textos	Formas (Types)	Itens (Tokens)
Antropologia	25	6.642	55.117
Arquitetura	94	12.442	191.961
Artes	44	8.257	78.717
Ciências Políticas	81	10.913	159.946
Desenvolvimento Internacional	18	4.361	35.944
Direito	66	9.914	137.720
Economia	94	11.737	201.246
Educação	50	7.282	92.952
Estudos do Gênero	29	6.310	62.439
Estudos de Mídia	22	5.150	41.302
Estudos Linguísticos	28	5.463	53.807
Filosofia	28	5.964	58.834
Gestão de Pessoas	41	6.892	77.973
História	14	4.215	28.622
Jornalismo	24	5.358	38.545
Literatura	14	4.785	28.699
Música	38	6.264	54.137
Sociologia	13	4.353	32.829
Total:	723	34.094 ⁵³	1.430.790

Fonte: Elaborado pelo autor.

Após concluir a descrição dos processos envolvidos na compilação dos *corpora* de estudo, a seção subsequente apresentará os procedimentos de análise empregados por meio do *software* AntConc®, ferramenta principal na análise dos *corpora* de estudo.

3.4 Ferramentas computacionais adotadas para exploração dos *corpora* de estudo

Nesta seção, apresento as funcionalidades do AntConc® empregadas no presente estudo. A versão adotada nesta tese é a 4.2.0. *Windows* (Anthony, 2020a). Algumas características notáveis deste *software* incluem:

⁵³ Número de *types* refere-se ao *corpus* TED-SHAPE como um todo e não a somatória de *types* nas disciplinas individuais.

- disponibilidade gratuita;
- consumo reduzido de espaço e memória do computador (190 MB);
- oferta do conjunto de ferramentas necessárias para realizar buscas e cálculos estatísticos pertinentes aos objetivos desta tese;
- interface relativamente simples e intuitiva, facilitando a alternância entre ferramentas com apenas alguns cliques;
- compatibilidade com os sistemas operacionais *Windows*®, *Mac OS X*® e *Linux*®;
- ausência de necessidade de conexão com a Internet para utilização do *software*.

No que concerne à extração e análise de pacotes lexicais, as principais ferramentas utilizadas foram a *N-Gram* e *Concordance*. Portanto, tais funcionalidades são descritas com maiores detalhes nos parágrafos subsequentes.

A Figura 15 ilustra o modo com as informações são exibidas ao usuário. Para obter a relação de pacotes lexicais utilizando o AntConc®, é preciso inicialmente selecionar todos os arquivos a serem analisados. Em seguida, o usuário deve clicar na aba *N-Gram* (A) e configurar parâmetros como tamanho da sequência (B), frequência mínima (C) e dispersão nos textos (D), desta forma, o programa produzirá listas de multipalavras conforme os critérios estabelecidos pelo usuário (E).

Figura 15 – Ferramenta N-Gram

The screenshot shows the AntConc N-Gram tool interface. The 'N-Gram' tab is selected. The main window displays a table of N-gram types. The table has columns for Rank, Freq, and Range. The first 23 items are listed, starting with 'to be able to' at rank 1 and frequency 424. Below the table, the search filters are visible: 'N-Gram Size' is set to 4, 'Min. Freq' is set to 30, and 'Min. Range' is set to 25. The 'Search Query' field is empty, and the 'Sort by' dropdown is set to 'Frequency'.

	Type	Rank	Freq	Range
1	to be able to	1	424	263
2	in the united states	2	336	201
3	at the same time	3	294	215
4	one of the most	4	274	221
5	at the end of	5	249	195
6	the end of the	6	245	190
7	it turns out that	7	239	161
8	if you want to	8	230	171
9	if you look at	9	221	155
10	for the first time	10	219	171
11	i want you to	11	202	134
12	a lot of people	12	200	161
13	is one of the	13	195	169
14	a little bit of	14	184	148
15	when it comes to	15	183	138
16	is going to be	16	179	139
17	the rest of the	17	174	138
18	in the middle of	18	173	144
19	all over the world	19	170	129
20	one of the things	20	167	121
21	and you can see	21	161	120
22	you look at the	22	155	120
23	thank you very much	23	154	141

Fonte: Elaborado pelo autor.

No presente estudo, o AntConc® foi configurado para extrair apenas pacotes lexicais formados por quatro palavras, com frequência relativa mínima de 20 ocorrências por milhão de palavras, e que apresentassem dispersão mínima de 4% nos textos dos *corpora*⁵⁴. É importante destacar que, neste estudo, o AntConc® foi configurado para contabilizar ocorrências como *don't* e *won't* como uma única palavra, em vez das formas separadas *do not* e *will not*, respectivamente.

Ao clicar em um dos pacotes lexicais na lista apresentada, o AntConc® mudará automaticamente para a tela da ferramenta *Concordance* (cf. Figura 16) e exibirá as linhas de concordância do item selecionado. Esse modo de visualização é chamado de *Key Word in Context*, KWIC, e possibilita ao usuário extrair, observar e estudar padrões sintagmáticos

⁵⁴ Na próxima seção, serão detalhados critérios de extensão, frequência e dispersão estabelecidos para extrair os pacotes lexicais dos *corpora* TED-STEM e TED-SHAPE.

(horizontalmente – à esquerda e à direita) e paradigmáticos (verticalmente) do pacote lexical escolhido. A utilização de linhas de concordância representa uma das principais abordagens na análise de *corpora*, pois, frequentemente, essas linhas servem como ponto de partida para a condução de outros tipos de investigações linguísticas no *corpus* em questão.

Figura 16 – Visualização da Key Word in Context pela ferramenta Concordance

	File	Left Context	Hit	Right Context
1	TED_STEM_Biolo...	and they've come from a common ancestry. But	if you look at	nucleotide diversity, which is more genetics that are passed
2	TED_STEM_Biolo...	ough the parents, you see that basking sharks,	if you look at	the first study, was order of magnitude less diverse even tha
3	TED_STEM_Biolo...	nd this has caused a very low diversity. And yet,	if you look at	the whale shark, which is the other plankton-eating large sh
4	TED_STEM_Biolo...	s illegal. They actually did get a fine for that. So	if you look at	all those studies I showed you, the total number of samples
5	TED_STEM_Biolo...	hey found that the diversity was incredibly low.	if you look at	the first line, really, you can see that all these different shark
6	TED_SHAPE_Eco...	illy ever give you an invoice? Or for that matter,	if you look at	the genetic level, 60 percent of medicines were prospected,
7	TED_SHAPE_Eco...	ie. They were the ones that didn't get paid. And	if you look at	the species level, you saw about fish. Today, the depletion o
8	TED_SHAPE_Eco...	nd convert them to shrimp farms. But of course,	if you look at	exactly what those profits are, almost 8,000 of those dollars
9	TED_STEM_Biolo...	a human breast cancer patient. It turns out that	if you look at	cancers, if you look at the immune system, if you look at agi
10	TED_STEM_Biolo...	patient. It turns out that if you look at cancers,	if you look at	the immune system, if you look at aging, if you look at devel

Fonte: Elaborado pelo autor.

Em síntese, a ferramenta AntConc® demonstra-se como um recurso valioso e versátil para análise de *corpora* linguísticos, proporcionando uma gama de funcionalidades que permitem a realização de buscas, extração de pacotes lexicais e análises estatísticas de modo eficiente e intuitivo. A descrição detalhada das principais características e funcionalidades do software apresentadas nesta seção visa fornecer um panorama sobre as possibilidades de análise e os métodos empregados nesta tese. A utilização do AntConc® na análise dos *corpora* TED-STEM e TED-SHAPE possibilita um maior entendimento das nuances linguísticas presentes nos diferentes domínios disciplinares e contribui para aprofundar o conhecimento acerca das especificidades do discurso nas TED Talks.

3.5 Critérios para extração dos pacotes lexicais

Embora os pacotes lexicais possam abranger um número diverso de palavras, frequentemente se observa na literatura trabalhos que se concentram em sequências de quatro palavras (Biber; Barbieri, 2007; Hyland, 2008a; Liu; Chen, H., 2020a; 2020b; Nesi; Basturkmen, 2006; Reppen; Olson, 2020). Tal escolha fundamenta-se nos estudos iniciais conduzidos por Biber *et al.* (1999), nos quais se constatou que sequências de três palavras,

embora numerosas, não apresentam natureza frasal. Por outro lado, sequências com cinco ou mais palavras são consequentemente mais frasais, porém pouco frequentes. Em outras palavras, há uma relação inversamente proporcional⁵⁵ entre o tamanho do pacote lexical e sua frequência no *corpus*. Diante disso, optei por analisar exclusivamente pacotes lexicais formados por quatro palavras, uma vez que, em geral, possibilitam a extração de uma quantidade adequada de dados e abarcam muitos dos pacotes lexicais de três palavras em suas estruturas.

Após a definição do tamanho dos pacotes lexicais a serem analisados, o passo subsequente consistiu em determinar a frequência mínima que uma sequência deveria apresentar para ser selecionada como um pacote lexical nos *corpora* de estudo. Entretanto, o critério de frequência mínima tende a ser arbitrário e estabelecido conforme o escopo de cada pesquisa. Conforme mencionado anteriormente, Biber *et al.* (1999) optaram por adotar um ponto de corte de dez ocorrências por um milhão de palavras; Hyland (2008a) e Biber, Conrad e Cortes (2004), por sua vez, estabeleceram cortes mais elevados de 20 e 40 ocorrências por um milhão de palavras, respectivamente. Com base na discussão supracitada e reconhecendo que as dimensões dos *corpora* TED-STEM e TED-SHAPE são relativamente pequenas para os padrões atuais, decidi por empregar um critério de frequência moderado de 20 ocorrências por milhão de palavras, similar aos estudos de Cortes (2004), Hyland (2008a) e Hyland e Jiang (2018).

Paralelamente à frequência, é necessário também que os pacotes lexicais estejam distribuídos em um número mínimo de textos a fim de evitar idiosincrasia. De maneira similar aos critérios de frequência, a dispersão mínima que um pacote lexical precisa apresentar em um *corpus* também aparenta ser arbitrária. Em geral, números absolutos são empregados para determinar o corte mínimo de dispersão. Biber *et al.* (1999) e Biber, Conrad e Cortes (2004), por exemplo, optaram por sequências ocorrendo em, no mínimo, cinco textos, independentemente do tamanho do *corpus*.

Entretanto, selecionar de forma arbitrária um número absoluto (e baixo) de textos para determinar a dispersão dos pacotes lexicais em meus *corpora* de estudo seria pouco pertinente para o contexto desta pesquisa. De fato, Biber, Conrad e Cortes (2004) e Conrad e Biber (2004) reconhecem que aplicar um critério de dispersão de cinco textos não provoca

⁵⁵ Estatisticamente, é esperado que cada componente adicionado ao pacote lexical reduza o drasticamente o número de ocorrências no *corpus* (Biber, 2009).

nenhum efeito prático, uma vez que, mesmo pacotes lexicais menos frequentes em seus estudos (*i.e.* 40 ocorrências/milhão) aparecem em pelo menos 20 textos. Nesse sentido, a adoção de um critério de dispersão relativa, aparenta ser um método mais confiável para assegurar que um pacote lexical não esteja associado apenas a um grupo restrito de falantes.

A decisão de utilizar uma taxa percentual de dispersão é consistente com os estudos de Hyland (2008a) e Hyland e Jiang (2018), nos quais são adotados um ponto de corte de dispersão de 10% dos textos no *corpus*, e Liu e Chen, H. (2020a; 2020b) que consideraram que um pacote lexical deveria ocorrer em um mínimo de 5% dos textos no *corpus*. Neste estudo, optei por aplicar um critério de dispersão mínima de 4%, o que equivale a 32 textos no *corpus* TED-STEM e 29 textos no *corpus* TED-SHAPE.

Após ter determinado as diretrizes acerca do tamanho, frequência e dispersão necessários para extração dos pacotes lexicais, o próximo passo foi a adoção de uma taxonomia que descrevesse a estrutura interna dessas sequências. Para o contexto desta tese, adotei a taxonomia proposta por Biber, Conrad e Cortes (2004). Conforme explicado no capítulo anterior, a motivação para me basear em tal taxonomia se deu por ela ser comumente utilizada na literatura acadêmica e por oferecer uma categorização adaptável para pacotes lexicais em diversos registros linguísticos (*cf.* Capítulo 2, Seção 2.4). É importante destacar também que essa classificação foi decidida previamente à análise dos *corpora* de estudo (teoria → dados).

Após a escolha pela taxonomia de categorização estrutural, o passo subsequente foi a adoção de uma classificação para as funções discursivas desempenhadas pelos pacotes lexicais. Desta forma, os pacotes lexicais encontrados nos *corpora* de estudo TED-STEM e TED-SHAPE foram classificados de acordo com uma das três grandes categorias funcionais propostas por Biber, Conrad e Cortes (2004) e detalhadamente descritas no Capítulo 2, Seção 2.4.

No capítulo anterior, foram expostas as principais similaridades e diferenças entre as taxonomias funcionais propostas nos trabalhos de Biber, Conrad e Cortes (2004) e Hyland (2008a). Com base nela, acredito que a primeira proposta apresenta melhor adequação ao escopo desta pesquisa pelos seguintes motivos: (i) estar alinhada com a taxonomia estrutural adotada para classificar os pacotes lexicais; (ii) ter sido proposta com base em uma ampla variedade de registros linguísticos e não apenas em textos escritos.

3.6 Refinamento dos pacotes lexicais relacionados ao tópico e em contexto de sobreposição

Adotando os critérios de frequência relativa mínima de 20 ocorrências por milhão e dispersão de 4% entre os textos dos *corpora* de estudo, identifiquei, inicialmente, 158 pacotes lexicais distintos no *corpus* de estudo TED-STEM, totalizando 6.119 ocorrências por milhão de palavras, e 128 pacotes lexicais no *corpus* TED-SHAPE, com um total de 4.517 ocorrências por milhão de palavras. Todavia, a identificação automatizada de pacotes lexicais requer cautela do pesquisador, pois ela abre a possibilidade de os dados quantitativos extraídos estarem inflacionados devido a pacotes lexicais em contexto de sobreposição, algo raramente reportado na literatura (Chen, Y. 2009; Chen, Y.; Baker, 2010; Grondona, 2015).

Em seu estudo, Chen, Y. (2009) propõe a exclusão manual de dois tipos de pacotes lexicais: os *pacotes lexicais relacionados ao tópico*, e *pacotes lexicais em contexto de sobreposição*. A autora esclarece que o primeiro tipo inclui sequências dependentes do contexto ou que incorporam nomes próprios (e.g. *financial and non financial, in the UK and*), embora tais sequências possam aparecer com certa frequência dentro do *corpus*, elas não deveriam ser consideradas como uma estrutura pré-montada da língua, pois não realizam funções distintas e informam pouco sobre a produção linguística dos autores. No *corpus* TED-STEM, os pacotes lexicais *in the United States* e *of the United States* foram desconsiderados para o presente estudo por serem entendidas como expressões dependentes de contexto, já no *corpus* TED-SHAPE, os padrões desconsiderados foram: *in New York City, in the United States, of the United States* e *to the United States*.

O segundo tipo de exclusão manual abrange duas ou três sequências que, na realidade, fazem parte de um encadeamento maior e, ao serem consideradas como sequências distintas, elas acabam inflacionando os dados quantitativos. A fim de determinar se o número de ocorrências de um pacote lexical deveria ser refinado e minimizar a inflação dos dados extraídos de modo automatizado, empreguei a metodologia proposta por Chen, Y. (2009). A autora explica que a sobreposição de pacotes lexicais pode ocorrer de três formas:

- *Sobreposição completa*: ocorre quando duas sequências, muito similares, apresentam o mesmo número de ocorrências e as linhas de concordância revelam que elas derivam de um encadeamento maior (relação um-para-um). A título de exemplificação, a autora demonstra como as ocorrências das sequências *this may be due* e *may be due*

to não ocorrem isoladamente, e, na verdade, elas fazem parte do encadeamento maior *this may be due to*.

- *Subsunção completa*: acontece quando há sobreposição de duas sequências, mas suas respectivas ocorrências no *corpus* são distintas. No estudo de Chen, Y. (2009), a sequência *as a result of* ocorreu 17 vezes, enquanto *a result of the* ocorreu apenas 5 vezes. Nestes casos, a sequência com menor frequência (*i.e.* 5) é incorporada e apenas a maior ocorrência é contabilizada da seguinte forma: *as a result of + (the)* com 17 ocorrências.
- *Subsunção parcial*: verifica-se quando a sobreposição de duas sequências ocorre de modo parcial. O número de ocorrências do encadeamento maior é descontado do conjunto de sequências menores que ele incorpora, a sequência maior é utilizada e seu número de ocorrências é refinado. No estudo de Chen, Y. (2009), por exemplo, as sequências *the end of the* e *at the end of* ocorreram 10 e 6 vezes, respectivamente, enquanto o encadeamento *at the end of the* ocorreu apenas 3 vezes. Nesse caso, Chen, Y. (2009) propõe que a soma dos conjuntos menores seja reduzida da sequência maior, ou seja, $10 + 6 - 3$, após isso, as sequências devem ser combinadas (*i.e.* *(at) + the end of the*) e sua ocorrência refinada para 13.

As Tabelas 5 e 6, a seguir, ilustram os pacotes lexicais em contexto de sobreposição que tiveram as suas frequências relativas (FR) refinadas (por milhão de palavras) para o presente o estudo.

Tabela 5 – Pacotes lexicais em contexto de sobreposição no *corpus* TED-STEM

(continua)

# palavras	Pacotes lexicais sobrepostos					FA	FR	DiA	DiR%	Encadeamento lexical refinado	FA	FR	DiA	DiR%
4	what	I	want	to		77	48	65	8,1%					
4		I	want	to	do	37	23	33	4,1%	what I want to + (do)	89	55	74	9,3%
5	what	I	want	to	do	25	16	24	3%					
4		we	need	to	do	53	33	44	5,5%					
4	what	we	need	to		37	23	34	4,2%	(what) + we need to do	64	40	55	6,9%
5	what	we	need	to	do	26	16	23	2,8%					
4	if	you	look	at		135	84	86	10,8%					
4		you	look	at	the	92	57	75	9,4%	if you look at + (the)	161	100	109	13,6%
5	if	you	look	at	the	66	41	52	6,5%					
4		going	to	show	you	98	61	70	8,7%					
4	I'm	going	to	show		91	57	67	8,4%	(I'm) + going to show you	108	67	78	9,8%
5	I'm	going	to	show	you	81	50	59	7,4%					
4	I	want	to	show		54	34	39	4,9%					
4		want	to	show	you	51	32	36	4,5%	I want to show + (you)	56	35	40	5%
5	I	want	to	show	you	49	30	35	4,4%					
4	to	be	able	to		294	183	168	21,1%					
4		be	able	to	do	41	25	35	4,4%	to be able to + (do)	308	191	180	22,6%
5	to	be	able	to	do	27	17	23	2,8%					
4	I	want	to	talk		75	47	53	6,6%					
4		want	to	talk	about	45	28	32	4%	I want to talk + (about)	79	49	56	7%
5	I	want	to	talk	about	41	25	29	3,6%					
4	to	talk	to	you		62	38	53	6,6%					
4		talk	to	you	about	47	29	40	5%	to talk to you + (about)	65	40	55	6,9%
5	to	talk	to	you	about	44	27	38	4,7%					
4		going	to	tell	you	63	39	54	6,7%					
4	I'm	going	to	tell		52	32	45	5,6%	(I'm) + going to tell you	65	40	55	6,9%
5	I'm	going	to	tell	you	50	31	44	5,5%					

(conclusão)

# palavras	Pacotes lexicais sobrepostos					FA	FR	DiA	DiR%	Encadeamento lexical refinado	FA	FR	DiA	DiR%
4	I	want	to	tell		55	34	45	5,6%	I want to tell + (you)	57	35	47	5,9%
4		want	to	tell	you	52	32	43	5,4%					
5	I	want	to	tell	you	50	31	41	5,1%					
4		going	to	talk	about	59	37	48	6%	(I'm) + going to talk about	78	48	64	8%
4	I'm	going	to	talk		57	35	45	5,6%					
5	I'm	going	to	talk	about	38	24	29	3,6%					
4		it	turns	out	that	183	114	116	14,5%	(and) + it turns out that	212	132	138	17,3%
4	and	it	turns	out		67	42	55	6,9%					
5	and	it	turns	out	that	38	24	33	4,1%					
4	let	me	give	you		55	34	47	5,9%	let me give you + (an example)	77	48	68	8,5%
4			give	you	an example	40	25	39	4,9%					
6	let	me	give	you	an example	18	11	18	2,2%					
4	if	you	think	about		53	33	46	5,7%	if you think about + (it)	71	44	61	7,6%
4		you	think	about	it	42	26	38	4,7%					
5	if	you	think	about	it	24	15	23	2,8%					
4	one	of	the	things		108	67	76	9,5%	one of the things + (that)	132	82	97	12,1%
4		of	the	things	that	72	45	58	7,2%					
5	one	of	the	things	that	48	30	37	4,6%					
4		is	one	of	the	111	69	93	11,6%	(this) + is one of the	132	82	109	13,6%
4	this	is	one	of		69	43	55	6,9%					
5	this	is	one	of	the	48	30	39	4,9%					
4	there	are	a	lot		38	24	36	4,5%	there are a lot + (of)	38	24	36	4,5%
4		are	a	lot	of	36	22	34	4,2%					
5	there	are	a	lot	of	36	22	34	4,2%					
4	at	the	end	of		123	76	98	12,3%	at the end of + (the)	169	105	134	16,8%
4		the	end	of	the	118	73	94	11,8%					
5	at	the	end	of	the	72	45	58	7,2%					
4	in	the	middle	of		94	58	78	9,8%	in the middle of + (the)	104	65	85	10,6%
4		the	middle	of	the	58	36	47	5,9%					
5	in	the	middle	of	the	48	30	40	5%					

Fonte: Elaborado pelo autor

Tabela 6 – Pacotes lexicais em contexto de sobreposição no *corpus* TED-SHAPE

(continua)

# palavras	Pacotes lexicais sobrepostos					FA	FR	DiA	DiR%	Encadeamento lexical refinado	FA	FR	DiA	DiR%
4		going	to	have	to	57	40	32	4,4%	(we're) + going to have to	73	51	44	6%
4	we're	going	to	have		48	34	30	4,1%					
5	we're	going	to	have	to	32	22	18	2,4%					
4	if	you	look	at		86	60	69	9,5%	if you look at + (the)	110	77	85	11,7%
4		you	look	at	the	63	44	45	6,2%					
5	if	you	look	at	the	39	27	29	4%					
4		is	going	to	be	81	57	61	8,4%	(this) + is going to be	99	69	78	10,7%
4	this	is	going	to		34	24	31	4,2%					
5	this	is	going	to	be	16	11	14	1,9%					
4	I	want	to	talk		68	48	55	7,6%	I want to talk + (about)	80	56	65	8,9%
4		want	to	talk	about	52	36	42	5,8%					
5	I	want	to	talk	about	40	28	32	4,4%					
4	I	want	to	tell		43	30	34	4,7%	I want to tell + (you)	45	31	36	4,9%
4		want	to	tell	you	41	29	34	4,7%					
5	I	want	to	tell	you	39	27	32	4,4%					
4		going	to	talk	about	35	24	32	4,4%	(I'm) + going to talk about	48	34	43	5,9%
4	I'm	going	to	talk		34	24	30	4,1%					
5	I'm	going	to	talk	about	21	15	19	2,6%					
4	if	you	think	about		62	43	46	6,3%	if you think about + (it)	77	54	59	8,1%
4		you	think	about	it	44	31	37	5,1%					
5	if	you	think	about	it	29	20	24	3,3%					
4	let	me	give	you		46	32	43	5,9%	let me give you + (an example)	61	43	55	7,6%
4			give	you	an example	32	22	28	3,8%					
6	let	me	give	you	an example	17	12	16	2,2%					
4		is	one	of	the	84	59	76	10,5%	(this) + is one of the	97	68	87	12%
4	this	is	one	of		33	23	30	4,1%					
5	this	is	one	of	the	20	14	19	2,6%					

(conclusão)

# palavras	Pacotes lexicais sobrepostos					FA	FR	DiA	DiR%	Encadeamento lexical refinado	FA	FR	DiA	DiR%
4	one	of	the	things		59	41	45	6,2%	one of the things + (that)	80	56	61	8,4%
4		of	the	things	that	49	34	40	5,5%					
5	one	of	the	things	that	28	20	24	3,3%					
4	there	are	a	lot		29	20	29	4%	there are a lot + (of)	31	22	30	4,1%
4		are	a	lot	of	30	21	29	4%					
5	there	are	a	lot	of	28	20	28	3,8%					
4	for	the	first	time		105	73	87	12%	for the first time + (in)	113	78	93	12,8%
4		the	first	time	in	35	24	32	4%					
5	for	the	first	time	in	27	18	26	3,5%					
4		the	end	of	the	127	89	96	13,2%	(at) + the end of the	191	133	141	19,5%
4	at	the	end	of		126	88	97	13,4%					
5	at	the	end	of	the	62	43	52	7,1%					
4	in	the	middle	of		79	55	66	9,1%	in the middle of + (the)	85	59	72	9,9%
4		the	middle	of	the	49	34	46	6,3%					
5	in	the	middle	of	the	43	30	40	5,5%					

Fonte: Elaborado pelo autor

É interessante observar que não houve nenhum caso de sobreposição completa, ou seja, uma relação um-para-um, nos *corpora* de estudo e apenas um caso de subsunção completa no *corpus* TED-STEM, em que as 36 ocorrências do padrão *are a lot of* estavam abrangidas dentro das 38 ocorrências do pacote lexical *there are a lot*, resultando assim no encadeamento *there are a lot + (of)*. Os demais padrões apresentados nas Tabelas 5 e 6 refletem situações de subsunção parcial – apenas uma fração das ocorrências de um pacote lexical foi abrangida por outro. Por exemplo, no *corpus* TED-SHAPE, a frequência da sequência de cinco palavras *we're going to have to* precisou ser ajustada para evitar inflação na contabilização dos dados. As 32 ocorrências absolutas dessa sequência foram subtraídas da soma de dois componentes menores: *going to have to* (FA= 57) e *we're going to have* (FA= 48), respectivamente, resultando na frequência refinada de 73 para o encadeamento (*we're*) + *going to have to* (FA = 57 + 48 – 32).

Após o refinamento dos dados, o conjunto de dados no *corpus* TED-STEM diminuiu de 158 para 138 e totalizou 5.483 ocorrências por milhão de palavras (redução de 11,5% em relação à extração automática). Já o conjunto de 128 pacotes lexicais distintos no *corpus* TED-SHAPE diminuiu para 114, resultando em 3.992 ocorrências por milhão de palavras (redução de 13% em relação à extração automática). A relação final dos pacotes lexicais extraídos, incluindo suas respectivas frequências e dispersões, estão detalhadas nos Apêndices A e B.

3.7 Material e métodos: Considerações finais

Este capítulo concentrou-se na apresentação e descrição dos materiais e métodos empregados na presente investigação. Inicialmente, abordei os principais motivos que levaram à compilação dos *corpora* TED-STEM e TED-SHAPE, bem como os procedimentos adotados na organização dos metadados e na coleta dos textos. Apesar dos avanços tecnológicos, o *design* e compilação de um *corpus* de estudo ainda representam um processo meticuloso (Viana; O'Boyle, 2022) e demandaram um tempo considerável para sua construção. Contudo, é possível afirmar que o esforço foi recompensado, visto que os *corpora* TED-STEM e TED-SHAPE possibilitaram a análise de pacotes lexicais em TED Talks sob uma perspectiva disciplinar, aspecto que, até onde foi possível verificar, não recebeu devida atenção em outros estudos. Após a descrição dos *corpora* de estudo, expliquei como a seleção

e refinamento dos pacotes lexicais foram conduzidos e quais taxonomias estruturais e funcionais foram adotadas para categorizá-los.

Em síntese, a compilação dos *corpora* de estudo seguiu as etapas a seguir:

- i. *Identificação das TED Talks associadas às disciplinas acadêmicas*: seleção das etiquetas utilizadas no site oficial e atribuição, por parte do pesquisador, aos domínios disciplinares STEM ou SHAPE.
- ii. *Seleção das TED Talks elegíveis para compor os corpora de estudo*: palestras sem a transcrição oficial, produzidas em idiomas que não o inglês, animações ou que fossem organizadas em forma de debates ou entrevistas não foram selecionadas para o presente estudo.
- iii. *Organização dos metadados*: informações adicionais acerca dos textos coletados (autores, título da palestra, ano de produção, dentre outros) foram salvas em planilha do Microsoft Excel® a fim de evitar duplicação de textos e facilitar a filtragem de diversas informações sempre que necessário.
- iv. *Extração dos textos e codificação SGML*: utilização da ferramenta *WebBootCaT* do *Sketch Engine*® para extração das transcrições das TED Talks selecionadas a partir de suas respectivas URLs. Em seguida, informações não linguísticas presentes nas transcrições foram ocultadas manualmente utilizando codificação SGML. Por último, as transcrições foram segmentadas em arquivos de textos individuais e salvas em formato *.txt*.
- v. *Determinação dos critérios para extração dos pacotes lexicais*: após a compilação dos *corpora* de estudo, o *software* AntConc® foi utilizado para extrair as sequências fixas de quatro palavras, com frequência mínima de 20 ocorrências por milhão de palavra e dispersas em, pelo menos, 4% dos textos em cada *corpus*.
- vi. *Escolha da taxonomia para análise dos dados*: Por último, decidiu-se que a taxonomia proposta por Biber, Conrad e Cortes (2004) seria aplicada tanto para análise estrutural quanto funcional dos pacotes lexicais extraídos.
- vii. *Refinamento de pacotes lexicais em contexto de sobreposição e dependentes de contexto*: segundo a proposta de Chen, Y. (2009), Chen, Y. e Baker (2010) e Grondona (2015).

Tendo apresentado e discutido os materiais e escolhas metodológicas empregados nesta tese, abordo, no capítulo seguinte, os resultados encontrados durante a análise dos dados extraídos dos *corpora* de estudo.

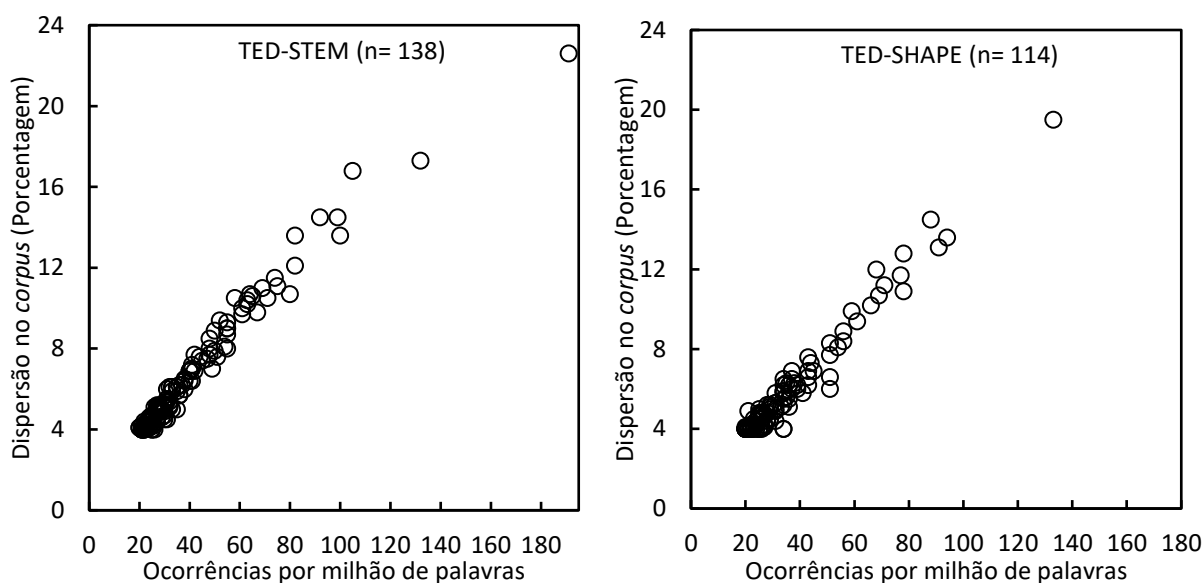
4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Neste capítulo, a ênfase recai sobre a análise dos pacotes lexicais derivados dos *corpora* de estudo, compilados a partir de TED Talks vinculadas aos domínios acadêmicos STEM e SHAPE, conforme delineado no capítulo anterior. Com o propósito de alcançar tal objetivo, na Seção 4.1, abordo as variações pertinentes às estruturas gramaticais e recorrência de pacotes lexicais distintos nos *corpora* de estudo em questão. Visando debater de forma mais detalhada as funções desempenhadas no discurso, as classificações funcionais sugeridas por Biber, Conrad e Cortes (2004) estão apresentadas em seções separadas. Na Seção 4.2, discuto os pacotes lexicais que expressam posicionamento, enquanto os que organizam o discurso são contemplados na Seção 4.3 e os que expressam referências são examinados na Seção 4.4. As implicações decorrentes dos dados obtidos são apresentadas e discutidas ao final de cada uma dessas três respectivas seções.

4.1 Diferenças na recorrência e estrutura dos pacotes lexicais extraídos dos *corpora* TED-STEM e TED-SHAPE

Considerando as dimensões similares dos *corpora* de estudo no tocante à quantidade de textos, itens e formas; é interessante investigar o porquê de haver um conjunto maior de pacotes lexicais em TED Talks associadas às áreas STEM em comparação com a contraparte SHAPE. Para tanto, o Gráfico 2 a seguir, compara os pacotes lexicais extraídos de ambos os *corpora* de estudo no âmbito de suas frequências relativas e dispersão nos textos.

Gráfico 2 – Distribuição da FR e dispersão dos dados entre os *corpora* de estudo



As informações apresentadas acima contribuem para uma compreensão aprimorada de como os pacotes lexicais variam em termos de frequência relativa e dispersão entre os *corpora* de estudo. É notável, por exemplo, que em ambos os *corpora* de estudo, a maior parte dos pacotes lexicais ocorrem entre 20-40 vezes por milhão de palavras e estão dispersos em até 8% dos textos. É possível perceber, no entanto, uma concentração discretamente maior de pacotes lexicais com frequência relativa acima de 60 vezes por milhão de palavras no *corpus* TED-STEM, sinalizando uma maior recorrência de expressões convencionalizadas no discurso das TED Talks associadas a essa área.

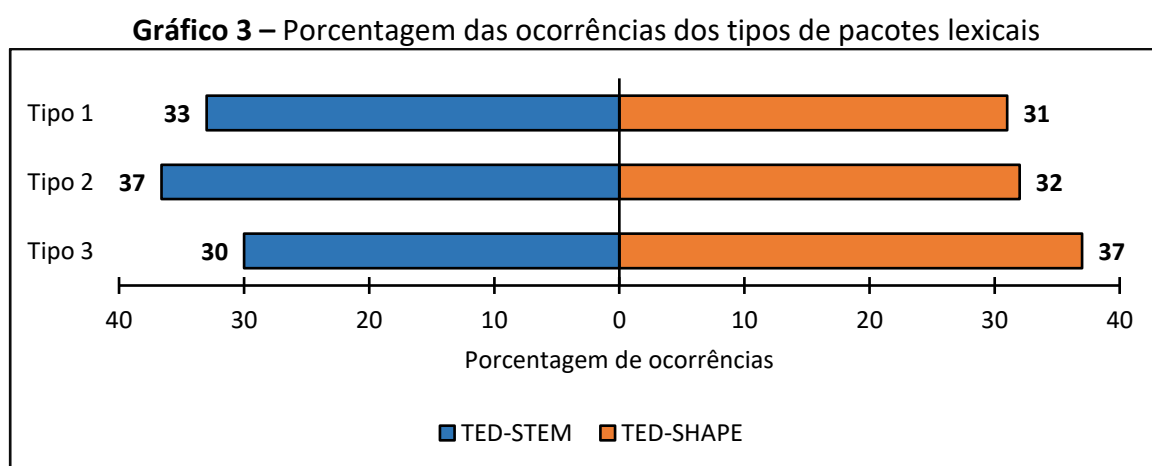
Uma possível explicação para isso poderia estar relacionada à natureza disciplinar das *hard sciences*, que, por lidarem com estudos de leis universais, tendem a possuir fundamentos teóricos rígidos e menor competição interna de seus paradigmas (Becher; Trowler, 2001; Trowler, 2012; 2014), o que se refletiria em um maior consenso discursivo entre os membros da comunidade disciplinar e, conseqüentemente, em uma estrutura discursiva mais padronizada, repetitiva e generalizada. Por outro lado, o menor conjunto de pacotes lexicais encontrado no *corpus* TED-SHAPE é coerente com a maior diversidade lexical e subjetividade associada ao discurso das *soft sciences*. As divergências teóricas são mais evidentes entre os membros dessa comunidade/tribo, seu discurso apresentaria fundamentos teóricos menos rígidos, com objetos de estudos restritos ao contexto da pesquisa (Becher; Trowler, 2001;

Trowler, 2012; 2014). Consequentemente, o discurso de suas pesquisas tende a ser mais idiossincrático e interpretativo.

No tocante à categorização da estrutura interna dos pacotes lexicais, optei por adotar a taxonomia estrutural proposta por Biber, Conrad e Cortes (2004) e, conforme fora explanado anteriormente (*cf.* Capítulo 3, Seção 3.5), esses autores categorizam os pacotes lexicais em três tipos principais de estruturas, a saber:

- Tipo 1 – fragmentos de sintagmas verbais (SV) (*e.g. and I'm going to*);
- Tipo 2 – fragmentos de orações subordinadas (OS) (*e.g. you can see that*), e
- Tipo 3 – fragmentos de sintagmas nominais/preposicionais (SN/SP) (*e.g. at the same time*).

No que concerne à estrutura dos pacotes lexicais, o Gráfico 3 contrapõe a porcentagem de pacotes lexicais distintos identificados entre os *corpora* de estudo.



Fonte: Elaborado pelo autor.

Os dados acima evidenciam que pacotes lexicais do Tipo 1 (*e.g. and I'm going to*) ocorrem em proporções semelhantes em ambos os *corpora* de estudo, aproximadamente 30%. Por sua vez, os pacotes lexicais que incorporam orações subordinadas (Tipo 2) exibiram uma proporção ligeiramente maior no *corpus* TED-STEM, indicando uma padronização de verbos vinculados às estruturas infinitivas (*e.g. to come up with*) e *wh-clauses* (*e.g. what we do is*), por exemplo. De fato, o padrão mais recorrente no *corpus* TED-STEM foi o encadeamento *to be able to + (do)*, com uma frequência relativa de 191 ocorrências por

milhão de palavras, sugerindo um enfoque discursivo nos objetivos e ações realizadas pelos palestrantes.

Uma variação notável consiste na maior proporção de pacotes lexicais compostos por fragmentos SN/SP (*e.g. (at) + the end of the*) no *corpus* TED-SHAPE. Tais estruturas tendem a ser mais recorrentes em registros acadêmicos escritos (Biber, 2006a), indicando que palestrantes TED recorrem a estruturas tanto de registros orais quanto escritos, similar ao que ocorre em aulas expositivas.

Levando em consideração que pacotes lexicais de natureza verbal tendem a ser mais prevalentes em registros orais da língua (Biber, 2006a), a maior proporção destes em TED-STEM poderia ser associada a um discurso mais centrado na interação orador e público-alvo (*e.g. you can see the*) e na fragmentação do conteúdo informativo, possivelmente visando facilitar a compreensão do público-alvo. Já a elevada proporção de pacotes lexicais de natureza nominal em TED-SHAPE aponta para uma linguagem possivelmente mais formal e elaborada, com maior integração e densidade informacional do discurso apresentado ao público-alvo (*e.g. one of the most*) e na manutenção da coesão lógica do texto (*e.g. at the same time*), similar ao uso de pacotes lexicais nos registros escritos.

Nas próximas seções, apresento e discuto os resultados da análise funcional dos pacotes lexicais encontrados nos *corpora* de estudo, seguindo a taxonomia funcional proposta por Biber, Conrad e Cortes (2004).

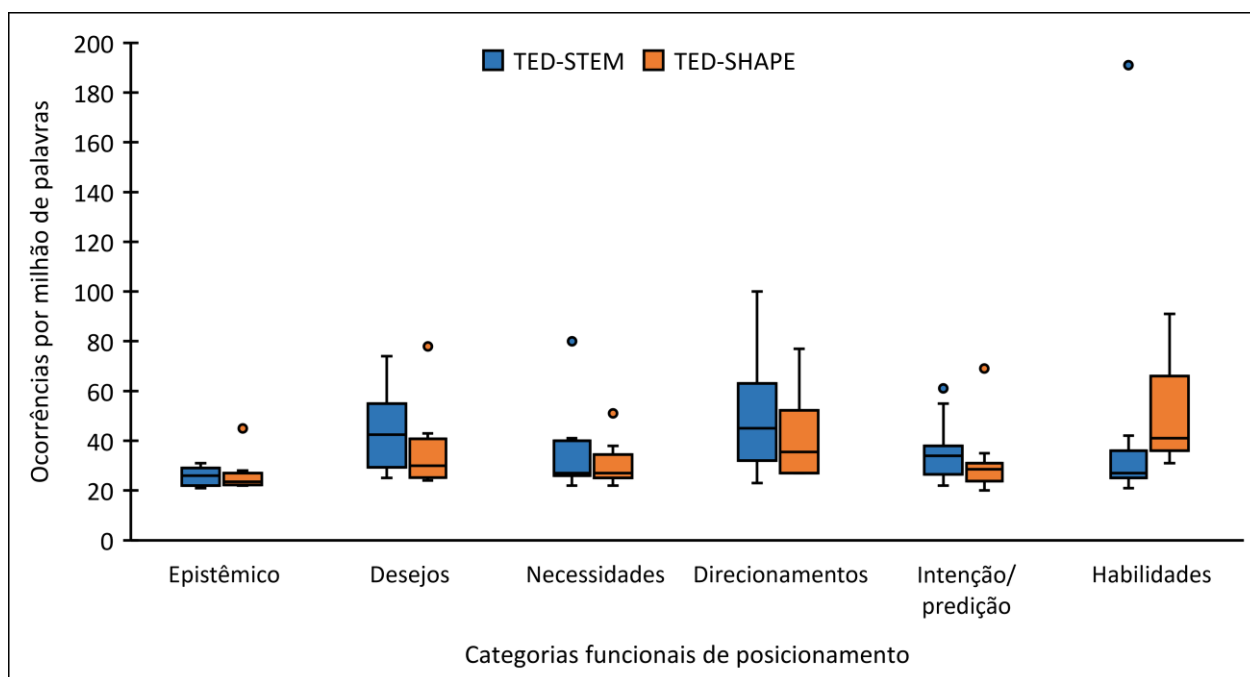
4.2 Pacotes lexicais de expressão de posicionamento

Considerando que as relações sociais entre os membros de uma comunidade acadêmica são reconhecidas, construídas e negociadas por meio do discurso (Hyland, 2005), nesta seção, abordo os pacotes lexicais mais recorrentes entre TED Talks das áreas STEM e SHAPE para expressarem posicionamento, ou seja, sinalizarem credibilidade, promoverem engajamento, apresentarem avaliações e pontos de vistas pessoais, tornando os argumentos apresentados mais convincentes perante o público-alvo (Biber, 2006b).

Em relação ao aspecto de frequência, o Gráfico 4 apresenta uma perspectiva macro da distribuição das ocorrências dos pacotes lexicais da categoria de posicionamento entre os *corpora* de estudo. Tais dados, apresentam informações relevantes acerca da tendência central, indicado pela linha da mediana no centro do quadrado e dispersão do conjunto de

dados, representada pelo tamanho da caixa, entre TED Talks associadas às áreas STEM e SHAPE.

Gráfico 4 – Distribuição de ocorrências de pacotes lexicais de posicionamento entre os *corpora* de estudo



Fonte: Elaborado pelo autor.

Os dados acima evidenciam, por exemplo, uma baixa variação (tamanho da caixa) nos conjuntos de pacotes lexicais epistêmicos em ambos os *corpora* de estudo, apesar de haver um valor discrepante no *corpus* TED-SHAPE, há pouca variação na distribuição do restante do conjunto de pacotes lexicais identificados. As categorias de necessidades e intenção/predição apresentam pouca variação interna na distribuição de seus dados, especialmente no *corpus* TED-SHAPE. Contudo, valores discrepantes no conjunto de pacotes lexicais dessas categorias foram identificados em ambos os *corpora* de estudo.

Já o conjunto de expressões que denotam desejos e direcionamentos exibiram uma tendência central alta no *corpus* TED-STEM. Embora não tenham sido identificados valores discrepantes nessas categorias, a maior dispersão entre os valores máximos e mínimos (tamanho da caixa) sugere uma diferença maior na recorrência de certos padrões dentre os textos no *corpus* TED-STEM, quando comparados à contraparte SHAPE.

No *corpus* TED-SHAPE, a categoria de pacotes lexicais de habilidades apresentou a maior tendência central, porém há uma grande variação interna nas ocorrências deste

conjunto. Em contrapartida, a categoria de habilidades do *corpus* TED-STEM apresenta pouca variação interna nas ocorrências do conjunto, indicando que o conjunto de pacotes lexicais dessa categoria tende a ser empregado com frequência similar. Destaca-se, no entanto, o valor discrepante de um pacote lexical em relação ao conjunto de padrões dessa categoria.

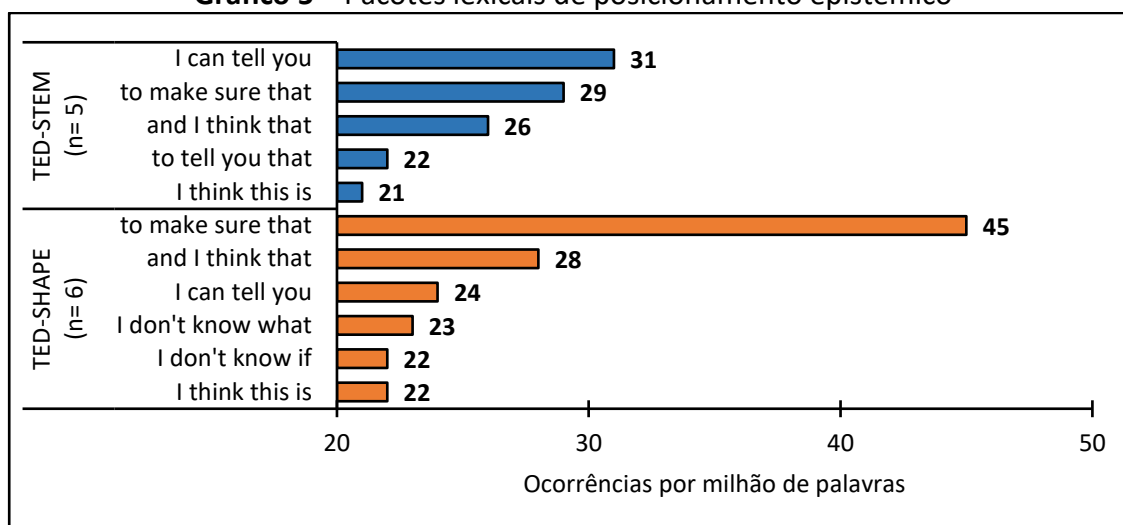
A presente seção está estruturada da seguinte forma: na Seção 4.2.1, conduzo uma análise mais detalhada dos pacotes lexicais de natureza epistêmica – aqueles que expressam o grau de assertividade do falante relação a seu discurso. Ao longo das Seções 4.2.2 – 4.2.6, trato das subcategorias de posicionamento atitudinal, isto é, as sequências utilizadas para expressar, respectivamente, desejos e motivações pessoais, necessidades/obrigações, direcionamentos, intenções/predições e habilidades. Por fim, na Seção 4.2.7, encerro com uma síntese das principais características discursivas das TED Talks STEM e SHAPE que foram evidenciadas por pacotes lexicais de posicionamento.

4.2.1 Posicionamento epistêmico

Nos parágrafos subsequentes, analiso detalhadamente os pacotes lexicais relacionados à função de posicionamento epistêmico, isto é, aqueles pacotes lexicais que palestrantes TED utilizam para sinalizar suas opiniões, certezas, dúvidas e avaliações acerca das informações apresentadas em seus discursos.

O Gráfico 5 exibe o conjunto de pacotes lexicais distintos, com frequência relativa acima de 20 ocorrências por milhão de palavras, encontrados em cada *corpus* de estudo. Em relação ao conjunto de dados, destaca-se a alta frequência relativa do padrão *to make sure that* no *corpus* TED-SHAPE quando comparado aos demais pacotes lexicais dessa categoria.

Gráfico 5 – Pacotes lexicais de posicionamento epistêmico



Fonte: Elaborado pelo autor

Ao abordar como os palestrantes de TED Talks demonstram conhecimento e se estabelecem como autoridades perante o público-alvo, o pacote lexical *I can tell you* revelou ser um dos padrões mais recorrentes em ambos os *corpora* de estudo. Em geral, o emprego desse pacote lexical auxilia os palestrantes no compartilhamento de experiências pessoais que promovam sua imagem de especialista e reforçam a credibilidade discursiva perante o público-alvo.

Nos exemplos a seguir, é possível observar como os palestrantes empregam o padrão *I can tell you* em conjunto com a apresentação de suas credenciais profissionais (*i.e. as a flag expert* e *I'm an economist*), a fim de fortalecer seus argumentos e persuadir o público-alvo de que possuem a credibilidade necessária para opinar sobre o assunto em questão.

[01] Now, as a flag expert, **I can tell you** this is officially ugly. OK? It's OK to laugh at this flag.
(TED_STEM_Design_01)

[02] So thank God, because I'm an economist, and **I can tell you**, value is not just price.
(TED_SHAPE_Economia_50)

O pacote lexical *to tell you that*, por sua vez, evidenciou uma maior assertividade discursiva por parte dos palestrantes STEM. No exemplo abaixo, é interessante observar como a referência à qualificação profissional (*i.e. heart specialist*) e a expressão *I am here today to* são utilizadas conjuntamente para sinalizar envolvimento pessoal e promover a imagem de autoridade do palestrante e reforçar a credibilidade de seus argumentos e afirmações frente

ao público-alvo. Nesse sentido, a função que pacote lexical *to tell you that* desempenha no discurso é semelhante ao uso do padrão *I can tell you*.

[03] But what is this link? Is it real or purely metaphorical?
As a heart specialist, I am here today **to tell you that** this
link is very real. Emotions, you will learn, can and do have
a direct physical effect on the human heart.
(TED_STEM_Medicina_101)

Embora a apresentação de credenciais profissionais possa conferir credibilidade de maneira mais assertiva, a maior parte das ocorrências de *I can tell you* e *to tell you that* aponta que palestrantes TED preferem manter suas credenciais profissionais implícitas no discurso e recorrem a esses pacotes lexicais para sinalizar que possuem vasta experiência tácita. Essa abordagem mais sutil na apresentação de experiência e conhecimento permite que os palestrantes estabeleçam sua autoridade e credibilidade sem parecerem pretensiosos ou condescendentes. No Exemplo 04, o uso da expressão *from my observations* junto ao pacote lexical *I can tell you* sugere o intuito do palestrante de evitar uma imagem arrogante perante o público-alvo, o que, conseqüentemente, poderia prejudicar o engajamento, conexão e proximidade com o discurso.

[04] And from my observations, **I can tell you**, those that eat
together more frequently, exhibit higher levels of
communitas.
(TED_SHAPE_Arquitetura_31)

O exemplo acima se alinha ao argumento de Caliendo (2014) de que, independentemente da autoridade ou prestígio que palestrantes TED possam ter, o uso de qualificações profissionais explícitas ocorre de forma mínima e instrumental em ambos os *corpora* de estudo. Em vez disso, eles optam por expressões mais dialógicas que estabelecem uma relação mais simétrica entre palestrantes e o público leigo, mesmo que de maneira implícita. Possivelmente, isso se deve ao entendimento de que as TED Talks são eventos voltados à divulgação e discussão de ideias potencialmente relevantes para a nossa sociedade e não para promoção acadêmica, profissional ou social do palestrante.

Outro padrão recorrente na promoção de credibilidade frente ao público-alvo nas TED Talks foi o uso do pacote lexical *to make sure that*. As ocorrências desse padrão revelam, no entanto, interesses distintos entre palestrantes das áreas STEM e SHAPE. No *corpus* TED-

STEM, observei que o emprego de *to make sure that* tende a identificar os objetivos e resultados pretendidos pelos palestrantes por meio de suas pesquisas e/ou projetos, conforme exemplificado abaixo:

[05] And the idea is not to cure autism. That's not the idea. What we want is **to make sure that** those individuals with autism can be free from the devastating consequences that come with it at times.

(TED_STEM_Psicologia_08)

Já os palestrantes SHAPE recorrem ao pacote lexical *to make sure that* não para explicitar objetivos desejáveis, mas para enfatizar ações que precisam ou devem ser realizadas para alcançar determinados fins.

[06] And we've got to do better. The evaluation, we have to start with kids earlier, we have **to make sure that** we provide the support to young people. We've got to give them all of these opportunities.

(TED_SHAPE_Educação_19)

O pacote lexical *to make sure that* também é empregado quando o palestrante busca explicitar ao público-alvo o motivo pelo qual certos procedimentos técnicos foram adotados. O exemplo a seguir ilustra como esse pacote lexical auxilia o palestrante na explicação da razão de ter sido utilizado um grupo de controle durante a realização de um experimento.

[07] We had to run a control group **to make sure that** coming into the laboratory didn't improve your Mandarin skills. So a group of babies came in and listened to English.

(TED_SHAPE_Estudos_Linguísticos_19)

Nesses contextos, o padrão *to make sure that* desempenha um papel importante na comunicação efetiva das intenções e justificativas metodológicas do palestrante, contribuindo para a compreensão e a transparência do processo pelos ouvintes. Em tais casos, a função discursiva realizada por esse pacote lexical é semelhante à categoria de elaboração e clarificação do discurso, que será discutida na Seção 4.3.2.

Outros pacotes lexicais de posicionamento epistêmico também informam acerca de como palestrantes TED inserem opiniões e comentários de cunho pessoal e sinalizam marcas de atenuação (*hedge*) em seus discursos. Os padrões *and I think that* e *I think this is*, por

exemplo, são interessantes, pois evidenciam como palestrantes TED evitam um forte comprometimento em seus discursos ao sinalizarem opiniões e/ou incertezas em relação à informação apresentada. No exemplo a seguir, o uso do padrão *and I think that* sugere que a proposição apresentada (*i.e. the use of sound in astronomy*) não deve ser entendida como um conhecimento certo e imutável no campo da astronomia, mais sim um ponto de vista subjetivo e particular ao palestrante.

[08] When we give people the opportunity to succeed without limits, that will lead to personal fulfilment and prospering life. **And I think that** the use of sound in astronomy is helping us to achieve that and to contribute to science.
(TED_STEM_Astronomia_44)

De modo similar, o uso do pacote lexical *I think this is* também expressa as opiniões de cunho pessoal e julgamentos de valor por parte dos palestrantes. Conforme é possível observar no exemplo a seguir, o padrão *I think this is* auxiliou o palestrante na exposição de sua opinião acerca da organização dos sistemas educacionais ao redor do mundo.

[09] There isn't an education system on the planet that teaches dance every day to children the way we teach them mathematics. Why? Why not? **I think this is** rather important. I think math is very important, but so is dance. Children dance all the time if they're allowed to, we all do. We all have bodies, don't we?
(TED_SHAPE_Educação_38)

Os exemplos acima revelam uma estratégia retórica empregada por palestrantes TED para estabelecer um equilíbrio entre a autoridade do palestrante e a abertura para o diálogo e o debate. A recorrência desses pacotes lexicais entre palestrantes de ambas as áreas indica um certo grau de prudência no discurso de palestrantes TED e sugere que eles estariam receptivos a dialogar com pontos de vistas, argumentos e interpretações alternativas/contrárias as suas.

Um aspecto interessante revelado pelos dados foi o maior conjunto de expressões padronizadas distintas para sinalizar atenuação no *corpus* TED-SHAPE, quando comparado à contraparte STEM. No exemplo a seguir, o palestrante recorre ao pacote lexical *I don't know what* para sinalizar uma certa dificuldade em explicar ou definir um determinado objeto para seu público-alvo.

[10] Okay, what is this object? I have no idea. **I don't know what** that thing is. It looks terrible. Is it a little hot plate? I don't get it. Does anyone know?
(TED_SHAPE_Arquitetura_37)

Já o pacote lexical *I don't know if* evidencia momentos de maior interação entre palestrantes das áreas SHAPE e o público-alvo. Em aproximadamente 40% das ocorrências, esse padrão foi empregado com o objetivo de averiguar se determinada informação seria ou não de conhecimento do público-alvo, conforme pode ser observado no exemplo a seguir.

[11] Now, **I don't know if** anybody knows who this is. This is ... does somebody know? This is Beethoven.
(TED_SHAPE_Música_28)

Ainda em relação aos exemplos apresentados acima, é interessante observar que os palestrantes fazem uso de perguntas retóricas (*i.e. does anyone/somebody know?*) para verificar se alguém presente no auditório seria capaz de reconhecer o conteúdo em questão. Isso reforça a intenção de promover o conhecimento de maneira colaborativa e simétrica junto ao seu público-alvo, evidenciando a importância do diálogo e da troca de perspectivas nos eventos das TED Talks.

Em uma perspectiva macro, os dados referentes aos pacotes lexicais de posicionamento epistêmico sugerem que, embora sejam publicamente reconhecidos como especialistas em seus respectivos campos de atuação, palestrantes de ambas as áreas tendem a não depender de credenciais profissionais e titulações acadêmicas como argumento de autoridade. Em vez disso, esses palestrantes recorrem a um conjunto considerável de expressões padronizadas que os auxiliam a comentar acerca da veracidade das informações e, conseqüentemente, promover credibilidade discursiva junto ao público-alvo. Essa abordagem retórica contribui para o estabelecimento de um ambiente de aprendizado colaborativo e de engajamento mútuo, alinhado aos objetivos das TED Talks.

Em relação à variação disciplinar, observei também um posicionamento epistêmico mais interpessoal, avaliativo e prudente entre palestrantes das áreas SHAPE quando comparados à contraparte STEM. Considerando que a natureza humana e social das *soft sciences* permite um maior número de interpretações e abordagens divergentes sobre um mesmo fenômeno (Becher, 1994; Becher; Trowler, 2001; Hyland, 2006) é plausível inferir que os palestrantes SHAPE sinalizam mais aberturas em seus discursos para expor e elaborar suas

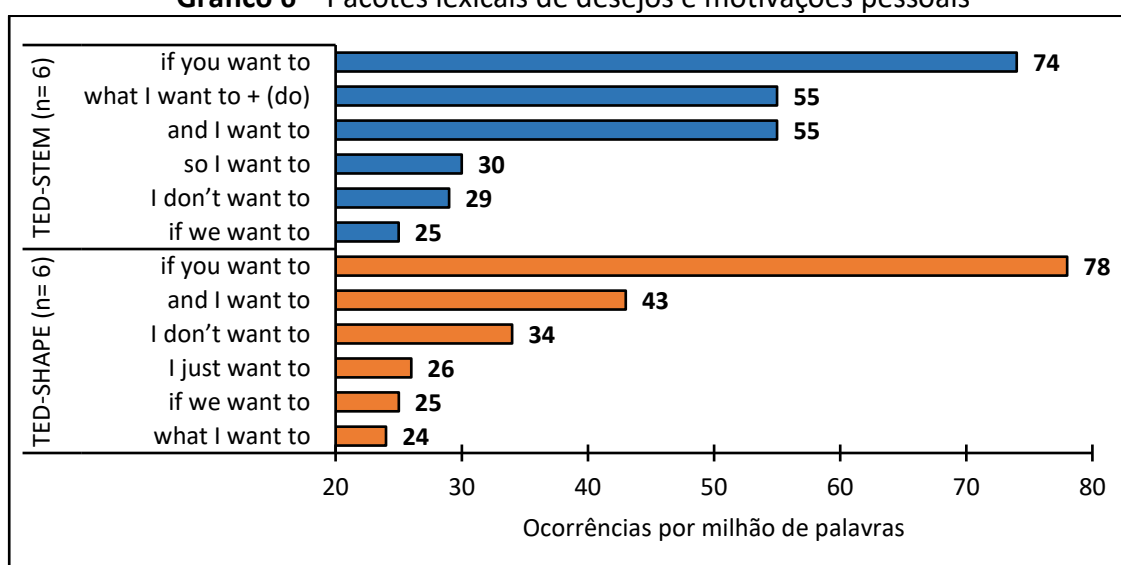
opiniões. Por outro lado, tal característica seria menos evidente nas *hard sciences*, uma vez que os interesses desses palestrantes tendem a ser fenômenos de natureza empírica, quantitativa e discreta, isto é, passíveis de serem estudados mesmo quando isolados de um todo. Em outras palavras, disciplinas da área STEM tendem a apresentar informações de modo linear e preciso, limitando assim as aberturas para que seus palestrantes apresentem posicionamentos e opiniões em seus discursos.

Tendo concluído a discussão sobre o conjunto de pacotes lexicais de posicionamento epistêmico extraídos de ambos os *corpora* de estudo, analiso, nas seções seguintes, as categorias de pacotes lexicais de posicionamento atitudinal, iniciando pelos padrões utilizados para expressar desejos e motivações pessoais dos palestrantes diante do público-alvo.

4.2.2 Posicionamento atitudinal: desejos e motivações pessoais

Os dados exibidos no Gráfico 6 indicam que palestrantes TED de ambas as áreas recorrem a um conjunto semelhante de expressões padronizadas para sinalizar desejos e motivações pessoais em seus discursos. As duas variações observadas foram os padrões *so I want to* e *I just want to*, que ocorreram com maior frequência no *corpus* TED-STEM e TED-SHAPE, respectivamente. Vale destacar também a alta recorrência do pacote lexical *if you want to* em relação as demais sequências de desejos e motivações encontradas em ambos os *corpora* de estudo.

Gráfico 6 – Pacotes lexicais de desejos e motivações pessoais



Fonte: Elaborado pelo pesquisador

No que diz respeito ao pacote lexical *if you want to*, observei que a utilização desse padrão em TED Talks apontou para interesses distintos entre os palestrantes das áreas STEM e SHAPE. Nos exemplos a seguir, é possível observar, respectivamente, que palestrantes STEM tendem a empregar esse pacote lexical em conjunto com verbos de natureza cognitivas (*i.e. learn*) e de natureza técnica e empírica (*i.e. find*), indicando um interesse maior na promoção do espírito científico no público-alvo e no estímulo à aquisição de novos conhecimentos. O Exemplo 12 também ilustra como palestrantes STEM buscam recontextualizar práticas científicas de maneira simples e direta para o público leigo.

[12] **If you want to** learn about thermodynamics, a duck is a good place to start, for example, why their feet don't get cold.
(TED_STEM_Física_34)

[13] Now, **if you want to** find dinosaurs, don't look in the Paleozoic, you won't find them. They hadn't evolved yet.
(TED_STEM_Geologia_02)

Já em TED Talks associadas às áreas SHAPE, o uso de *if you want to* indicou um maior enfoque em estimular o público-alvo a adotar práticas voltadas a melhoria do convívio social. O Exemplo 14 ilustra como o palestrante emprega esse padrão para propor uma ação comportamental, qualitativa e intrínseca às relações humanas (*i.e. establish trust*).

[14] And we put all the questions I got and all the answers on an open Facebook because we thought transparency is important **if you want to** establish trust.
(TED_SHAPE_Ciências_Políticas_34)

É relevante destacar também como a mudança pronominal de *you* → *we*, no pacote lexical *if we want to*, revela diferentes focos discursivos entre as áreas STEM e SHAPE. Na primeira, observei que o pronome *we* tende a se referir ao palestrante e demais especialistas. O exemplo a seguir ilustra como *if we want to* evidencia a participação do palestrante como membro de uma comunidade acadêmica e apresenta/discute ações que estariam além da responsabilidade de um público leigo, consequentemente, reforçando a imagem de autoridade do palestrante.

- [15] So we can get results in real time and analyze them with our contextual information. **If we want to** better understand and decode diseases like cancer, we need to stop treating them as acute, isolated episodes, and consider and measure everything that affects our health on a permanent basis.
(TED_STEM_Biotecnologia_04)

Os palestrantes vinculados às áreas SHAPE, por outro lado, tendem a incluir seu público-alvo como agentes da ação apresentada junto ao pacote lexical *if we want to*. Nesse sentido, a diferença entre os pacotes lexicais *if you want to* e *if we want to* seria menos evidente no discurso das TED Talks SHAPE, uma vez que ambos sugerem uma interação mais direta entre palestrantes e o público-alvo.

- [16] We know we are in a global climate crisis, but globally, our waters are also threatened, and we are facing a global water crisis, and **if we want to** address these crises in our lifetime, we need to change.
(TED_SHAPE_Direito_33)

O pacote lexical *I don't want to* também revelou algumas características interessantes no discurso nas TED Talks. Uma delas está relacionada à maneira como situações e/ou acontecimentos que os palestrantes desejam evitar são apresentados no discurso. O exemplo a seguir ilustra como o palestrante emprega esse pacote lexical para expressar suas inseguranças e vulnerabilidades frente ao público-alvo.

- [17] **I don't want to** get there and then still feel like a fraud. **I don't want to** feel like an impostor. **I don't want to** get there only to feel like I'm not supposed to be here.
(TED_STEM_Psicologia_09)

O exemplo acima destaca uma característica considerada fundamental no discurso das TED Talks: o estabelecimento de empatia e conexão emocional entre palestrante e seu público-alvo (Caliendo; Compagnone, 2014; Sugimoto et al., 2013). Em ambos os *corpora* de estudo, palestrantes TED recorrem ao padrão *I don't want to* para expor medos e incertezas que, embora sejam particulares e subjetivos aos palestrantes, o público-alvo pode facilmente se relacionar.

Além disso, o pacote lexical *I don't want to* também se mostrou como um padrão recorrente quando os palestrantes buscam orientar seu público-alvo sobre como as ideias apresentadas em seus discursos deveriam ser compreendidas. No exemplo a seguir, o pacote

lexical *I don't want to* auxilia o palestrante a primeiro minimizar possíveis equívocos de interpretação por parte de seu público-alvo para, em seguida, clarificar o seu posicionamento e atitude em relação ao assunto por meio do pacote lexical *I just want to* e do verbo *suggest*

- [18] **I don't want to** romanticize poverty. **I just want to** suggest that this informal urbanization is not just the image of precariousness.

(TED_SHAPE_Arquitetura_84)

Por fim, é relevante destacar como alguns dos pacotes lexicais dessa categoria desempenham múltiplas funções discursivas. Os padrões *and I want to*, *I just want to*, *so I want to* e *what I want to + (do)* são exemplos de expressões recorrentes que, simultaneamente, expressam desejos dos palestrantes e introduzem novos tópicos ao discurso.

- [19] We have ideas for how to make things better, **and I want to** share three of them that we've picked up in our own work.

(TED_SHAPE_Jornalismo_14)

- [20] Don't worry. I know it's very complicated. I don't want to take you through the gory details, **I just want to** tell you about some of the amazing features of this product.

(TED_SHAPE_Economia_53)

- [21] Well, maybe some of you have already seen this picture. I also brought it to you in person, **So I want to** kindly introduce you to the 3-D microprinter, which looks like this.

(TED_STEM_Engenharia_45)

- [22] The law leans so severely towards punishment these days that it's part of the problem. And that's **what I want to** talk about here.

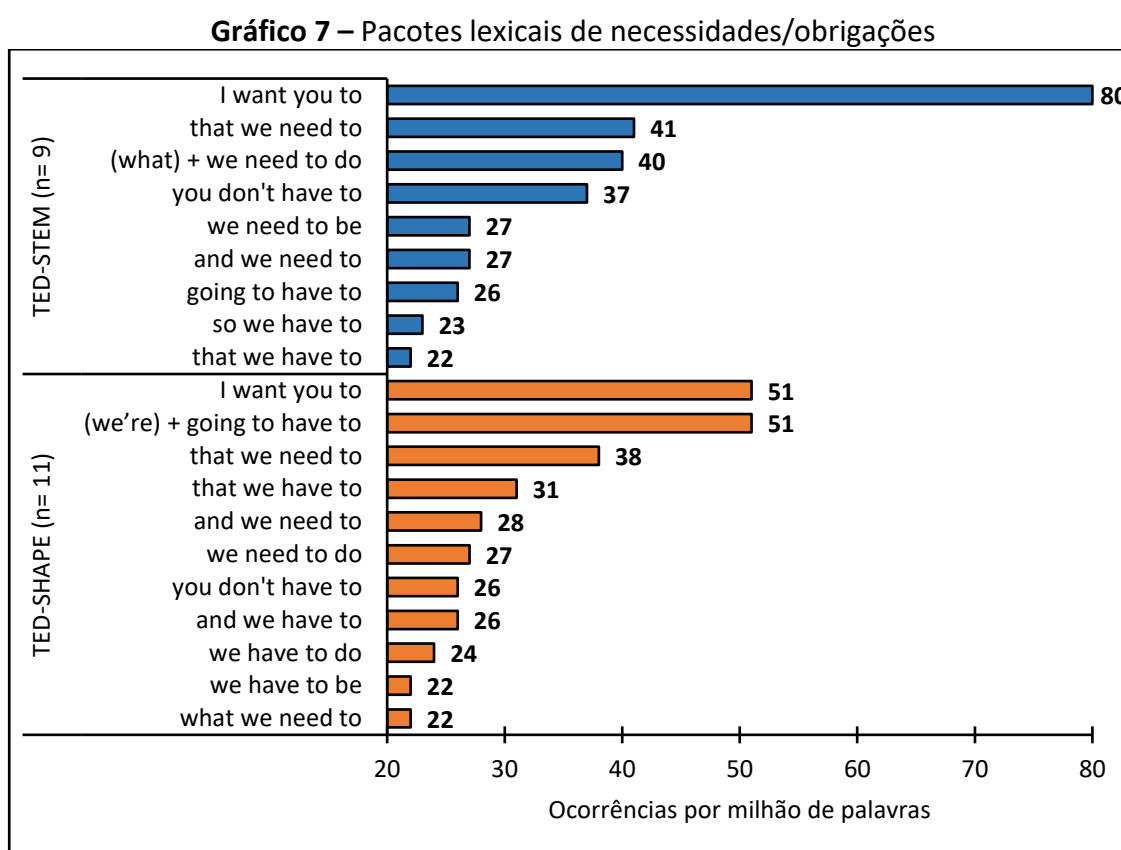
(TED_SHAPE_Direito_44)

Os exemplos acima ressaltam a importância do engajamento e da conexão que o palestrante busca estabelecer com seu público-alvo, uma vez que eles se esforçam para compartilhar suas ideias, esclarecer conceitos e apresentar aspectos interessantes de suas respectivas áreas de conhecimento. Tal estratégia discursiva é empregada tanto nas áreas SHAPE quanto nas áreas STEM, revelando um elemento comum na comunicação e na abordagem dos palestrantes.

Ao concluir a análise dos pacotes lexicais referentes às expressões de desejos, abordo, na seção seguinte, os pacotes lexicais classificados na subcategoria de necessidades/obrigações. Devido ao grande conjunto de padrões distintos relacionados à manutenção da atenção do público-alvo, optei por adaptar a taxonomia funcional proposta por Biber, Conrad e Cortes (2004) e distinguir os pacotes lexicais que expressam necessidades/obrigações no discurso (Seção 4.2.3) daqueles associados ao direcionamento da atenção do público-alvo. Estes últimos, serão discutidos posteriormente na Seção 4.2.4.

4.2.3 Posicionamento atitudinal: necessidades/obrigações

Nos registros acadêmicos, tanto orais quanto escritos, as expressões que sinalizam necessidades e obrigações transmitem geralmente significados relacionados a deveres pessoais. Isso inclui a identificação dos responsáveis por realizar determinadas ações, ou ações que indicam conclusões lógicas com base nas evidências apresentadas no discurso (Biber, 2006a). No contexto das TED Talks, o Gráfico 7 exibe os pacotes lexicais mais recorrentes entre as áreas STEM e SHAPE para realizar tais funções.



Fonte: Elaborado pelo pesquisador

Entre o conjunto de padrões apresentado acima, o pacote lexical *I want you to* se destacou como um dos padrões mais frequentes, especialmente no *corpus* TED-STEM, para sinalizar ações ou atividades que o palestrante deseja ser realizadas pelo seu público-alvo. Uma variação notável em relação a esse pacote lexical envolve a maior proporção de verbos de natureza cognitiva ou reflexiva (*i.e. realize*) no *corpus* TED-SHAPE quando comparado à contraparte STEM. Mais da metade das ocorrências de *I want you to* no *corpus* TED-SHAPE estiveram associados às atividades dessa natureza.

- [23] I want to push this, because the reason is the men automatically – they get – not automatically, but they tend to get more support. But **I want you to** realize that resources in the hands of African women is a powerful tool.
(TED_SHAPE_Desenvolvimento_Internacional_16)

Por outro lado, o exemplo a seguir ilustra como palestrantes STEM tendem a utilizar *I want you to* para interagir com o público-alvo e explicar ações e processos a serem realizados.

- [24] All right, now what **I want you to** do is to recite the last five letters of the English alphabet backwards. You should have started with Z. All right?
(TED_STEM_Psicologia_100)

Palestrantes TED também recorrem aos demais pacotes lexicais desta categoria para identificar ações e procedimentos necessários para resolver um problema ou alcançar determinado resultado. Nesse sentido, foi possível estabelecer uma relação entre os pacotes lexicais de necessidades/obrigações e o passo retórico *chamada para ação* (*call for action*) proposto por Chang e Huang (2015). Os autores observaram que é comum haver um momento no discurso das TED Talks no qual palestrantes propõem possíveis caminhos e ações que consideram as melhores dentro de seus respectivos contextos. Os exemplos a seguir sugerem que palestrantes TED tendem a sinalizar chamadas para ações de responsabilidade coletiva, instigando o público-alvo a refletir criticamente sobre o problema apresentado no discurso.

- [25] The climate is changing, everyone knows that. And it's here and it's bad and it's serious, **and we need to** do something about it. But I feel pretty optimistic that we can do this.
(TED_STEM_Botânica_03)

[26] And it's an important issue, because it means **that we have to**, as a society, think about what we value. And if it's just papers and Phys. Rev. letters, we've got a problem.
(TED_STEM_Química_06)

[27] We're either all losers or we all can be winners. But we are **going to have to** decide between zero and sum. **We're going to have to** decide between zero benefit for all or living life as the sum of all of us.
(TED_SHAPE_Sociologia_03)

[28] And if you want to have an inclusive America, you have to fight for the vote, and that is basically **what we need to** see right now.
(TED_SHAPE_Ciências_Políticas_45)

Uma diferença notável, no entanto, foi a maior recorrência do pacote lexical *so we have to* no *corpus* TED-STEM. Ao analisar esse padrão, observei que, embora seu uso também evidencie problemas que necessitam de soluções, ele tende a posicionar o palestrante e outros membros da comunidade científica como aqueles que, eventualmente, serão os responsáveis por encontrar as soluções, semelhante ao que foi evidenciado pelo pacote lexical *if we want to* (cf. Seção 4.2.2). O exemplo a seguir sugere que problema apresentado junto ao padrão *so we have to* indica uma responsabilidade que estaria além do público-leigo.

[29] So we can't use airbags. **So we have to** come up with another ingenious idea of how do we land it.
(TED_STEM_Astronomia_09)

Como apontado por Hyland (2004b), enfatizar a filiação do pesquisador a uma comunidade dedicada à busca de novos conhecimentos e soluções de problemas é uma maneira eficaz de atribuir maior credibilidade discursiva. Nesse contexto, a recorrência dos pacotes lexicais *if we want to* e *so we have to* sugere que essa estratégia de persuasão aparenta ser mais comum entre os palestrantes de TED Talks da área STEM.

Por último, o pacote lexical *you don't have to* evidencia uma tentativa dos palestrantes TED em aproximar o fazer científico do público leigo. Os exemplos a seguir ilustram como palestrantes de ambas as áreas recorrem a este pacote lexical com o intuito de explicitar para seu público-alvo que não é necessário possuir um alto grau acadêmico ou se preocupar com complexidade de certos procedimentos para entender como a ciência funciona.

[30] Hanny is a citizen scientist, or as I like to call them, "citizen warriors." And she shows that **you don't have to** be a fancy astrophysicist to participate. You just have to be curious.

(TED_STEM_Astronomia_24)

[31] When people ask you, "What is the evidence for evolution?" **You don't have to** think about some exotic examples, or fossils. You just have to look at your skin.

(TED_SHAPE_Antropologia_16)

Além de auxiliar o palestrante a desmitificar o fazer científico como algo extremamente complexo e fora do alcance do público leigo, o pacote lexical *you don't have to* também é empregado durante a clarificação de informações que possam não ter sido bem compreendida pelo público-alvo, conforme demonstram os exemplos abaixo.

[32] So **you don't have to** do amazing things with your money to make yourself happy. You can do small, trivial things and still get the benefits from doing this.

(TED_STEM_Psicologia_89)

[33] So if we go back to the trust stack, you still have to trust the idea, you have to trust the platform, but **you don't have to** trust the other person in the traditional sense.

(TED_SHAPE_Economia_64)

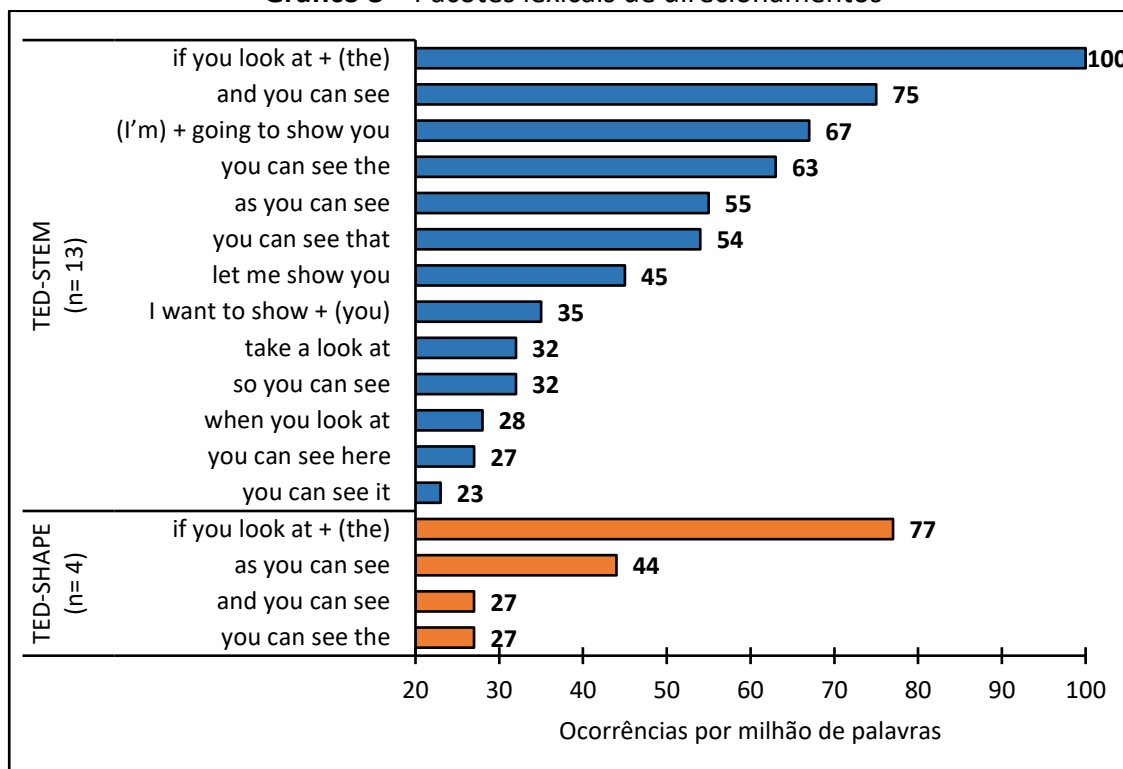
Tendo discutido os pacotes lexicais que sinalizam necessidades/obrigações no discurso das TED Talks, abordo, na seção seguinte, a categoria os pacotes lexicais que auxiliam esses palestrantes no direcionamento da atenção de seu público-alvo a elementos visuais e/ou informações importantes para a compreensão das ideias apresentadas no discurso

4.2.4 Posicionamento atitudinal: direcionamentos

O Gráfico 8 ilustra o conjunto de padrões aos quais palestrantes TED recorrem a fim de gerenciar a atenção do público-alvo no decorrer de seus discursos. Levando em consideração a natureza expositiva das TED Talks e a alta sofisticação audiovisual presente nesse gênero (Anderson, 2016; Kedrowicz; Taylor, 2016; Scotto di Carlo, 2014a), não é surpreendente que muitos dos pacotes lexicais de direcionamento encontrados sejam compostos por verbos relacionados ao campo semântico visual/percepção (*i.e. look, see e show*). O maior conjunto de pacotes lexicais distintos encontrado no *corpus* TED-STEM indica uma padronização discursiva mais acentuada desses palestrantes no momento de utilizarem

recursos visuais como instrumentos na elaboração de seus discursos e na clarificação das ideias expostas.

Gráfico 8 – Pacotes lexicais de direcionamentos



Fonte: Elaborado pelo autor

Em especial, pacotes lexicais compostos pelo verbo lexical *see* tendem a ser empregados no fornecimento de informações/explicações adicionais que facilitem o processo de compreensão do público-alvo. Adicionalmente, é possível observar nos exemplos a seguir que o uso desses padrões sugere uma tentativa de envolver o público-alvo como testemunhas visuais dos resultados apresentados.

[34] Three-dimensional printing devices that allow us to print the components exactly to the micron detail of what the child's anatomy will look like. **You can see here** the skull of this child being printed in the hours before we performed this surgery.

(TED_STEM_Medicina_86)

[35] He's got a prominent Adam's apple. He's got angular bones to the face, a mustache, **and you can see** there's a height difference, because he's gone through a growth spurt that she won't get.

(TED_SHAPE_Estudos_Gênero_29)

[36] So we mold the blocks, **as you can see** in the photo, and we cure it using water steam.

(TED_STEM_Engenharia_50)

[37] Actually, what happens is that a very intricate structure appears. And I hope **you can see that** there are some ring-like structures in the middle image, which is water.

(TED_STEM_Biologia_44)

No que se refere ao encadeamento *if you look at + (the)*, este se revelou como o padrão mais recorrente que palestrantes TED empregam para, simultaneamente, direcionar a atenção do público-alvo aos elementos visuais no palco e iniciar a discussão sobre o tópico em seu discurso, conforme ilustram os exemplos a seguir.

[38] This is my home state of Connecticut. Connecticut used to be home to a lot of wild salmon. But **if you look at** this map of Connecticut, every dot on that map is a dam. There are over 3,000 dams in the state of Connecticut.

(TED_STEM_Ecologia_19)

[39] And if you look, wind and concentrated solar have the lowest CO2 emissions, **if you look at the** graph. Nuclear – there are two bars here. One is a low estimate, and one is a high estimate.

(TED_SHAPE_Economia_76)

O emprego de *if you look at + (the)* em TED Talks também se mostrou coerente com os resultados apresentados por Neely e Cortes (2009) em palestras acadêmicas, já que esse padrão também é utilizado com a função de focar um tópico, especialmente em situações nas quais ele não faz referências a elementos visuais no palco. Neely e Cortes (2009) esclarecem que, nesses casos, o significado de *if you look at* será semelhante ao da expressão “*if you consider*”, conforme é possível observar nos exemplos a seguir.

[40] At the same time, if you ask, "Has that impacted the efficiency, how we've been able to develop drugs?" the answer is not really. **If you look at the** cost of developing drugs, how that's done, it basically hasn't budged that.

(TED_STEM_Genética_21)

[41] **If you look at** science on one hand, science is a very rational approach to its surroundings, whereas art, on the other hand, is usually an emotional approach to its surroundings.

(TED_SHAPE_Artes_06)

Outros padrões compostos pelo verbo lexical *look* foram as sequências *take a look at* e *when you look at*. Tais pacotes lexicais mostraram-se recorrentes apenas no *corpus* TED-STEM e auxiliam os palestrantes no direcionamento da atenção do público-alvo para, em seguida, elaborar explicações acerca das informações visuais ressaltadas. Nos exemplos a seguir, é interessante notar como o uso do marcador discursivo *let's* (Exemplo 42) e da pergunta retórica (*i.e. what do you see?* – Exemplo 43) revelam diferentes artifícios discursivos empregados por palestrantes STEM para reconhecer a presença do público-alvo, direcionar sua atenção e estabelecer maior proximidade e engajamento.

- [42] So let's **take a look at** this plant here, Centella asiatica. It's a weed. We call it a weed. Now, Centella asiatica grows across the world in many different habitats – in Africa, in Asia – and this plant has been instrumental in providing a solution to that dreadful disease called leprosy in Madagascar in the 1940s.

(TED_STEM_Botânica_01)

- [43] **When you look at** this image, what do you see? Well, I see a highly sophisticated biological system, that through the use of enzymes, can move and place atoms more quickly and precisely than anything we've ever engineered.

(TED_STEM_Biologia_96)

Palestras STEM também recorrem a um conjunto maior de pacotes lexicais compostos pelo verbo lexical *show*. A maior recorrência dos padrões (*I'm*) + *going to show you*, *I want to show* + (*you*) e *let me show you*; reforça um aspecto importante da variação disciplinar nesse registro apontada por Caliendo e Compagnone (2014) de que os palestrantes de TED Talks relacionadas às *hard sciences* preocupam-se em apresentar evidências diretas e tangíveis para atribuir maior credibilidade a seus discursos. Os próximos exemplos ilustram como palestrantes STEM se preocupam em não apenas exibirem aparatos visuais estáticos, como fotografias e gráficos, mas também em demonstrar visualmente o funcionamento de processos, assegurando assim a compreensão das informações transmitidas.

- [44] In a second, **I'm going to show you** a slow motion of a single firefly so that you can get a sense.

(TED_STEM_Matemática_31)

- [45] For trying to map the past, I knew that I had to see differently. So **I want to show you** an example of how we see differently using the infrared.

(TED_STEM_Geologia_05)

[46] So we get quantum locking. And **let me show you** how this works. I have here a superconductor, which I wrapped up so it'd stay cold long enough. And when I place it on top of a regular magnet, it just stays locked in midair.
(TED_STEM_Física_07)

No tocante ao conjunto de pacotes lexicais compostos pelos verbos lexicais *look* e *show* no *corpus* TED-STEM, é possível inferir que a maior padronização discursiva na maneira como esses palestrantes gerenciam a atenção de seu público-alvo esteja relacionada à predominância de recursos visuais (*e.g.* diagramas, fórmulas e equações) no discurso acadêmico oral das *hard sciences* (Rowley-Jolivet, 2002b, 2004). Poderia ser argumentado, portanto, que a maior recorrência de pacotes lexicais visuais em TED Talks da área STEM reflete o estilo desses palestrantes em divulgar o conhecimento apresentando evidências empíricas que fundamentam a elaboração de seus argumentos e, consequentemente, os tornam mais difíceis de serem refutados. Em oposição, a construção discursiva das *soft sciences* tende a ser realizada de forma argumentativa, reiterativa e mais focada na interpretação subjetiva do pesquisador sobre de ideias e conceitos abstratos que, em geral, não necessitam de suporte visual para serem explicados ao público-alvo (Kashiha; Heng, 2013; Rowley-Jolivet, 2002b, 2004; Simpson-Vlach, 2006). Essas diferenças refletem as características intrínsecas de cada área do conhecimento e as formas pelas quais os palestrantes recorrem a diferentes pacotes lexicais para engajar e informar seu público-alvo.

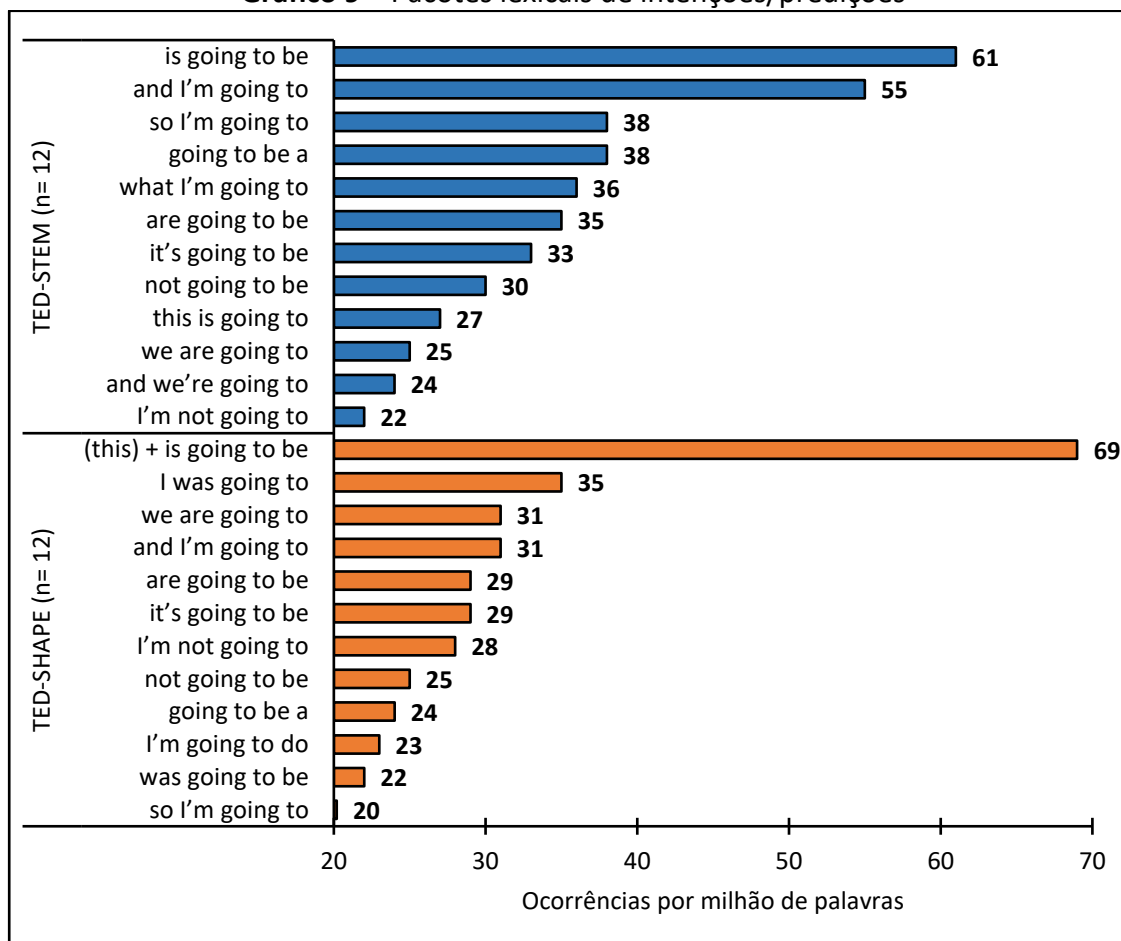
Tendo finalizado a discussão dos pacotes lexicais que sinalizam direcionamentos no discurso das TED Talks, abordo na seção seguinte os padrões mais recorrentes empregados por palestrantes TED para comunicar suas expectativas e projeções em relação aos temas abordados em suas apresentações.

4.2.5 Posicionamento atitudinal: intenções/predições

Nesta seção, abordo os pacotes lexicais empregados na sinalização das intenções, antecipação de tópicos e apresentação de previsões relacionadas a desenvolvimentos e eventos futuros. Os dados ilustrados no Gráfico 9 demonstram que, em ambas as áreas, essas funções discursivas são executadas por meio de um conjunto considerável de pacotes lexicais. Ao entender a disseminação de ideias e iniciativas de inovação tecnológica como uma das principais missões das TED Talks, é plausível inferir que a alta padronização dessa função

esteja vinculada ao propósito dos palestrantes em apresentar, ao público-alvo, uma gama de possibilidades oriundas de suas pesquisas.

Gráfico 9 – Pacotes lexicais de intenções/predições



Fonte: Elaborado pelo autor

Em relação à variação na frequência das ocorrências, vale destacar a maior recorrência do encadeamento *(this) + is going to be*, em comparação aos demais pacotes lexicais do *corpus* TED-SHAPE. Ainda que padrão *is going to be* também seja o mais empregado por palestrantes STEM, a diferença entre os demais pacotes lexicais do conjunto se mostrou menos acentuada quando comparada à contraparte SHAPE. Tal variação sugere que palestrantes STEM empregam tanto padrões que expressam predições (*i.e. is going to be*) quanto intenções (*i.e. and I'm going to*) em seus discursos. Por outro lado, a elevada recorrência do encadeamento *(this) + is going to be* no *corpus* TED-SHAPE sugere que esses palestrantes tendem a adotar um discurso mais impessoal e voltado à predição de eventos.

No que concerne à predição de eventos, observei que, em ambas as áreas, palestrantes TED recorrem a diversos pacotes lexicais para inserir avaliações de caráter pessoal acerca de situações e eventos futuros prováveis. É interessante notar também o uso de diversos adjetivos e locuções que veiculam avaliações positivas (*i.e. amazing, pretty exciting*), que denotam um certo otimismo por parte dos palestrantes, conforme demonstram os exemplos a seguir.

- [47] CRISPR **is going to be** amazing, in terms of the number of different scientific advances that it's going to catalyze.
(TED_STEM_Biotecnologia_12)
- [48] And so, I'm pretty optimistic about it. I think **it's going to be** a pretty exciting time to be alive.
(TED_SHAPE_Educação_36)
- [49] But we actually believe **this is going to be** a very powerful set of tools and we're already starting in numerous avenues to use this tool.
(TED_STEM_Biologia_136)
- [50] I see that we certainly do not know how **we are going to** solve those problems yet. But we can take a page out of what we have done on climate change and we can understand that we have got to reinterpret the zero-sum mentality.
(TED_SHAPE_Sociologia_03)

Há também uma notável padronização na maneira como palestrantes TED antecipam ações, atividades e tópicos que serão abordados no decorrer de suas palestras para o público-alvo. Em geral, essas funções são desempenhadas por pacotes lexicais de caráter pessoal (*e.g. and I'm going to*). Os exemplos a seguir ilustram momentos nos quais palestrantes TED antecipam para o público-alvo as atividades que pretendem realizar durante seu tempo de fala (*i.e. tell, recite e prompt*). Nesses casos, é possível observar uma sobreposição entre a categoria de intenções e predições com a macro função de organizadores discursivos.

- [51] So among and within bats is a huge variation in their ability to use sensory perception. Now **this is going to be** important for **what I'm going to** tell you later during the talk.
(TED_STEM_Genética_25)
- [52] So throughout the talk, I'm going to put lyrics up on the screen, **and I'm going to** recite some **and I'm going to** prompt you when it's your turn to do the response, OK?
(TED_SHAPE_Educação_15)

No que diz respeito ao pacote lexical *I'm not going to*, palestrantes TED de ambas as áreas empregam esse padrão como forma de explicar o motivo pelo qual alguma ação não será realizada durante sua apresentação. Devido o limite de tempo disponível para cada apresentação (aproximadamente 18 minutos), é necessário que os palestrantes priorizem quais informações são essenciais para serem transmitidas ao público-alvo. Tal necessidade explicaria, ao menos parcialmente, a recorrência considerável desse padrão em ambos os *corpora* de estudo.

[53] Now, **I'm not going to** go into discussing changing consumption patterns or reducing food waste, because those conversations have been going on for some time now.
(TED_STEM_Agronomia_18)

[54] **I'm not going to** go into the reasons for these tragedies, but they are part of the reason why technological change happens quite slowly.
(TED_SHAPE_Arquitetura_32)

Durante a análise da subcategoria de direcionamentos, aponte para o fato de palestrantes associados às áreas STEM utilizarem recursos visuais com maior frequência quando comparados à contraparte SHAPE. Essa diferença também foi reforçada pelo conjunto de pacotes lexicais de intenção e predição, uma vez que palestrantes STEM empregam pacotes lexicais dessa categoria para antecipar a exibição de algum tipo de demonstração.

[55] So **what I'm going to** do here is, I cracked open an episode of my radio show, "99% Invisible," **and I'm going to** reconstruct it here on stage.
(TED_STEM_Design_10)

[56] Even though we launched these just a couple of weeks ago, we've already got some initial imagery from the satellites **and I'm going to** show it publicly for the first time right now. This is the very first picture taken by our satellite.
(TED_STEM_Astronomia_46)

O conjunto de pacotes lexicais associados à categoria de intenção/predição ainda apontam para outras variações entre palestrantes TED das áreas STEM e SHAPE. Na primeira, palestrantes tendem a adotar um discurso mais linear e direto, com foco nos aspectos visuais e demonstrativos (*e.g. and I'm going to*). Por outro lado, a recorrência dos pacotes lexicais *I was going to* e *was going to be*, sugerem, por exemplo, que palestrantes da área SHAPE

tendem a priorizar a narração e contextualização de eventos para o público-alvo, destacando a importância da compreensão abrangente das circunstâncias envolvidas.

[57] When we started out, I thought **I was going to** find so much corruption, I was literally going to either die or get killed in the process.

(TED_SHAPE_Desenvolvimento_Internacional_08)

[58] I want to tell you about the last conversation that I had with Will. It was the day that he **was going to be** executed, and we were just talking. There was nothing left to do in his case.

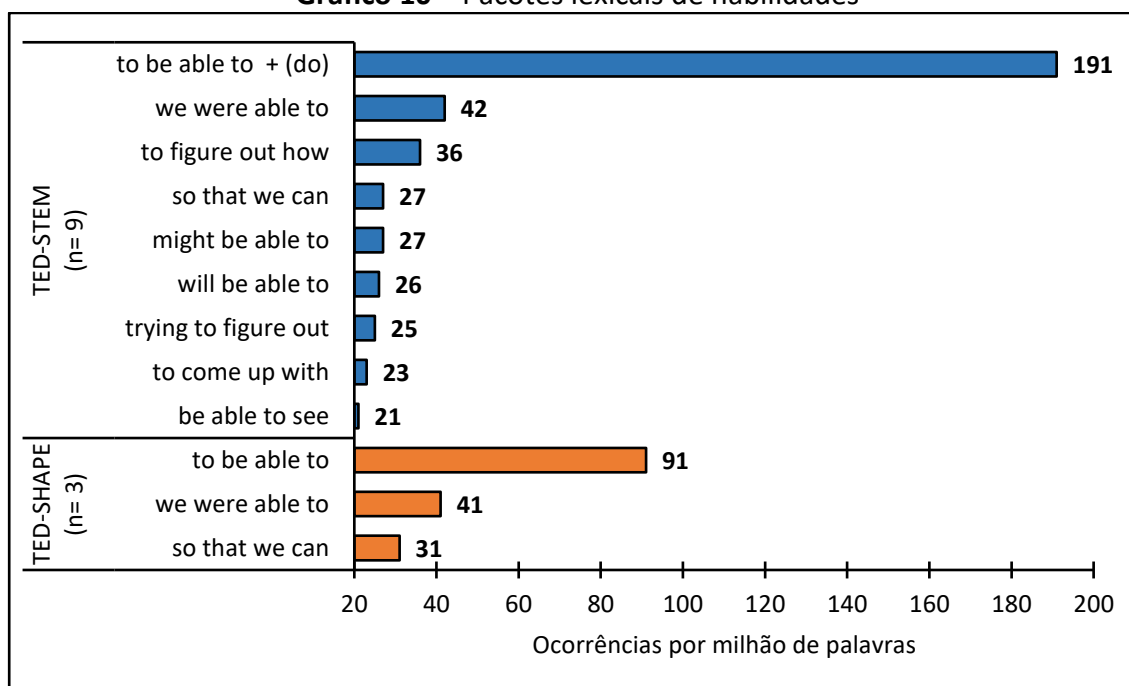
(TED_SHAPE_Direito_11)

Tendo explorado as diferenças entre o discurso de palestrantes das TED Talks STEM e SHAPE no que se refere ao uso de padrões com a função de sinalizar intenções e predições, passo agora à análise de outra dimensão do discurso: os pacotes lexicais relacionados a habilidades. Na seção seguinte, investigo como palestrantes de ambas as áreas abordam e destacam habilidades específicas, ressaltando a importância de tais competências em suas respectivas áreas de conhecimento e como elas contribuem para o desenvolvimento e aplicação prática de suas pesquisas e projetos.

4.2.6 Posicionamento atitudinal: habilidades

A última subcategoria de posicionamentos atitudinais agrupa os pacotes lexicais que indicam habilidades. O Gráfico 10 ilustra os padrões distintos encontrados em ambos os *corpora* de estudo a partir de suas frequências relativas.

Gráfico 10 – Pacotes lexicais de habilidades



Fonte: Elaborado pelo autor

No que concerne ao conjunto de dados apresentados acima, destaco, primeiramente, o conjunto três vezes maior de padrões distintos que expressam habilidades em TED Talks STEM em comparação a área SHAPE. A segunda variação a ser destacada é a notável discrepância entre a recorrência do encadeamento *to be able to + (do)* e os demais pacotes lexicais dessa categoria, especialmente em palestrantes das áreas STEM. Os exemplos a seguir sugerem que, possivelmente, a elevada recorrência desse padrão seria um reflexo do propósito de divulgação, inerente às TED Talks, de potenciais aplicações advindas de inovações e desenvolvimento tecnológicos para o público-alvo.

[59] Now if we could scale this, you could actually use it to run your Prius, which is kind of my dream – **to be able to** drive a virus-powered car.

(TED_STEM_Física_06)

[60] We're going **to be able to** see a child's brain as they experience an emotion, as they learn to speak and read, as they solve a math problem, as they have an idea. And we're going **to be able to** invent brain-based interventions for children who have difficulty learning.

(TED_SHAPE_Estudos_Linguísticos_19)

O conjunto de pacotes lexicais associados à expressão de habilidades também apontam para uma maior padronização no discurso das TED Talks vinculadas às disciplinas

STEM no que se refere à indicação de lacunas e desafios a serem superados. Para tanto, os exemplos a seguir demonstram como o padrão *to be able to* tende a ser empregado como uma extensão de expressões que sinalizam necessidades ou obrigações (*i.e. have to* e *ought to*).

[61] I absolutely understand the importance of ritual, especially when it comes to the people that we love. But we have **to be able to** create and practice this ritual without harming the environment, which is why we need new options.

(TED_STEM_Ecologia_02)

[62] Well, you know, anytime a new health threat enters our community, we ought **to be able to** rapidly identify it, contain it, and if we can mitigate it, for sure, and eliminate it if possible.

(TED_STEM_Medicina_32)

No que tange a apresentação de resultados obtidos, observei uma tendência de sinalização junto ao padrão *we were able to*. Novamente, os dados apontam que, em TED Talks associadas às áreas STEM, o uso do pronome *we* frequentemente refere-se à filiação do palestrante como membro de uma comunidade acadêmica. Essa diferença reforça uma variação já mencionada durante a discussão dos pacotes lexicais *if we want to* e *so we have to* (*cf.* Seções 4.2.2 e 4.2.3, respectivamente) de que, ao se posicionarem como membros de uma comunidade acadêmica, palestrantes STEM [indiretamente] sinalizam ao seu público-alvo que os resultados e argumentos reportados foram validados por outros membros de seus respectivos grupos de pesquisa (*i.e. I and my colleagues*) e, conseqüentemente, seriam confiáveis.

[63] And I and my colleagues cut into it, and **we were able to** determine, by looking at lines of arrested growth, some lines in it, that B-rex had died at the age of 16.

(TED_STEM_Genética_24)

[64] Working with an awesome team at Ginkgo, and others at UC Santa Cruz, **we were able to** extract a little bit of the DNA from a tiny sliver of this plant specimen and to sequence the DNA inside.

(TED_STEM_Engenharia_14)

No *corpus* TED-SHAPE, embora o uso do pacote lexical *we were able to* também esteja relacionado à ideia de trabalho em equipe, o foco dos palestrantes revelou-se mais direcionado ao relato de resultados qualitativos e que não necessitam, obrigatoriamente, de

validação externa para serem aceitos como confiáveis.

- [65] So with the help of the city and the parks, **we were able to** raise funds and design and build a half-court, right on the police station parking lot. It's a start.
(TED_SHAPE_Arquitetura_44)
- [66] And it was only through this medical and legal advocacy that **we were able to** help restore his legal status and his rights, that he could do that, all through asylum.
(TED_SHAPE_Direito_27)

O último pacote lexical de habilidades recorrente em ambos os *corpora* de estudo foi *so that we can*, que indica uma preocupação dos palestrantes em esclarecer ao público-alvo quais decisões foram tomadas para que os resultados desejáveis pudessem ser alcançados. No Exemplo 67, o emprego do padrão *so that we can* auxilia o palestrante a conectar uma ação (*i.e. capture and anesthetize*) com o resultado pretendido (*i.e. install GPS collars and follow their movements*). De modo similar, no Exemplo 68, o uso de *so that we can* conecta uma ação realizada (*i.e. come up with a petition*) e o propósito por trás disso (*i.e. demand new elections*).

- [67] We have to capture and anesthetize them **so that we can** install GPS collars around their necks and follow their movements.
(TED_STEM_Biologia_100)
- [68] One of the very first actions of our Faure Must Go movement was to come up with a petition, asking citizens to sign **so that we can** demand new elections, as the constitution allows.
(TED_SHAPE_Ciências_Políticas_28)

No que concerne ao *corpus* TED-STEM, a recorrência do padrão *might be able to* evidencia que palestrantes dessa área demonstram um certo grau de cautela ao especularem sobre determinados assuntos, similar ao uso dos pacotes lexicais de posicionamento epistêmico *and I think that* e *I think this is* (cf. Seção 4.2.1). No exemplo a seguir, o emprego do padrão *might be able to*, juntamente com a expressão anguladora *they speculate* expressa um certo distanciamento por parte do palestrante em relação à proposição que está sendo apresentada (*i.e. possibilidade de encontrar formas de vida em outros planetas*).

[69] And they speculate that up to one fifth of them have an Earth-like planet that **might be able to** harbor life, but we haven't seen any of them. We've only detected them indirectly.

(TED_STEM_Astronomia_22)

O emprego do pacote lexical *might be able to* também tende a estar relacionado às especulações de ideias e projetos ainda em estágio experimental e/ou que necessitariam de mais dados empíricos para serem confirmados, conforme ilustra o exemplo a seguir.

[70] If we could figure out how these cells make the decision to become one cell type or another, we **might be able to** harness them to generate cells that we need to repair diseased or damaged tissue.

(TED_STEM_Biotecnologia_21)

Já os pacotes lexicais *to come up with*, *to figure out how* e *trying to figure out* apontam uma variação interessante no modo como palestrantes das áreas STEM enfatizam o tipo de problema que almejam ou necessitam solucionar por meio de suas pesquisas. O padrão *to figure out how*, por exemplo, tende a estar relacionado à necessidade de compreender melhor o funcionamento de processos e fenômenos, conforme ilustra o exemplo a seguir.

[71] I believe that the answer here is **to figure out how** to create superintelligent A.I. such that even if – when – it escapes, it is still safe because it is fundamentally on our side because it shares our values.

(TED_STEM_Inteligência_Artificial_09)

Já os pacotes lexicais *to come up with* e *trying to figure out* emergem como escolhas recorrentes no momento em que palestrantes STEM sinalizam lacunas que ainda necessitam ser solucionadas. Nos exemplos abaixo, é interessante notar como esses padrões também sinalizam um certo grau de cautela dos palestrantes ao comentarem sobre áreas de pesquisas que ainda necessitam de estudos adicionais.

[72] And so what we're doing now is just **trying to figure out** on which types of exoplanets, which gases could be attributed to life.

(TED_STEM_Astronomia_37)

[73] And the aim of all this is to raise a generation of people able **to come up with** new ideas to face the challenges in front of us. We all know. I think it's a long way to go, but the questions, and the tools, are now there, adjacent and possible.

(TED_STEM_Matemática_35)

No decorrer desta seção, foi possível observar como o maior conjunto de pacotes lexicais que expressam habilidades em TED Talks associadas às áreas STEM revela uma preocupação maior desses palestrantes em delimitar de forma clara e precisa os problemas, procedimentos, resultados e limitações relacionados às suas disciplinas/pesquisas, quando comparados à contraparte SHAPE.

Após a análise detalhada dos pacotes lexicais de posicionamento em TED Talks das áreas STEM e SHAPE, é fundamental sintetizar e contextualizar os achados nesta seção, a fim de lançar luz sobre as implicações gerais e as possíveis conclusões que possam ser extraídas do presente estudo. Sendo assim, apresento, na seção seguinte, uma discussão abrangente sobre o que os dados revelam em termos de padrões, variações e similaridades entre as duas áreas de estudo, considerando os pacotes lexicais de posicionamento discutidos. Essa síntese permitirá uma melhor compreensão dos diferentes estilos discursivos e estratégias adotadas por palestrantes nas áreas STEM e SHAPE, bem como uma perspectiva sobre o papel que esses padrões desempenham na comunicação efetiva dentro do contexto das TED Talks.

4.2.7 Expressão de posicionamento: o que os dados nos revelam?

Os números indicam que os pacotes lexicais que expressam posicionamentos encontrados em ambos os *corpora* de estudo foram empregados principalmente para a sinalização de aparatos visuais, ênfase na apresentação dos resultados obtidos e promoção da autoridade dos palestrantes como especialistas. Tais estratégias serviram para conferir credibilidade às informações apresentadas nas TED Talks analisadas. Ao longo desta seção, apresentei a análise dos dados relacionados aos pacotes lexicais que auxiliam os palestrantes TED a expressar opiniões, avaliar o grau de conhecimento (certeza, incerteza e possibilidade) e sinalizar suas atitudes em relação às informações compartilhadas em seus discursos.

Na literatura acadêmica, a maneira como palestrantes TED expressam posicionamentos em seus discursos já foi examinada sob o conceito de “proximidade” (Caliendo, 2014; Scotto di Carlo, 2014b), assim como pelo uso de verbos lexicais epistêmicos

(i.e. *see, show, know* e *think*) (Caliendo; Compagnone, 2014) e de adjetivos avaliativos (Scotto di Carlo, 2015). No entanto, o presente estudo procurou responder a uma questão que não foi diretamente abordada previamente, isto é, *como se configura o discurso disciplinar das TED Talks por meio dos pacotes lexicais*; na presente seção, essa pergunta foi abordada exclusivamente pelo prisma dos pacotes lexicais de expressão de posicionamento. Para tanto, acredito que a variação disciplinar entre os *corpora* de estudo possa ser apresentada a partir dos aspectos da frequência e função discursiva. No que diz respeito à função discursiva de posicionamento, acredito que os pacotes lexicais identificados permitem os seguintes apontamentos acerca do discurso das TED Talks:

- **Palestrantes STEM empregam um número maior de expressões padronizadas em comparação com a contraparte SHAPE;**

As distribuições distintas das subcategorias de posicionamento evidenciam propósitos comunicativos diversos entre palestrantes TED de diferentes áreas disciplinares. Nas TED Talks conduzidas por palestrantes da área STEM, há um enfoque mais acentuado em apresentar soluções para problemas práticos ao público-alvo (e.g. *to be able to* + *(do)*) e destacar a utilidade e aplicações de projetos em desenvolvimento (e.g. *will be able to*). Palestrantes das áreas SHAPE, por sua vez, preocupam-se em contextualizar o público-alvo sobre os problemas (e.g. *I was going to*) e em ressaltar ações a serem realizadas (e.g. *(we're)* + *going to have to*).

Uma variação observada entre os *corpora* de estudo foi o maior número de pacotes lexicais distintos que incorporam o pronome “*you*” no *corpus* TED-STEM, possivelmente, isso se deve a quantidade de expressões utilizadas por palestrantes STEM no gerenciamento da atenção de seu público-alvo (e.g. *let me show you, so you can see, when you look at*, entre outros), evidenciando uma tentativa desses em estabelecer uma interação mais próxima com o público-alvo.

Palestrantes SHAPE, por outro lado, empregam um maior número de pacotes lexicais que incorporam o pronome “*we*”, o que sugere um discurso mais coletivo, no qual o público-alvo tende a ser incluído nas ações apresentadas (e.g. *(we're)* + *going to have to, we have to do*). No entanto, é importante ressaltar que palestrantes SHAPE interagem com o público-alvo por meio de pacotes lexicais distintos; a combinação *I don't know if* + *(you)*, por exemplo,

mostrou-se como um padrão recorrente para tal função.

- **A alta padronização de direcionamentos no *corpus* TED-STEM revela uma estreita relação entre o uso de aparatos visuais para as convenções epistemológicas nessas disciplinas;**

Conforme fora ilustrado no Gráfico 4, a categoria de direcionamento exibiu os maiores valores centrais de frequência relativa no *corpus* TED-STEM. É possível que a alta recorrência desses pacotes lexicais seja um reflexo da necessidade que esses palestrantes têm em utilizar diferentes tipos de aparatos visuais como facilitadores na recontextualização do discurso para o público-alvo. Embora palestrantes de ambas as áreas busquem direcionar a atenção de seu público-alvo, o amplo conjunto de pacotes lexicais distintos no *corpus* TED-STEM revelou que o gerenciamento da percepção (*e.g. as you can see, take a look at*) e a demonstração visual das informações apresentadas (*e.g. I'm going to show, let me show you*) são características notavelmente mais acentuadas nas TED Talks STEM. Essas características estariam alinhadas às práticas epistemológicas das *hard sciences*, que elaboram e defendem seus argumentos junto a instrumentos visuais para descrever com o máximo de precisão e detalhes os fenômenos estudados (Becher, 1987; Rowley-Jolivet, 2000, 2002a, 2002b). Por outro lado, a baixa recorrência desses pacotes lexicais em TED Talks SHAPE aponta para um discurso mais textualmente interpretativo e descritivo por parte desses palestrantes.

- **Intenções/predições são altamente padronizadas em ambas as áreas;**

Palestrantes STEM e SHAPE recorrem a um amplo conjunto de pacotes lexicais que os auxiliam a justificar, perante o público-alvo, a importância de suas pesquisas e de que forma os investimentos que receberam trarão benefícios para a sociedade. A alta recorrência de padrões do tipo: (*this*) + *is going to be* e *it's going to be* reflete o interesse em ambas as áreas de divulgar ideias e projetos, possivelmente ainda em fase experimental, mas já considerados de alto valor potencial.

- **Há um conjunto considerável de expressões que sinalizam habilidades em TED Talks STEM. Contudo, essa é a categoria com menor número de pacotes lexicais distintos no *corpus* TED-SHAPE;**

Possivelmente, a extrema recorrência do encadeamento *to be able to* + (*do*), assim como a presença de outros de pacotes lexicais tais como *to come up with* e *to figure out how* seria um reflexo da ênfase dada por palestrantes STEM em relatar realizações de processos e aplicações práticas de conhecimentos visando à resolução de problemas, algo considerado como mais típico à natureza disciplinar das *hard sciences* (Becher; Trowler, 2001; Trowler, 2012). Possivelmente, o fato de as *hard sciences* valorizarem a aptidão dos pesquisadores em solucionar problemas ajudaria a explicar o maior conjunto de pacotes lexicais distintos que sinalizam habilidades no *corpus* TED-STEM.

- **O movimento retórico *chamada para ação* tende a ser realizado de maneira mais padronizada entre palestrantes SHAPE;**

As *chamadas para ação* expressas nas TED Talks SHAPE indicam um interesse maior em atividades relacionadas às interações humanas e sociais do que à promoção da ciência. Por exemplo, o pacote lexical *if you want to* demonstrou que palestrantes SHAPE buscam instigar, em seu público-alvo, mudanças no âmbito das práticas sociais (*e.g. if you want to establish trust*). Palestrantes STEM, por outro lado, mostraram-se mais interessados em promover práticas científicas e em instigar seu público-alvo a buscar novos conhecimentos. Em comparação, o mesmo pacote lexical no *corpus* TED-STEM tende a promover atividades de natureza técnica (*e.g. if you want to build a robot*). Adicionalmente, palestrantes STEM também buscam desmitificar o conceito de ciência como algo complexo e restrito aos muros das universidades e laboratórios (*e.g. you don't have to have a PhD in biology*).

É válido observar também a ocorrência do plural majestático (*royal we*) na maioria dos pacotes lexicais utilizados para expressar obrigações em ambos os *corpora* de estudo (*e.g. that we have to, we need to do*). Considerando que um dos propósitos das palestras TED é a divulgação de temas relevantes tanto para a área acadêmica quanto para o cotidiano do público-alvo (Caliendo, 2014), o fato de o plural majestático ser recorrente em pacotes lexicais dessa categoria sinaliza a tentativa desses palestrantes em estabelecer interesses em comum com seu o público-alvo. Ou seja, por meio da noção de coletividade, os palestrantes expressam a necessidade de uma cooperação mais próxima entre sociedade e comunidade científica e compartilham com o público-alvo as responsabilidades para a solução de seus problemas. Tais padrões refletem uma das missões da organização TED que é “convidar a

todos a se envolverem com ideias e empregá-las em suas comunidades⁵⁶ (TED, 2022).

- **Palestrantes STEM e SHAPE empregam pacotes lexicais distintos para promover autoridade, transmitir conhecimentos e adotar posicionamentos em seus discursos.**

Diversos pacotes lexicais de posicionamento revelam que palestrantes TED buscam promover sua imagem como especialistas ao longo de seus discursos. Observei, por exemplo, que sequências como *I can tell you, if we want to* e *we were able to* são padrões recorrentes que auxiliam palestrantes a se promoverem como autoridades sobre determinado assunto diante do público-alvo. No entanto, é interessante destacar que palestrantes TED evitam associação a títulos e méritos acadêmicos e profissionais, apresentando-se, em vez disso, como pessoas interessadas em compartilhar conhecimentos e experiências únicas com o público-alvo.

No que se refere à transmissão de conhecimento, os dados revelaram que palestrantes das TED Talks STEM recorrem a diversos pacotes lexicais para recontextualizarem os conceitos, teorias e aplicações relacionadas às suas disciplinas de maneira mais tangível possível, a fim de facilitar a compreensão e direcionar/convencer o público-alvo a chegarem às mesmas conclusões que os palestrantes. Os padrões *so you can see* e *you can see that* são exemplos de pacotes lexicais recorrentes em TED Talks STEM que auxiliam na realização de tais funções.

A atenuação do discurso por meio de pacotes lexicais que sinalizam inserções de opiniões dos palestrantes também foi uma característica recorrente em ambos os *corpora* de estudo e indica uma preocupação de palestrantes TED em auxiliar seu público-alvo a diferenciar informações factuais, prováveis e duvidosas. Em ambos os *corpora* de estudo, os pacotes lexicais de posicionamento epistêmico, como *and I think that* e *I think this is*, são exemplos de padrões que sinalizam marcas de atenuação no discurso dos palestrantes. Entretanto, palestrantes STEM também recorrem a expressões de habilidades (*e.g. trying to figure out* e *might be able to*) para sinalizar atenuação. Visto que palestrantes STEM tendem a demonstrar um maior interesse em propor soluções para problemas, é compreensível que, em determinados momentos, os palestrantes manifestem cautela e evitem afirmações “exageradas” que possam ser rejeitadas pelo público-alvo. Assim, as marcas de atenuação em

⁵⁶ No original: “we invite everyone to engage with ideas and activate them in your community”.

STEM podem ser interpretadas como uma forma manter as interpretações e expectativas do público-alvo alinhadas ao que os dados sugerem.

Já no *corpus* TED-SHAPE, as marcas de atenuação foram evidenciadas por pacotes lexicais da categoria de posicionamento epistêmico (*e.g. I don't know what* e *I think this is*). Tais sequências também sinalizam a cautela dos palestrantes SHAPE em não apresentar conhecimentos absolutos, mas sim abertos a diferentes interpretações.

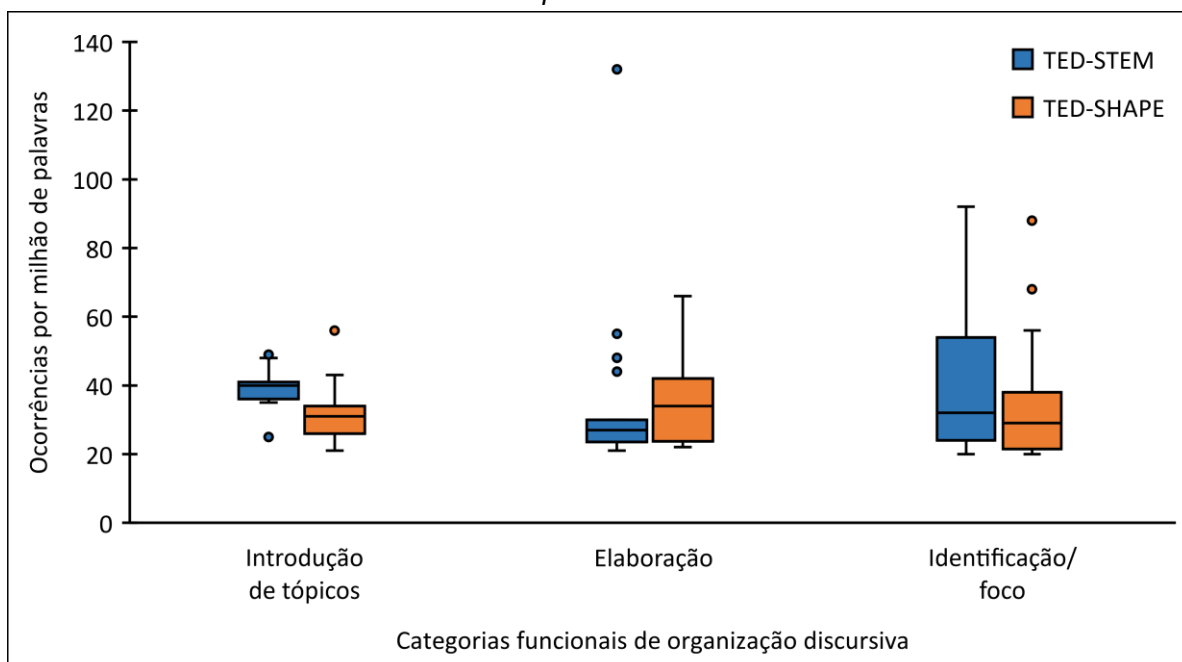
Tendo analisado as nuances e características presentes nos pacotes lexicais de posicionamento, tanto em palestras TED-STEM quanto em TED-SHAPE, discuto, na seção seguinte, acerca dos pacotes lexicais encontrados com a função de auxiliar os palestrantes a (re)organizar os seus discursos. Tais padrões operam como elementos coesivos que auxiliam na transição e conexão entre ideias, garantindo uma apresentação mais fluida e compreensível do discurso para o público-alvo.

4.3 Pacotes lexicais organizadores discursivos

Após ter apresentado os dados referentes às expressões de posicionamento em TED Talks das áreas STEM e SHAPE, nesta seção, analiso os pacotes lexicais mais recorrentes que os palestrantes TED empregam para estabelecer relações coesas entre as proposições em seus discursos. Na literatura acadêmica, os pacotes lexicais categorizados como organizadores discursivos são geralmente associados a três principais funções: (i) introdução de novos tópicos no discurso; (ii) elaboração ou esclarecimento das relações pragmáticas entre as ideias apresentadas ao longo do discurso; e (iii) identificação ou foco em entidades relevantes no discurso (Biber, 2006a; Biber; Barbieri, 2007).

Nos parágrafos a seguir, abordo, a partir de uma perspectiva macro, a estruturação do discurso disciplinar em TED Talks sob o prisma da categoria dos pacotes lexicais organizadores discursivos. Inicialmente, apresento no Gráfico 11 uma comparação entre as proporções de ocorrências dos pacotes lexicais nas subcategorias de organização discursiva em cada um dos *corpora* de estudo investigados.

Gráfico 11 – Distribuição de ocorrências de pacotes lexicais de organização discursiva entre os *corpora* de estudo



Fonte: Elaborado pelo autor

Os dados acima revelam algumas particularidades interessantes a respeito de como palestrantes das áreas STEM e SHAPE recorrem a pacotes lexicais para organizar seus discursos. Na categoria de introdução de tópicos, embora um pequeno número de pacotes lexicais apresente valores discrepantes de ocorrências, a elevada tendência central e baixa dispersão interna do conjunto de dados indica que os padrões adotados para realizar tal função discursiva tendem a ocorrer com frequências similares e apontam para a alta padronização em ambos os *corpora* de estudo, sobretudo no *corpus* TED-STEM.

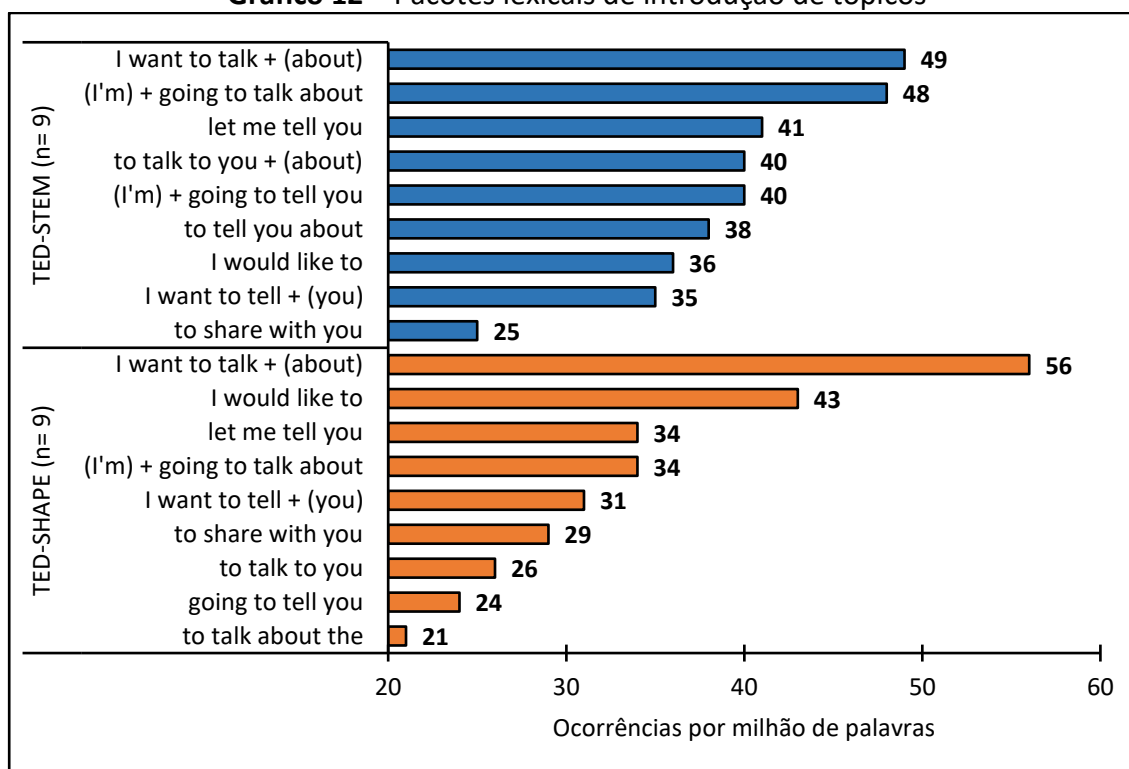
Quanto aos pacotes lexicais utilizados para elaboração do discurso, o tamanho da calda e a amplitude da caixa no gráfico evidencia uma maior dispersão nos valores das ocorrências do conjunto do *corpus* TED-SHAPE, ao passo que palestrantes STEM parecem recorrer ao conjunto dessa categoria de modo mais padronizado.

Já o conjunto de pacotes lexicais utilizados para identificação/foco apresentam uma maior dispersão nas ocorrências do conjunto do *corpus* TED-STEM, ou seja, alguns dos pacotes lexicais dessa categoria são utilizados com frequência maior em palestras STEM do que em SHAPE. Apesar disso, não foram identificados valores discrepantes nesse *corpus*. Adicionalmente, é interessante observar que o valor da linha de tendência central para a categoria de identificação/foco é semelhante em ambos os *corpora* de estudo, sinalizando

uma frequência de uso similar entre palestrantes de ambas as áreas. Essas funções estão discutidas em detalhes nas Seções 4.3.1, 4.3.2 e 4.3.3, respectivamente. Na Seção 4.3.4, concluo com as considerações finais sobre o que os dados encontrados revelaram a respeito da organização discursiva dos palestrantes das áreas STEM e SHAPE.

4.3.1 Introdução de tópicos

O Gráfico 12 compara as ocorrências dos pacotes lexicais associados à função de introduzir novos tópicos ao discurso nas TED Talks das áreas STEM e SHAPE. O conjunto considerável de pacotes lexicais distintos encontrados nessa categoria sugere que o estabelecimento de uma “linha mestra” para auxiliar os palestrantes na sinalização dos limites e da estrutura esquemática do discurso ocorre de maneira altamente padronizada em ambas as áreas. Isso pode ser atribuído a necessidade de diminuir a gama de assuntos abordados pelo palestrante para que o discurso possa ser desenvolvido de modo adequado dentro do tempo de fala disponível. É preciso considerar também que parte dessa padronização se deve aos ensaios, preparação e até mesmo memorização do discurso, assim como ao fato de que, ao evidenciarem regularmente os tópicos de seus discursos, palestrantes TED evitam ambiguidades na mensagem transmitida e facilitam a compreensão do público-alvo.

Gráfico 12 – Pacotes lexicais de introdução de tópicos

Fonte: Elaborado pelo autor

Com base no gráfico acima, observa-se que os palestrantes de ambas as áreas recorrem a um conjunto semelhante de pacotes lexicais para anunciar seus objetivos discursivos, antecipar etapas do texto e sinalizar mudanças de tópicos. Além disso, muitos dos padrões utilizados para iniciar novos tópicos são extensões das expressões sinalizadoras de desejos e intenções (*cf.* Seções 4.2.2 e 4.2.5, respectivamente) e elucidam de forma mais clara⁵⁷ o intuito dos palestrantes em iniciar novos tópicos em seus discursos.

Em relação ao uso dessas expressões no discurso, os dados apontam que palestrantes TED tendem a explicitar o tema central logo no início de suas apresentações. Em ambos os *corpora* de estudo, aproximadamente 15% das ocorrências do conjunto de pacotes lexicais relacionados à introdução de tópicos ocorrem nos momentos iniciais das palestras⁵⁸ e são marcadas por expressões de cumprimento (*i.e. good morning*), conforme ilustrado no exemplo a seguir.

⁵⁷ Apesar de haver sobreposição, é possível perceber que a introdução de um novo tópico ao discurso se mostra mais evidente no encadeamento *I want to talk + (about)* do que no caso do pacote lexical *and I want to*, que foi considerado com expressão de desejo.

⁵⁸ O termo abertura é utilizado aqui de modo genérico para se indicar as 50 primeiras palavras de um texto, uma vez que este é o limite máximo de co-texto disponibilizado nas linhas de concordância do AntConc®.

- [74] Good morning, **I'm going to talk about** architecture and prefabrication. How a house could be built? And how we could improve with the current technology, the quality of our living spaces, the affordability, and share these with greater markets and people.

(TED_SHAPE_Arquitetura_08)

Embora encadeamentos como *I want to talk + (about)* e *I want to tell + (you)* evidenciem os verbos mais recorrentes na sinalização de novos tópicos ao discurso, o pacote lexical *I would like to* é notável por ser empregado em conjunto com uma ampla variedade de sintagmas verbais. Essa combinação serve não apenas para introduzir novos tópicos no discurso (Exemplo 75), mas também para sinalizar a conclusão da fala dos palestrantes (Exemplo 76) e expressar desejos (Exemplo 77).

- [75] Hello. **I would like to** start my talk with actually two questions, and the first one is: How many people here actually eat pig meat? Please raise your hand – oh, that's a lot. And how many people have actually seen a live pig producing this meat?

(TED_STEM_Engenharia_13)

- [76] **I would like to** end with this thought: That when we reject the single story, when we realize that there is never a single story about any place, we regain a kind of paradise.

(TED_SHAPE_Literatura_13)

- [77] And now **I would like to** ask you to formally help us, because we cannot do this alone.

(TED_SHAPE_Estudos_Gênero_13)

Já a elevada recorrência do pacote lexical *let me tell you* sugere que a introdução de novos tópicos por meio de narrativas (*storytelling*) constitui uma estratégia discursiva comum entre palestrantes de ambas as áreas para alcançar um público amplo e diversificado. Os exemplos a seguir ilustram como palestrantes TED recorrem a esse pacote lexical para sinalizar o início de uma história, possivelmente com o objetivo de promover o engajamento do público-alvo com a narrativa.

[78] This is me, age five, in front of that hospital. I don't think they thought I'd grow up to sell insurance. So **let me tell you** how that happened. In 2008, I was working for the Ministry of Agriculture of Rwanda, and my boss had just been promoted to become the minister. She launched an ambitious plan to start a green revolution in her country, and before we knew it, we were importing tons of fertilizer and seed and telling farmers how to apply that fertilizer and plant.
(TED_STEM_Agronomia_05)

[79] So what do I mean by "productive conflict"? Well, **let me tell you** a story. In 2016, a store employee for an outdoor retailer – I'll call her "Alex" – Alex approached her boss and asked for a raise. Now, she was told her pay was fairly standard for her position and that her boss didn't even have the authority to give such a raise. And that was supposed to be the end of the conversation.
(TED_SHAPE_Gestão_Pessoas_18)

Nos exemplos acima, é interessante notar que as narrativas são apresentadas de maneira semelhante. Em ambos os casos, por exemplo, o público-alvo é contextualizado temporalmente e é apresentada uma situação problema. Contextualizar tais informações é relevante, pois elas instigam a curiosidade do público-alvo, incentivando-os a prestar atenção na história que está sendo contada.

Outros pacotes lexicais pertencentes a essa categoria também apontaram a preocupação dos palestrantes TED em narrar histórias que despertem a curiosidade e mantenham ou recuperem a atenção e o interesse do público-alvo (*i.e. (I'm) + going to talk about, I want to tell + (you), to share with you*). É interessante observar como palestrantes de ambas as áreas empregam adjetivos/sintagmas adjetivais (*i.e. quite a story*), sintagmas verbais (*i.e. reveal them in a minute*) e expressões de pessoalidade (*i.e. a little of my story*) para atrair a atenção e curiosidade de seu público-alvo em relação ao desfecho da história apresentada.

[80] And the last story **I'm going to talk about** is quite a story.
(TED_STEM_Design_24)

[81] I'm going to reveal them to you in a minute. But first **I want to tell you** how they were discovered.
(TED_STEM_Biologia_106)

[82] So I'd like **to share with you** a little of my story today, how I ended up here, how it connects with Davinia and, hopefully, with you.
(TED_SHAPE_Ciências_Políticas_36)

A recorrência desses padrões corrobora o argumento de Anderson (2016) de que o uso de narrativas se configura como um recurso valioso, empregado pelos palestrantes TED para

inserir breves explicações que ilustrem ou reforcem uma ideia para o público-alvo. Essa abordagem contribui para a eficácia na comunicação das mensagens e promove maior engajamento por parte dos espectadores.

O pacote lexical *to share with you*, por sua vez, se mostrou como um padrão recorrente para iniciar tanto a narração de eventos, quanto em disseminar informações e compartilhar ideias e projetos de sucesso. A recorrência desse padrão em ambos os *corpora* de estudo evidencia a tentativa desses palestrantes em estabelecer uma conexão com seu público-alvo e fomentar a troca de conhecimento e a colaboração entre as partes envolvidas.

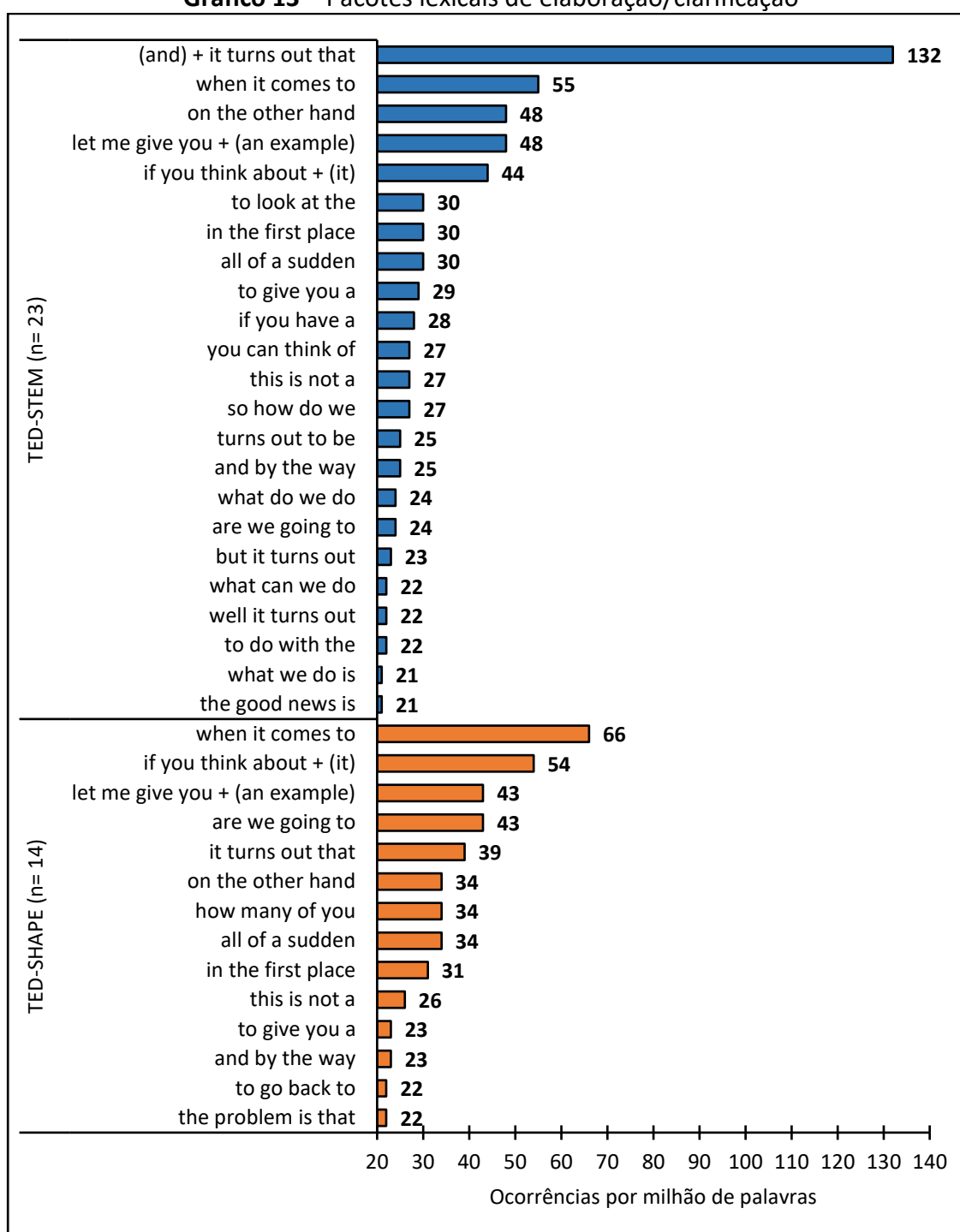
[83] As a speech scientist, I'm fascinated by how the voice is produced, and I have an idea for how it can be engineered. That's what I'd like **to share with you**.
(TED_STEM_Engenharia_66)

[84] Hello, Manchester. I'm so happy to be here. My name is Sophia Wallace. I'm an artist. And I'm here today **to share with you** a project that I hope will empower you personally, and by extension those that you love.
(TED_SHAPE_Estudos_Gênero_06)

Após concluir a análise dos pacotes lexicais mais frequentes utilizados para introduzir novos tópicos nas TED Talks, a seção subsequente abordará os pacotes lexicais identificados que desempenham as funções de elaboração e clarificação de tópicos no discurso. Essa discussão aprofundará a compreensão das estratégias empregadas pelos palestrantes para garantir a clareza e a coesão em suas exposições, contribuindo para a eficácia na comunicação das ideias apresentadas.

4.3.2 Elaboração/clarificação de tópico

A presente categoria abrange os pacotes lexicais que auxiliam palestrantes TED a indicar relações de coesão interna no texto, como adição, contraste, causa, exemplificação, elaboração e explicação, visando garantir que o público-alvo compreenda efetivamente o significado pretendido pelos palestrantes. O Gráfico 13 ilustra os pacotes lexicais distintos identificados em cada um dos *corpora* de estudo e revela que palestrantes TED de ambas as áreas empregam um amplo conjunto de padrões recorrentes para assegurar a clareza e a compreensão das ideias apresentadas.

Gráfico 13 – Pacotes lexicais de elaboração/clarificação

Fonte: Elaborado pelo autor

No que se refere à maior padronização no *corpus* TED-STEM, é interessante destacar também a maior quantidade de pacotes lexicais distintos compostos pelo padrão lexicogramatical *it turns out*. Especificamente, o encadeamento *(and) + it turns out that* revelou-se extremamente recorrente na elaboração do discurso por meio de explicações adicionais sobre o tópico em discussão. Nesse contexto, Simpson (2004, p. 56) esclarece que

em registros acadêmicos orais, esse padrão tende a “transformar os enunciados de fatos em enunciados sobre a descoberta desses fatos⁵⁹”. Logo, o amplo conjunto e recorrência de estruturas padronizadas compostas por *it turns out* no *corpus* TED-STEM sugerem um certo “ar de surpresa/imprevisibilidade” durante o discurso desses palestrantes, possivelmente com o objetivo de manter o interesse e a atenção de seu público-alvo.

Os exemplos a seguir demonstram como os padrões *but it turns out*, *it turns out to be* e *turns out to be* são empregados para indicar a apresentação de informações e/ou descobertas que contrariam, pelo menos em certa medida, conhecimentos e pressupostos previamente aceitos. Ao enquadrar informações dessa maneira, os palestrantes conseguem sinalizar ao público-alvo que o que está prestes a ser apresentado é algo surpreendente ou inesperado. Utilizando-se de uma abordagem narrativa e enfatizando revelações e descobertas, palestrantes TED prepararam seu público-alvo para conclusões mais impactantes em seus discursos, contribuindo para o envolvimento e atenção dos ouvintes.

[85] We grew up believing that the hymen is a proof of virginity. **But it turns out** we were wrong. What we discovered is that the popular story we're told about female virginity is based on two anatomical myths.

(TED_STEM_Medicina_78)

[86] I'm going to teach you how to stop a coup, OK? Now, where to begin? Alright, how about this: **it turns out that** one of the main safeguards of US democracy is not in the constitution at all. It's not in the law at all. It's actually just a little tradition, it's a little custom.

(TED_SHAPE_Direito_62)

[87] Now, rabies, as terrible as it is, **turns out to be** a pretty nice virus in this case.

(TED_STEM_Biologia_22)

Outro padrão empregado pelos palestrantes TED para sinalizar desdobramentos inesperados durante a narração de eventos foi o pacote lexical *all of a sudden*. Nos exemplos a seguir, observa-se como esse padrão também auxilia os palestrantes na sinalização de surpresa, imprevisibilidade e tensão em suas narrativas, contribuindo para a manutenção do interesse e da atenção do público-alvo.

⁵⁹ No original: “[...] *transforms them from mere statements of facts, to statements about the discovery of these facts*”.

[88] You may have seen some of this before, but we're behind a coral reef, and there's this rock out there, a lot of sand, fishes swimming around. And **all of a sudden**, this octopus appears, and now it flashes white, inks in my face and jets away.

(TED_STEM_Biologia_114)

[89] Later we found the leading cause of illness and death in this particular region is through respiratory failure. So **all of a sudden**, we had two problems. We were there originally to look at toilets and get human waste off the ground, that's fine. But **all of a sudden** now there was a second problem: How do we actually get the smoke down?

(TED_SHAPE_Arquitetura_67)

A recorrência dos pacotes lexicais apresentados acima reforça a argumentação de que a narração de histórias é uma característica comumente empregada por palestrantes TED para se conectarem com o público-alvo e transmitirem informações de maneira mais envolvente, acessível e de fácil memorização, promovendo maior empatia e entendimento.

No âmbito dos registros orais e discurso acadêmico, outra estratégia de engajamento frequente é o uso de perguntas retóricas (Flowerdew; Miller, 1997; Hyland, 2010), pois é através delas que pesquisadores buscam aproximar os interesses de seus estudos junto ao público-alvo, fortalecendo a conexão e o engajamento da audiência com o discurso. No que se refere às TED Talks, o pacote lexical *are we going to* é um padrão recorrente que evidencia o uso de perguntas retóricas como um meio de elaboração discursiva. Os exemplos a seguir demonstram como esse pacote lexical é empregado para antecipar e instigar questionamentos em seu público-alvo, além de, simultaneamente, fornecer informações adicionais que auxiliem na clarificação das ideias apresentadas no discurso.

[90] It has weeping water right now. Liquid water on its surface. It has organics. It has all the right conditions. So what **are we going to** do next? We're going to launch a series of missions to begin that search for life on Mars.

(TED_STEM_Astronomia_20)

[91] So how **are we going to** solve this problem of knowledge polarization? One obvious tactic is to try to fix our technology, to redesign our digital platforms, so as to make them less susceptible to polarization.

(TED_SHAPE_Filosofia_07)

Uma variação notável no uso desse pacote lexical consiste na maior recorrência do padrão *what + are we going to* em TED Talks das áreas STEM, abrangendo aproximadamente 47% das ocorrências desse padrão. Por outro lado, os palestrantes das áreas SHAPE

demonstraram uma preferência mais acentuada pelo padrão *how + are we going to*, observado em 40% das ocorrências desse padrão nesse *corpus*. Tal variação sugere que o discurso de palestrantes STEM tende a se concentrar na apresentação ou discussão de ações específicas [o que faremos], enquanto palestrantes SHAPE preferem elaborar sobre o processo de como algo será realizado [como faremos].

No *corpus* TED-SHAPE, um padrão particularmente recorrente que também promove um discurso mais interativo entre palestrantes e o público-alvo foi o pacote lexical *how many of you*. No exemplo a seguir, o palestrante recorre a esse padrão como meio de, inicialmente, levantar informação quantitativas junto ao público-alvo e, a partir disso, elaborar as informações que serão apresentadas em seus discursos.

[92] **How many of you** have heard someone say they don't care about their privacy because they don't have anything to hide? Go on. Now, **how many of you** use any kind of encryption software? Raise your hand.

(TED_SHAPE_Direito_31)

O maior conjunto de pacotes lexicais distintos encontrando no *corpus* TED-STEM indica, contudo, que o emprego de perguntas retóricas como meio iniciar elaborações ou clarificações do discurso ocorre de maneira mais padronizada em TED Talks relacionadas às áreas STEM. Os pacotes lexicais *so how do we* e *what do we do* são exemplos de padrões empregados no esclarecimento de como determinados procedimentos são realizados.

[93] So a future strategy to improve nanoparticles is to temporarily disarm the immune cells in the liver. **So how do we** disarm these cells? Well, we looked at drugs that were already clinically approved for other indications to see if any of them could stop the immune cells from eating the nanoparticles.

(TED_STEM_Biotecnologia_25)

[94] So, we've got now this flour. And **what do we do**? We add some water. In stage one, we weigh it. In stage two, we add water and salt to it, mix it together, and we create something that we call "clay."

(TED_STEM_Química_12)

Outra evidência da maior padronização de perguntas retóricas no *corpus* TED-STEM é a recorrência do pacote lexical *what can we do*. Diferentemente das expressões apresentadas acima, o emprego desse padrão sinaliza momentos nos quais os palestrantes enfatizam para

seu público-alvo as possíveis aplicações ou ações a serem realizadas em relação ao tópico em discussão. Nesse sentido, o uso desse pacote lexical estaria associado não apenas à função de elaborar o discurso, mas também de destacar habilidades perante o público-alvo.

- [95] Now the chip is really beautiful, but **what can we do** with it? We can get incredible functionality inside these little chips. Let me show you. We could, for example, mimic infection, where we add bacterial cells into the lung.
(TED_STEM_Medicina_33)

Embora os padrões compostos por *(it) turns out* encontrados não sejam estruturas interrogativas, palestrantes TED tendem a utilizá-los no momento de elaborar as respostas ou explicações para as perguntas retóricas apresentadas em seus discursos. Em especial, a recorrência do pacote lexical *well it turns out*, no *corpus* TED-STEM, enfatiza uma maior padronização no uso de perguntas retóricas no discurso desses palestrantes.

- [96] So how does electricity ignite residential fires? **Well, it turns out** that the main causes are faulty and misused appliances and electrical wiring. Our invention had to address all of these issues.
(TED_STEM_Engenharia_37)
- [97] When we say math is painful, there's some truth to it for some people. But where does this math anxiety come from? **It turns out that** math anxiety is contagious. When adults are worried about math, the children around them start worrying, too.
(TED_STEM_Biologia_120)
- [98] Music can cause riots, topple governments, raise the hairs on the back of your neck. What can science tell us about that? **It turns out that** when your brain is enjoying a piece of music that might even give you the chills, it's awash with a neurotransmitter called dopamine.
(TED_SHAPE_Música_07)

Além do uso perguntas retóricas, os dados apontam que palestrantes STEM também recorrem a padrões como *to look at the* e *what we do is* para sinalizar elaboração/clarificação de ações e procedimentos. Os exemplos a seguir ilustram como esses pacotes lexicais auxiliam os palestrantes na explicação de estágios ou processos metodológicos de modo a facilitar a compreensão do público-alvo. Vale destacar aqui que o significado do pacote lexical *to look at the*, no *corpus* TED-STEM, tende a estar mais associado a verbos de natureza mental (*e.g. investigate* e *consider*), do que ao conceito primário de visão/percepção (Exemplo 99).

[99] We put them in a materials testing machine **to look at the** stress strain analysis and showed that they can withstand forces 800 times their body weight.

(TED_STEM_Biologia_111)

[100] Now reverse osmosis, it's the reverse of this natural process. And in order to achieve this reversal, **what we do is** we apply a pressure to the high-concentration side and in doing so, we drive the water the opposite direction.

(TED_STEM_Química_18)

Outro padrão empregado na elaboração do discurso em TED Talks foi o pacote lexical *in the first place*. Em ambas as áreas, palestrantes TED recorrem a essa sequência para enfatizar a importância ou centralidade de determinada informação frente ao público-alvo. Os exemplos a seguir ilustram, respectivamente, como esse pacote lexical auxilia palestrantes TED a:

(i) enfatizar relações lógicas entre informações ou eventos apresentados no discurso;

[101] Aging is a side effect of being alive **in the first place**, which is to say, metabolism. This is not a completely tautological statement; it's a reasonable statement. Aging is basically a process that happens to inanimate objects like cars, and it also happens to us.

(TED_STEM_Engenharia_07)

(ii) reforçar, perante o público-alvo, a importância/necessidade de discutir determinado tópico;

[102] So every Saturday morning, when this little girl used to rush down the stairs to watch that show, she felt a little less alone because they had fulfilled a need to see herself in the world outside. So, let's talk about why that need existed **in the first place**.

(TED_SHAPE_Estudos_Mídias_07)

(iii) evidenciar as origens de determinada situação e estabelecer uma conexão com algo que foi discutido anteriormente.

[103] So in Stockton, as I mentioned, we have historically had problems with violent crime. In fact, that's why I decided to run for office **in the first place**.

(TED_SHAPE_Ciências_Políticas_51)

Palestrantes TED também procuram organizar seus discursos frente ao público-alvo, elaborando discussões de certos tópicos. Em particular, os pacotes lexicais *if you think about* + (*it*) e *when it comes to* são exemplos de padrões empregados quando se deseja transitar de uma discussão generalizada para aspectos específicos sobre um tema. Os exemplos a seguir ilustram como esses padrões são empregados para delimitar a discussão de pontos específicos do tema em questão: os custos na produção de energia e a relação entre tecnologias e o trabalho do agente comunitário de saúde, respectivamente.

[104] Now, **if you think about** how we deal with energy pricing, for instance, we consider emissions taxes, which means we're imposing the costs of pollution on people who actually use that energy.

(TED_SHAPE_Economia_65)

[105] People are worried that technology is going to steal our jobs, but **when it comes to** community health workers, technology has actually been vital for creating jobs. Without technology – without this smartphone, without this rapid test – it would have been impossible for us to be able to employ A.B. and Musu.

(TED_STEM_Medicina_92)

Ainda em relação ao encadeamento *if you think about* + (*it*), esse padrão também sinaliza a apresentação de conclusões lógicas ou pontos de vista do palestrante, de modo a tentar influenciar o público-alvo a aceitar/concordar com as opiniões e argumentos apresentados, conforme ilustra o exemplo a seguir.

[106] There's enormous marketing of prescription drugs to people like you and me, which, **if you think about it**, makes no sense at all, since we can't buy them. Why do they market to us if we can't buy them?

(TED_STEM_Psicologia_16)

Os dados também indicam que palestrantes TED utilizam certos padrões que sinalizam elaboração/clarificação do discurso por meio da explicação de como diferentes eventos estão relacionados. No *corpus* TED-STEM, o pacote lexical mais recorrente para tal propósito foi o fragmento da expressão [*have*] *to do with the*. No exemplo a seguir, o palestrante argumenta uma possível conexão entre dois eventos/elementos do discurso (compreensão e mudança de perspectiva); no entanto, a explicação sobre como esses eventos/elementos estão

relacionados tende a não ser detalhada, indicando que, possivelmente, a relação de causa e efeito apresentada estaria implícita.

[107] My claim is that understanding has **to do with the** ability to change your perspective. If you don't have that, you don't have understanding. So that is my claim.
(TED_STEM_Matemática_27)

Palestrantes SHAPE, por sua vez, recorrem ao padrão *the problem is that* para elucidar o motivo pelo qual o problema ou lacuna que eles abordam em seu discurso é particularmente importante.

[108] Our work in Botswana showed for a modest investment, the economy would gain 1.3 billion dollars over the lifetime of the children. That was 150 times return on investment. But part of **the problem is that** value is generated in the future, but we need the money now to deliver it.
(TED_SHAPE_Economia_08)

Possivelmente, a maior recorrência desse pacote lexical no *corpus* TED-SHAPE esteja relacionada ao fato de que os problemas apresentados por esses palestrantes não são tão evidentes para o público-alvo, especialmente quando comparados à natureza empírica e precisa da contraparte STEM. Portanto, a expressão *the problem is that* se mostra como um padrão útil ao qual palestrantes SHAPE recorrem para garantir que o público-alvo compreenda determinadas relações no discurso de forma clara e simplificada.

Outra particularidade associada à organização discursiva nas TED Talks das áreas SHAPE foi a maneira como certos assuntos são retomados ao longo do discurso. O pacote lexical *to go back to*, por exemplo, sinaliza a retomada de argumentos – mesmo que externos a palestra, conforme é possível observar no exemplo a seguir.

[109] We need a digital economy where our data and our attention is not for sale to the highest-bidding authoritarian or demagogue. So, **to go back to** that Hollywood paraphrase, we do want the prodigious potential of artificial intelligence and digital technology to blossom, but for that, we must face this prodigious menace, open-eyed and now.
(TED_SHAPE_Sociologia_11)

Os dados encontrados também revelaram alguns padrões recorrentes acerca de como palestrantes TED organizam seus discursos por meio de exemplificação, comparação e

contraste entre as ideias apresentadas no texto. Tais estratégias são importantes, especialmente em discursos de popularização da ciência, pois auxiliam os autores na simplificação da densidade informacional em seus discursos e facilitam a compreensão de um público majoritariamente leigo (Calsamiglia, 2003).

No que diz respeito à exemplificação no discurso das TED Talks, os padrões mais utilizados por palestrantes foram o encadeamento *let me give you + (an example)* e o pacote lexical *to give you a*. A alta frequência desses, em ambos os *corpora* de estudo, sinaliza que palestrantes TED recorrem a demonstrações que os ajudem a simplificar as explicações de conceitos e ideias complexas para seu público-alvo. No exemplo abaixo, com o objetivo de elucidar a relação entre empresas dos setores de comunicações, tecnologias e entretenimento, o palestrante faz uso do padrão *let me give you + (an example)* para relembrar um caso real que ilustre o argumento de como esses setores de negócios influenciam uns aos outros.

[110] There's been an uneasy relationship between the TV business and the tech business, really ever since they both turned about 30. We go through periods of enthrallment, followed by reactions in boardrooms, in the finance community best characterized as, what's the finance term? Ick pooey. **Let me give you an example** of this. The year is 1976, and Warner buys Atari because video games are on the rise. The next year they march forward and they introduce Qube, the first interactive cable TV system, and the New York Times heralds this as telecommunications moving to the home, convergence, great things are happening. Everybody in the East Coast gets in the pictures – Citicorp, Penney, RCA – all getting into this big vision.
(TED_SHAPE_Estudos_Mídias_19)

Já no Exemplo 111, o palestrante emprega o padrão *let me give you* para elaborar uma analogia que relaciona o comportamento de objetos inanimados em um modelo proposto pela mecânica quântica ao comportamento de pessoas em um elevador. Ao utilizar esse padrão para apresentar uma situação completamente distinta, porém de fácil compreensão, o palestrante consegue explicar um conceito extremamente complexo (*i.e.* sobreposição quântica) de forma rápida, simples e satisfatória para seu público-alvo.

[111] This device has the ability to be in a quantum superposition, but it needs a little help to do it. Here, **let me give you** an analogy. You know how uncomfortable it is to be in a crowded elevator? I mean, when I'm in an elevator all alone, I do all sorts of weird things, but then other people get on board and I stop doing those things because I don't want to bother them, or, frankly, scare them. So quantum mechanics says that inanimate objects feel the same way. The fellow passengers for inanimate objects are not just people, but it's also the light shining on it and the wind blowing past it and the heat of the room. And so we knew, if we wanted to see this piece of metal behave quantum mechanically, we're going to have to kick out all the other passengers.

(TED_STEM_Física_01)

Outro padrão recorrente para indicar exemplificações foi *to give you a*. Um aspecto interessante acerca desse pacote lexical é a sua associação com aparatos sensoriais, especialmente por meio do padrão *to give you a + sense of*. No exemplo a seguir, o palestrante emprega esse padrão para sinalizar uma breve demonstração sensorial [auditiva] que auxilia o público-alvo na melhor compreensão de um procedimento explicado previamente.

[112] In my lab in Princeton, we bring people to the fMRI scanner and we scan their brains while they are either telling or listening to real-life stories. And **to give you a** sense of the stimulus we are using, let me play 20 seconds from a story that we used, told by a very talented storyteller, Jim O'Grady.

(TED_SHAPE_Estudios_Linguísticos_11)

O uso de comparações, por outro lado, apresentou maior padronização no discurso das TED Talks vinculadas às áreas STEM, especialmente por meio dos pacotes lexicais *you can think of* e *if you have a*. Tais padrões apontam para uma preocupação em utilizar analogias que recontextualizem conceitos técnicos e elaborem explicações utilizando elementos do cotidiano que ajudem o público-alvo a relacionar ideias e compreender melhor o discurso.

[113] And there's a structure that sits on the left and the right side of your brain, called the hippocampus. And **you can think of** the hippocampus almost like the informational inbox of your brain. It's very good at receiving new memory files and then holding on to them.

(TED_STEM_Medicina_68)

[114] Stress resilience is the active biological process that allows us to bounce back after stress. Similar to **if you have a** cold and your immune system fights it off.

(TED_STEM_Química_14)

Por sua vez, o modo como palestrantes TED sinalizam contraste entre diferentes proposições do discurso foi evidenciado, principalmente, pelo pacote lexical *on the other hand*, um padrão comum tanto em registros de discurso acadêmico escrito (Bocorny et al., 2021; Durrant, 2015; Hyland, 2008a; Hyland; Jiang, 2018) quanto oral (Liu; Chen, H. 2020b; Nesi; Basturkmen, 2006; Wang, 2017). Os exemplos a seguir denotam como palestrantes TED recorrem a esse padrão para sinalizar diferenças entre argumentos, processos, modelos, para o público-alvo.

[115] We separate the elements from nature and convert them into an irreversible state. That's industrial production. Nature, **on the other hand**, works in a totally opposite way. The natural system produces by bringing elements together, atom by atom.

(TED_STEM_Botânica_07)

[116] There are three main types of communication, spoken, gestural – so things like sign language – and graphic communication. Spoken and gestural are by their very nature ephemeral. It requires close contact for a message to be sent and received. And after the moment of transmission, it's gone forever. Graphic communication, **on the other hand**, decouples that relationship. And with its invention, it became possible for the first time for a message to be transmitted and preserved beyond a single moment in place and time.

(TED_SHAPE_Antropologia_07)

Além da sequência *on the other hand*, palestrantes das áreas STEM também sinalizam contraste em seus discursos por meio dos pacotes lexicais *but it turns out* e *the good news is*. No que diz respeito ao primeiro, o exemplo abaixo apresenta um contraste entre o conhecimento convencional do público-alvo e as novas descobertas e avanços científicos a respeito do tópico.

[117] And if you look at Mars, it looks really dry, it looks like the entire planet is a desert. **But it turns out** that it's not. The soil alone on Mars contains up to 60 percent water.

(TED_STEM_Engenharia_73)

Já o pacote lexical *the good news is* é interessante, pois auxilia no destaque de uma solução ou uma perspectiva otimista frente a um problema ou desafio. No exemplo a seguir, o palestrante recorre a esse pacote lexical para atrair a atenção do público-alvo para uma proposta na qual acredita haver um grande potencial de contribuir para a solução do

problema em pauta, isso ajuda o palestrante a manter seu público-alvo envolvido e motivado com o discurso, mesmo quando o problema apresentado aparenta ser complexo e desafiador.

[118] But the proposal here is to use this defense mechanism to protect our rivers, so we don't need to rehabilitate them. And **the good news is** we have technology that can do this. It's with membranes. Membranes that can separate salt and water.

(TED_STEM_Química_18)

Ao ressaltarem notícias e aspectos positivos, observa-se que palestrantes STEM buscam estabelecer uma conexão emocional com seu público-alvo, fazendo com que seus ouvintes se sintam esperançosos frente aos problemas apresentados.

Palestrantes TED também elaboram/clarificam seus discursos ao apresentarem informações adicionais ao público-alvo, especialmente junto ao pacote lexical *and by the way*, que demonstrou ser um padrão recorrente em ambos os *corpora* de estudo. Vale apontar também que o uso desse padrão sugere uma necessidade de clarificação que surgiu de maneira espontânea e informal no decorrer do discurso.

[119] They have a very nice landscape, you see the Jordanian Mountains on the one hand, and the Jordan Valley on the other hand, and a good, mild winter and a nice, hot summer, which is an excellent condition to mass-produce those creatures. **And by the way**, mass-production – it is not genetic manipulation. There are no GMOs, genetically modified organisms, whatsoever.

(TED_STEM_Agronomia_19)

[120] Here's some advice from one of the largest investors in the United States. This guy runs the China Investment Corporation. He is the main buyer of U.S. Treasury bonds. And he gave an interview in December. Here's his first bit of advice. And here's his second bit of advice. **And, by the way**, the Chinese Prime Minister reiterated this at Davos last Sunday.

(TED_SHAPE_Economia_93)

Outro padrão recorrente entre palestrantes TED para a clarificação de ideias e conceitos foi o pacote lexical *this is not a*. Os exemplos a seguir ilustram como esse pacote lexical é utilizando junto a outros dispositivos de clarificação ou reiteração do discurso (*i.e. again* e *and I'm telling you*) para prevenir uma possível interpretação errônea por parte do público-alvo.

[121] Heart disease kills more women at all ages than breast cancer. And the breast cancer campaign – again, **this is not a** competition. We're trying to be as good as the breast cancer campaign. We need to be as good as the breast cancer campaign.

(TED_STEM_Medicina_80)

[122] I believe every candidate will have to disclose where they stand on the abolition of fossil fuels. Because this question has to enter our mainstream politics. And I'm telling you, **this is not a** question of climate policy or environmental agenda. It's about the country that we want and the cities that we have and the cities that we want and who makes that choice.

(TED_SHAPE_Desenvolvimento_Internacional_18)

Nesta seção, os dados apresentados evidenciam como pacotes lexicais constituem ferramenta poderosa para elaboração/clarificação do discurso em TED Talks. Palestrantes de ambas as áreas recorrem a um conjunto considerável de padrões distintos que os auxiliam na comunicação de ideias complexas que precisam ser apresentadas em um curto período de tempo. Logo, alta a recorrência de padrões linguísticos, como (*and*) + *it turns out that, when it comes to*, dentre outros, se mostraram essenciais para tornar o discurso mais claro, conciso e envolvente, mantendo o público-alvo engajado e interessado até o final da fala do palestrante.

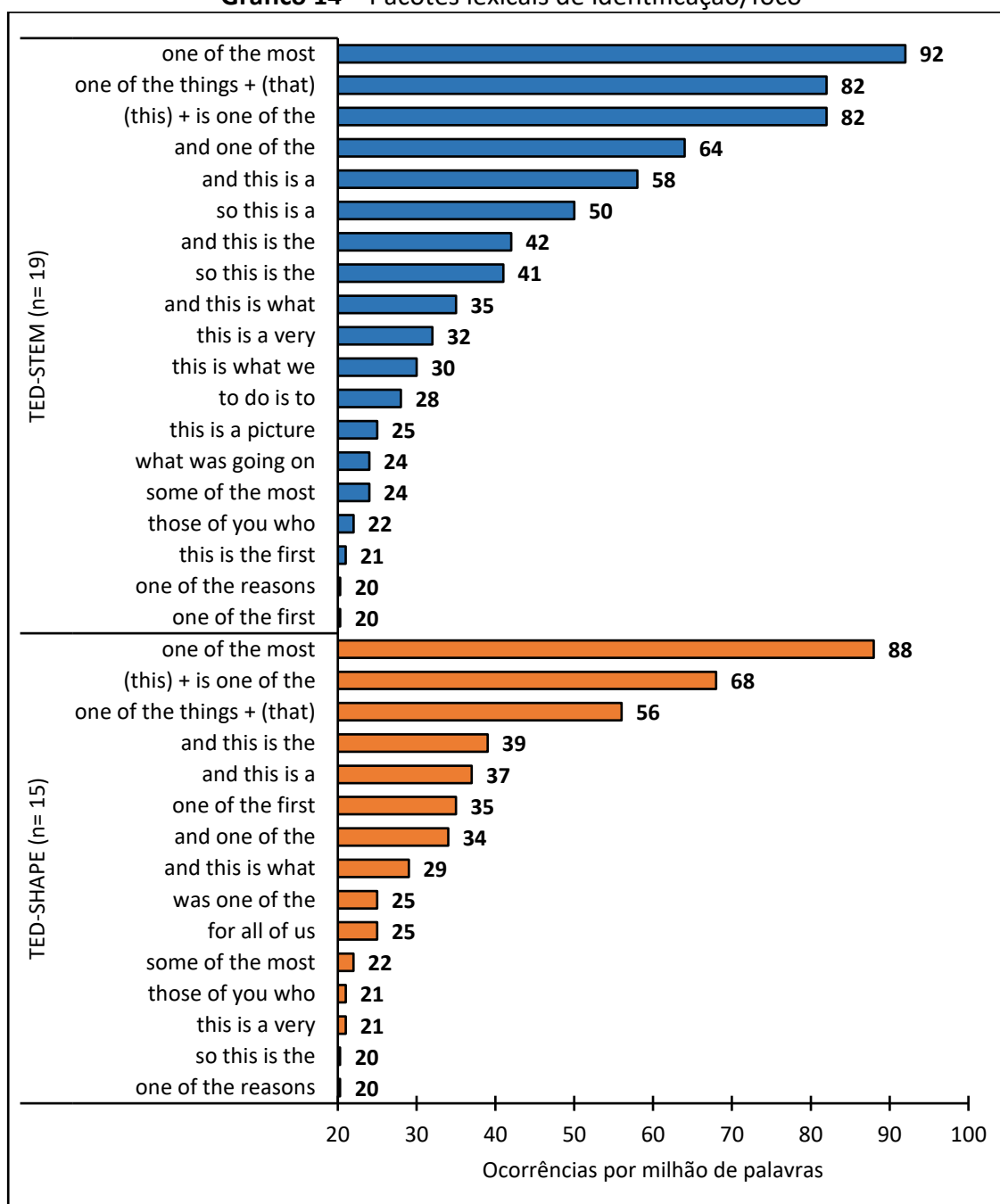
Na seção seguinte, discuto sobre os resultados obtidos sobre o uso dos pacotes lexicais com a função de identificação e como eles auxiliam os palestrantes de TED Talks na organização dos seus discursos.

4.3.3 Identificação/foco

A terceira categoria de organização discursiva compreende os pacotes lexicais relacionados à identificação e ao foco a entidades no discurso. Os padrões abordados nesta seção são comumente empregados para destacar elementos relevantes, estabelecer conexões lógicas e sintetizar conceitos e explicações no discurso (Biber, 2006a; Biber; Barbieri, 2007). No que tange às TED Talks, os dados ilustrados no Gráfico 14 revelam claramente que palestrantes de ambas as áreas empregam um amplo conjunto de padrões para ressaltar informações pertinentes e simplificar a compreensão de seu público-alvo. Nesse sentido, é possível estabelecer um paralelo entre a categoria de introdução/foco e os marcadores de relevância orientados à audiência (*audience-oriented relevance markers*), termo utilizado por

Crawford Camiciottoli (2014) para se referir às diferentes expressões (*e.g. this is important, the main point is*) que auxiliam palestrantes a indicarem ao público-alvo quais informações são particularmente importantes em seus discursos.

Gráfico 14 – Pacotes lexicais de identificação/foco



Fonte: Elaborado pelo autor

Os dados supracitados revelam que muitos dos padrões empregados para identificação/foco consistem em dêiticos discursivos (*e.g. and this is a, and this is what*) e, em

menor quantidade, em adjetivos de avaliação e relevância (*i.e. this is the first, one of the most*). Isso reflete a preocupação dos palestrantes em estabelecer engajamento e conexão emocional com seu público-alvo, alinhando-se à tendência discursiva dos gêneros de popularização da ciência de exaltar aspectos positivos e otimistas das ideias apresentadas (Caliendo, 2014; Scotto di Carlo, 2014).

No que diz respeito à sinalização de relevância no discurso, dois pacotes lexicais recorrentes em TED Talks de ambas as áreas foram os padrões *one of the most* e *some of the most*, sendo o primeiro altamente recorrente em ambos os *corpora* de estudo. Como pode ser observado nos exemplos a seguir, esses pacotes lexicais ajudam os palestrantes no destaque das principais características que diferenciam uma entidade, processo ou instrumento de seus respectivos grupos.

[123] The next time you see a spider web, please, pause and look a little closer. You'll be seeing **one of the most** high-performance materials known to man.

(TED_STEM_Engenharia_12)

[124] There was an article last year in the Proceedings of the National Academy of Sciences, which is **one of the most** prestigious scientific journals published in the West.

(TED_SHAPE_Sociologia_01)

[125] The Hubble Space Telescope, for the last 25 years it's been producing **some of the most** detailed views of our distant universe.

(TED_STEM_Astronomia_03)

[126] At MIT Media Lab they are doing **some of the most** cutting-edge work in biomechatronics, developing motorized limbs for amputees.

(TED_SHAPE_Artes_13)

Outro padrão empregado para sinalizar avaliação perante o público-alvo é o pacote lexical *this is a very*. Nos exemplos abaixo, pode-se notar que essa sequência auxilia os palestrantes TED a, primeiramente, ressaltar uma propriedade ou aspecto relevante do tema e, posteriormente, desenvolver a discussão sobre o assunto.

[127] Phosphorus is a chemical element that is essential to life. **This is a very** important point. I'd like everyone to understand precisely what the phosphorus issue is. Phosphorus is a key component in several molecules, in many of our molecules of life.

(TED_STEM_Agronomia_08)

- [128] There's ample evidence that gender- and ethnically diverse companies are more effective. But what about age? **This is a very** important question, because for the first time ever, we have five generations in the workplace at the same time, unintentionally.

(TED_SHAPE_Gestão_Pessoas_26)

Alguns dos pacotes lexicais encontrados também revelam que o pioneirismo e a inovação relacionados a certos eventos e/ou descobertas são aspectos frequentemente enfatizados por palestrantes TED. O padrão *one of the first*, por exemplo, é empregado para destacar precedência, pioneirismo e auxiliar na contextualização temporal de um acontecimento junto ao público-alvo, conforme ilustra os exemplos abaixo.

- [129] In 1965, Dr. Geiger founded **one of the first** two community health centers in this country, in a brutally poor area in the Mississippi Delta.

(TED_STEM_Medicina_94)

- [130] When I was 12 years old, I was **one of the first** few people in Singapore who underwent cochlear implantation.

(TED_SHAPE_Música_28)

Palestrantes STEM também recorrem ao pacote lexical *this is the first* para destacar inovações e avanços científicos. No exemplo a seguir, esse padrão é empregado de modo a primeiramente ressaltar a ausência de precedentes do procedimento realizado para que, em seguida, o palestrante possa elaborar a explicação para o público-alvo.

- [131] What is this? **This is the first** transplant of naked DNA, where you take an entire DNA operating system out of one cell, insert it into a different cell, and have that cell boot up as a separate species.

(TED_STEM_Biotecnologia_09)

Palestrantes TED também empregam um amplo conjunto de pacotes lexicais para organizar o fluxo informacional e sinalizar pontos importantes a serem considerados em seus discursos. Para tanto, um padrão altamente recorrente é o encadeamento *one of the things + (that)*, empregado na identificação de aspectos específicos dentro de uma discussão mais ampla e complexa. Nos exemplos a seguir, nota-se que esse padrão auxilia palestrantes TED na apresentação de informações relevantes, segmentadas em partes menores. Tal estratégia simplifica a organização das ideias e argumentos apresentados no discurso, tornando-o acessível ao público-alvo.

[132] **One of the things that** characterizes extroverts is they need stimulation. And that stimulation can be achieved by finding things that are exciting: loud noises, parties, and social events.

(TED_STEM_Psicologia_23)

[133] And **one of the things that** we're doing at Starbucks is really having structured mentorship. But the mentorship looks a lot more like being a sponsor. And so, our executive leaders will be responsible for the development of our young, diverse talent and making sure that they're getting that exposure and those opportunities.

(TED_SHAPE_Gestão_Pessoas_11)

Por sua vez, o pacote lexical *one of the reasons* demonstrou ser um padrão ao qual palestrantes TED recorrem para identificar razões que reforcem seus argumentos (Exemplo 134) e/ou elucidem as causas responsáveis por algum evento ou situação (Exemplo 135).

[134] **One of the reasons** NEOWISE is so valuable is that it sees the sky in the thermal infrared. That means that instead of seeing the sunlight that asteroids reflect, NEOWISE sees the heat that they emit.

(TED_STEM_Astronomia_08)

[135] But there are tons of people working on this, from financial institutions to technology companies, start-ups, and universities. And **one of the reasons** is that it's not just an economic evolution. It's also an innovation in computer science.

(TED_SHAPE_Economia_14)

Os dados obtidos também indicaram que palestrantes TED recorrem a um conjunto considerável de expressões padronizadas para sinalizar a conclusão de suas explicações. No Exemplo 136, o uso do pacote lexical *and this is what* consolida o argumento do palestrante e estabelece conexões entre as situações apresentadas previamente, facilitando assim a compreensão do público-alvo acerca do raciocínio exposto.

[136] Actually, this may be the biggest story that we have ever covered. We have covered many wars. We have covered a lot of tragedies, a lot of problems, a lot of conflict zones, a lot of hot spots in the region, because we were centered at the middle of it. But this is a story – it is a great story; it is beautiful. It is not something that you only cover because you have to cover a great incident. You are witnessing change in history. You are witnessing the birth of a new era. **And this is what** the story's all about.

(TED_SHAPE_Ciências_Políticas_68)

Já os exemplos a seguir ilustram como diversos pacotes lexicais de identificação/foco auxiliam os palestrantes na síntese dos principais argumentos e informações expostos previamente em seus discursos.

[137] Introducing crops with high biodiversity, adaptability and nutritional value will play an important role in creating food security, seed sovereignty and sustainable production in communities and countries that are experiencing hunger and malnutrition. Having nutritious food should not be a luxury. There is a need to ensure that there is accessible and affordable nutritious food for everyone. **And this is a** step towards making it a reality.

(TED_STEM_Agronomia_02)

[138] It tells us that we and everything we see around us are basically one large bubble **and this is the** idea of inflation— one large bubble surrounded by something else.

(TED_STEM_Física_04)

[139] In our imagination, it has to be light. It has to be ephemeral. It has to be formless. And it has to be free. **So this is the** big idea. It's a giant airbag.

(TED_SHAPE_Arquitetura_48)

[140] If one can read and write that code of DNA, then they can be reconstructed in a distant location. **This is what we** mean by biological teleportation.

(TED_STEM_Genética_04)

Uma variação interessante foi a maior recorrência do pacote lexical *this is a picture* entre palestrantes das áreas STEM, o que, possivelmente, poderia estar relacionada à natureza disciplinar quantitativa das *hard sciences* e à preferência desses palestrantes em simplificar a explicação de conceitos e informações complexas por meio da representação visual de instrumentos técnicos, equações, imagens, gráficos e mapas (cf. Seção 4.2.4). O exemplo abaixo ilustra como esse pacote lexical identifica/sinaliza uma representação visual que complementa a explicação do palestrante e torna seu discurso mais tangível e de fácil compreensão para o público-alvo.

[141] This monstrosity over here is the CPU, the chip, if you like. Of course, it's this big. Completely mechanical. This whole machine is mechanical. **This is a picture** of a prototype for part of the CPU which is in the Science Museum.

(TED_STEM_Engenharia_36)

A identificação de certos grupos entre os participantes do evento também se revelou um aspecto recorrente nas TED Talks. Palestrantes de ambas as áreas recorrem, por exemplo, ao pacote lexical *those of you who* para identificar ou se referir a diferentes subgrupos presentes no público-alvo.

[142] And I want to put special emphasis on the words "rapidly" and "easily" because many of the existing screening methods are exactly the opposite. And **those of you who** have undergone colonoscopy as part of a screening program for colorectal cancer will know what I mean.
(TED_STEM_Medicina_51)

[143] Now, **those of you who** have been to Istanbul have probably seen Topkapi Palace, which was the residence of Ottoman sultans for more than 400 years.
(TED_SHAPE_Literatura_05)

Nesse contexto, observei que palestrantes SHAPE tendem a se incluírem como parte de determinados grupos por meio do pacote lexical *for all of us*. A maior recorrência desse padrão sugere uma preocupação no discurso das TED Talks SHAPE em identificar implicações de natureza coletiva, conforme pode ser observado no exemplo a seguir.

[144] And it's time, **for all of us**, to take the social and emotional development of our kids seriously.
(TED_SHAPE_Educação_32)

Por sua vez, palestrantes STEM recorrem ao pacote lexical *what was going on* para identificar, de maneira generalizada, diversos tipos de atividades durante a narração de algum evento para o público-alvo.

[145] Something in the sample was killing the bacteria that were supposed to cause the disease. To find out **what was going on** he did an ingenious experiment.
(TED_STEM_Biologia_03)

Após realizar uma discussão mais centrada no uso individual dos pacotes lexicais levantados, apresento, na seção seguinte, uma perspectiva macro sobre o que esses dados apontam a respeito do discurso das TED Talks.

4.3.4 Organizadores discursivos: o que os dados nos revelam?

Considerando as características situacionais de produção das TED Talks (*e.g.* produção em tempo real, tempo limitado, disponibilidade de recursos audiovisuais, entre outros), era de se esperar que palestrantes TED recorressem a diversas expressões convencionalizadas para ordenar e relacionar as ideias apresentadas ao longo de seus discursos diante do público-alvo. No que diz respeito à função discursiva de posicionamento, acredito que os pacotes lexicais identificados permitem os seguintes apontamentos acerca do discurso das TED Talks:

- **Pacotes lexicais para identificação e foco de entidades no discurso são recorrentes em ambos os *corpora* de estudo.**

Os dados sugerem que palestrantes TED empregam um amplo conjunto de padrões para caracterizar entidades, ações e eventos (*e.g. one of the most*), bem como para evidenciar informações de natureza abstrata e interpretativa (*e.g. one of the reasons*) que facilitem a compreensão do discurso por parte do público-alvo. Adicionalmente, observei um maior interesse, por parte de palestrantes STEM, em identificar entidades de natureza concreta/técnica (*e.g. and this is a carbon-fiber handrail*) quando comparados à contraparte SHAPE, sinalizando uma preocupação em expor o público-alvo a elementos que tornem o discurso mais tangível e facilitem a compreensão. Aproximadamente 10% das ocorrências de identificação/foco encontradas no *corpus* TED-STEM estiveram relacionadas à identificação de substantivos concretos/técnicos (*e.g. molecule, solar cell, microscopy*), que tendem a não possuir outros sinônimos para serem utilizados no discurso.

Em contrapartida, é possível argumentar que o discurso das TED Talks SHAPE demande de seu público-alvo maior atenção e tempo para interpretar e compreender a carga informacional do discurso, uma vez que a identificação e elaboração de entidades de natureza abstrata e processos (*e.g. as a result of; the problem is that*) mostrou-se mais padronizada entre palestrantes dessas áreas. A identificação de substantivos animados (*e.g. people*) e coletivos (*e.g. government*) também se mostrou mais acentuada no *corpus* TED-SHAPE, indicando uma preferência maior desses palestrantes em identificar aspectos relacionados a pessoas, eventos e instituições.

Por último, a alta recorrência dessa categoria de pacotes lexicais pode ser atribuída às

características particulares dos registros acadêmicos orais, como produção do discurso em tempo real e uso de recursos audiovisuais. Logo, essa recorrência não estaria necessariamente relacionada à natureza disciplinar das TED Talks, uma vez que a maioria dos pacotes lexicais de identificação e foco foram comuns em ambos os *corpora* de estudo.

- **Palestrantes STEM empregam um conjunto muito maior de pacotes lexicais para realizar elaboração e clarificação do discurso em comparação com a contraparte SHAPE.**

A alta recorrência de padrões tais como: (*and*) + *it turns out that* e *let me give you + (an example)*; indica que em ambas as áreas há uma preocupação em elaborar o discurso por meio da narração de acontecimentos e apresentação de exemplos que clarifiquem e facilitem a compreensão do público-alvo acerca das informações apresentadas. Entretanto, os dados demonstraram que palestrantes das áreas STEM recorrem a um conjunto muito maior de pacotes lexicais distintos para elaborarem e clarificarem seus discursos. Em especial, pacotes lexicais compostos pelo padrão *it turns out*, juntamente com o maior número de expressões que sinalizam perguntas retóricas (*e.g. so how do we*) revelam um aspecto interessante no discurso de palestrantes STEM de que o “ar de suspense” durante a revelação de descobertas seria uma estratégia discursiva de manutenção da atenção do público-alvo mais evidente em TED Talks das áreas STEM.

Outra variação possível de ser apontada foi o maior interesse de palestrantes STEM em explicar procedimentos para o público-alvo, o pacote lexical *what we do is*, por exemplo, evidencia que esse tipo de informação seria mais empregada por palestrantes das áreas STEM quando comparados à contraparte SHAPE.

- **A introdução de novos tópicos ao discurso, por meio de pacotes lexicais, ocorre de maneira relativamente padronizada em ambos os *corpora* de estudo.**

A recorrência relativamente alta de pacotes lexicais que realizam a introdução de tópicos pode ser atribuída ao fato de que esses palestrantes produzem seus discursos em tempo real. Assim, ao sinalizarem periodicamente novos tópicos, palestrantes TED se preocupam em fragmentar o fluxo informacional de seus discursos, proporcionando maior

tempo de assimilação para o público-alvo, o que facilita o acompanhamento e o processo de compreensão do discurso (Crawford Camiciottoli, 2004; Liu; Chen, H. 2020b).

Em ambos os *corpora* de estudo, a introdução de novos tópicos no discurso tende a ser realizada por meio de encadeamentos com expressões de desejos (*e.g. I want to talk + (about)*) e intenções (*e.g. (I'm) + going to talk about*). Adicionalmente, aproximadamente 15% das ocorrências de introdução de tópicos foram empregadas na abertura das palestras (*i.e.* 50 primeiras palavras da transcrição), evidenciando uma tendência em ambas as áreas de adiantar os tópicos que serão discutidos, ajustando assim as expectativas do público-alvo. O uso de dêiticos discursivos (*e.g. I am here today*) e marcadores discursivos (*e.g. now, so e well*) também são características recorrentes no momento em que palestrantes TED iniciam novos tópicos em seus discursos.

- **Palestrantes de ambas as áreas utilizam narrativas de histórias para projetarem suas identidades pessoais e profissionais frente ao público-alvo.**

Um aspecto notável foi a alta recorrência do item lexical “*story*” junto a diversos dos pacotes lexicais encontrados, indicando que compartilhamento de acontecimentos pessoais com os ouvintes é uma estratégia utilizada por palestrantes TED de ambas as áreas. Nesse sentido, os pacotes lexicais *let me tell you* e *to share with you*, por exemplo, mostram-se como expressões com alta recorrência em ambos os *corpora* de estudo.

O discurso de natureza narrativa das TED Talks pode ajudar a manter a atenção do público-alvo durante toda a palestra, pois, quando os palestrantes contam histórias envolventes, o público é mantido envolvido e interessado, isso aumenta a probabilidade de que a mensagem transmitida seja lembrada. Conforme é apontado por Caliendo (2014, p. 130):

A narrativa das TED Talks é, de fato, povoada por referências aleatórias, embora constantes, às histórias de contextualização através das quais os palestrantes compartilham suas experiências, despertam curiosidade sobre o tema e desenvolvem empatia com o público⁶⁰.

⁶⁰ No original: *TED narrative is in fact populated by random though constant references to background stories through which the speakers share their experiences, arouse curiosity regarding the topic and develop empathy with the audience.*

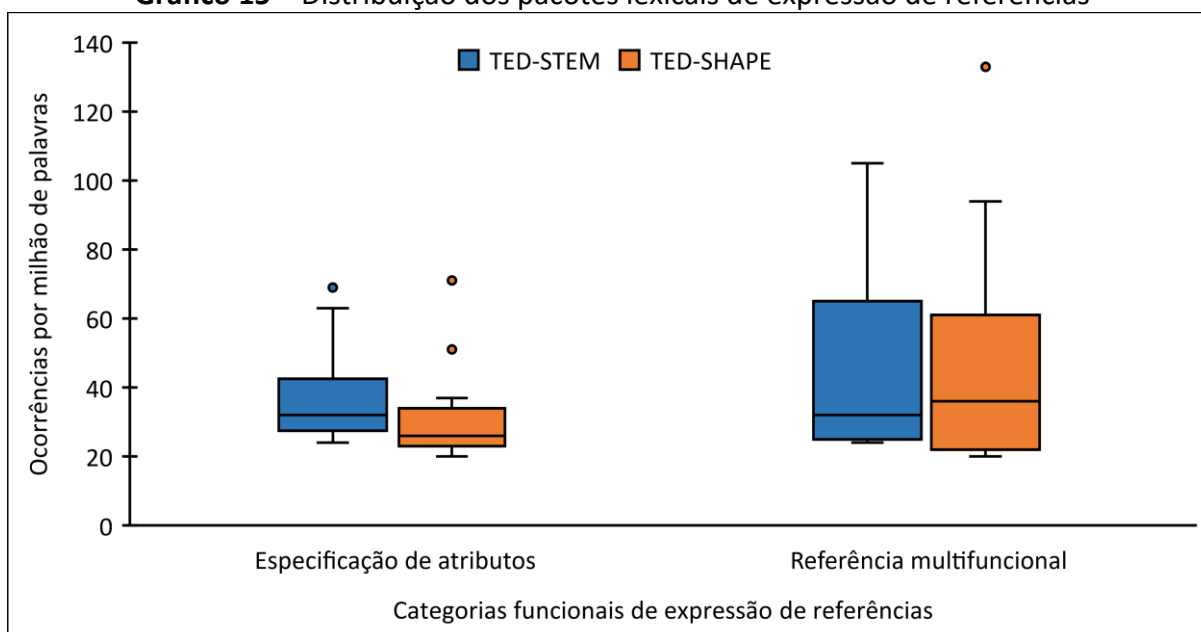
Ao compartilhar histórias pessoais, experiências e exemplos concretos, os palestrantes conseguem humanizar o conteúdo, tornando-o mais acessível e interessante (Nesi; Alsop, 2021). Isso facilita a compreensão das ideias apresentadas, mesmo quando se trata de temas complexos ou abstratos. Padrões recorrentes ao discurso das TED Talks como *(and) + it turns out that, let me tell you, all of a sudden* e *to share with you*, reforçam que a narração é uma técnica recorrente no estabelecimento da conexão emocional entre palestrante e público-alvo.

Na seção seguinte, examino a última categoria funcional dos pacotes lexicais levantados no presente estudo: a expressão de referência. Essa categoria aborda como os palestrantes TED recorrem a diversos padrões linguísticos para estabelecer conexões claras e coerentes entre as entidades mencionadas em seus discursos, permitindo uma melhor compreensão das ideias apresentadas para o público-alvo.

4.4 Pacotes lexicais de expressões de referências

Na presente seção, abordo em maiores detalhes o conjunto de pacotes lexicais relacionados à categoria de expressões referenciais encontrados nos *corpora* de estudo TED-STEM e TED-SHAPE. Os padrões descritos aqui têm como objetivo identificar ou atribuir características relevantes às entidades ou conceitos no discurso e serão discutidas a partir de duas subcategorias funcionais: especificação de atributos (Seção 4.4.1) e referências multifuncionais (Seção 4.4.2). Na Seção 4.4.3, concluo com as considerações finais sobre o que os dados encontrados revelaram a respeito da maneira expressões de referência são utilizadas na produção discursiva de palestrantes vinculados às áreas STEM e SHAPE.

Para facilitar essa discussão, o Gráfico 15 oferece uma comparação macro da proporção de ocorrências que expressam quantificações, atributos e referências de tempo, espaço e multifuncional encontradas nos *corpora* TED-STEM e TED-SHAPE.

Gráfico 15 – Distribuição dos pacotes lexicais de expressão de referências

Fonte: Elaborado pelo autor.

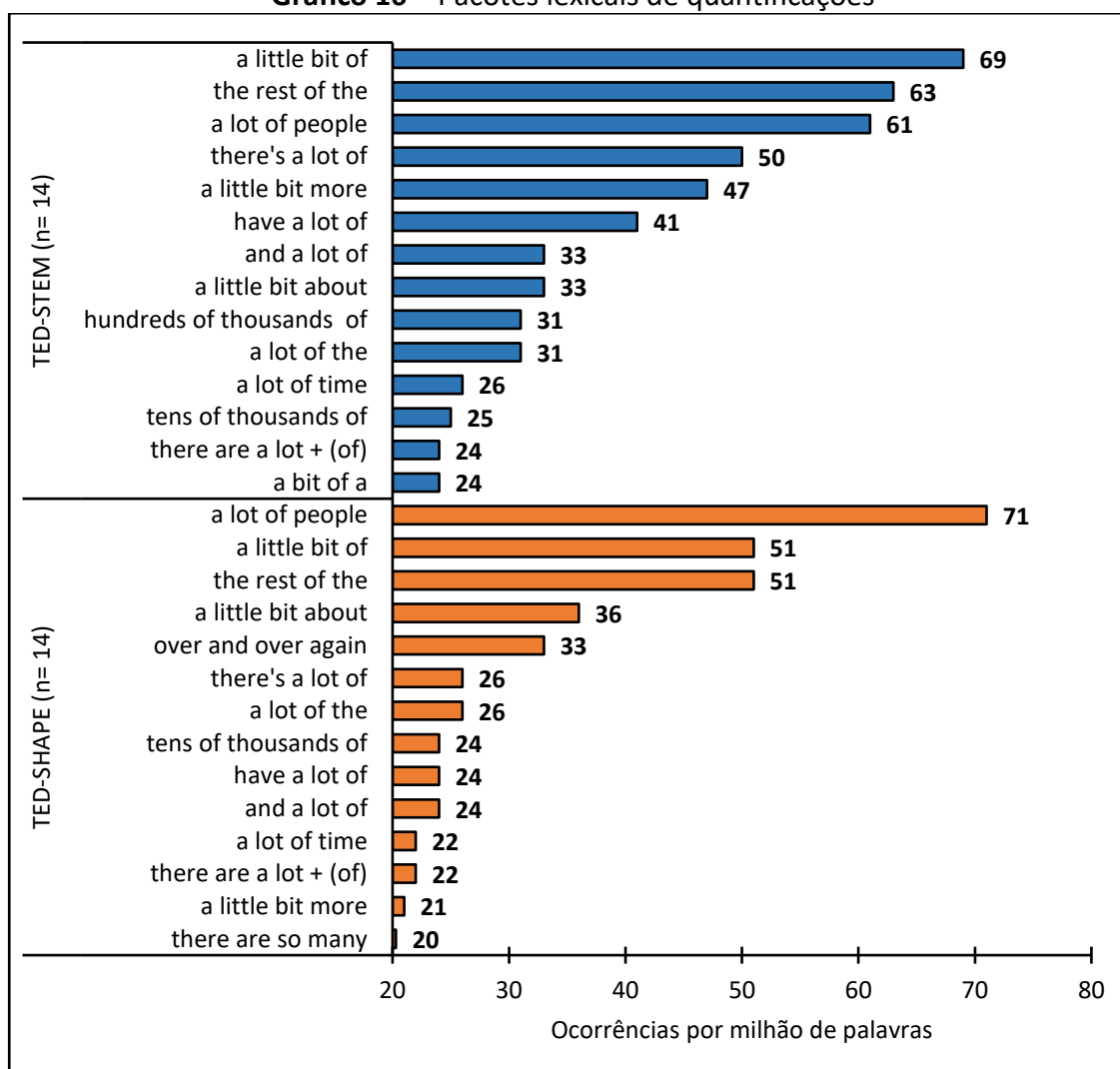
Os dados apresentados acima oferecem uma perspectiva interessante sobre a maneira como palestrantes TED de diferentes áreas do conhecimento recorrem a padrões linguísticos distintos para expressar atributos e referências em seus discursos. O conjunto de pacotes lexicais que especificam atributos a elementos do discurso (quantidades, tangíveis e intangíveis) apresentou uma maior dispersão interna nos valores dos dados no *corpus* TED-STEM, assim como uma tendência central de valores levemente mais elevada. Em comparação, essa categoria apresenta uma maior padronização interna no *corpus* TED-SHAPE, sugerindo que, apesar de haver alguns valores discrepantes, os pacotes lexicais desse conjunto tendem a ser empregados com frequência semelhante pelos palestrantes.

Já a categoria de pacotes lexicais multifuncionais apresentou uma dispersão interna no conjunto de dados semelhante em ambos os *corpora* de estudo (*i.e.* posição e tamanho da caixa). Contudo, a tendência central dos valores aparenta ser levemente mais elevada no *corpus* TED-SHAPE. Vale destacar ainda a acentuada discrepância na ocorrência de um dos pacotes lexicais em relação ao conjunto de referência multifuncional encontrados no *corpus* TED-SHAPE.

4.4.1 Especificação de atributos

Nesta subseção, examino os pacotes lexicais de especificações de atributos, isto é, os padrões mais recorrentes empregados por palestrantes TED ao expressar quantificações e descrições de atributos tangíveis e intangíveis referentes às entidades do discurso. Entre essas expressões, o conjunto considerável de pacotes lexicais que denotam quantificações em ambos os *corpora* de estudo (*cf.* Gráfico 16) evidencia a preocupação dos palestrantes em apresentar informações e dados mensuráveis, a fim de assegurar a transparência e confiabilidade do discurso perante o público-alvo.

Gráfico 16 – Pacotes lexicais de quantificações



Fonte: Elaborado pelo autor

Em relação à especificação de quantidades, vários pacotes lexicais identificados expressam quantificações genéricas do sintagma nominal seguinte, particularmente os

padrões compostos pelo quantificador *a lot of*. Os exemplos a seguir demonstram como os pacotes lexicais de quantificação auxiliam os palestrantes TED na simplificação das informações e durante a explicação de conceitos complexos, sem interromper o fluxo do discurso ou comprometer o engajamento com o público-alvo.

[146] And Plato, he had **a lot of the** same concerns that we did. He was worried about right and wrong. He wanted to know what is just. But he was worried that all we seem to be doing is trading opinions about this.

(TED_SHAPE_Filosofia_21)

[147] And flowers are a real bugger. They're really difficult for plants to produce. They take an enormous amount of energy **and a lot of** resources.

(TED_STEM_Botânica_04)

[148] I believe we'll see the biggest changes happen with changes to institutions like businesses, government, nonprofits, and society as a whole. **There's a lot of** evidence that this strategy is more effective.

(TED_STEM_Agronomia_15)

[149] And we all survived. It was cold, it was really tough and it was a great learning experience to discover that **there are a lot of** resources given by nature for free, and all that we need is our sensitivity to see them and our creativity.

(TED_SHAPE_Arquitetura_07)

[150] Now the DMCA had and is continuing to **have a lot of** complicated effects, but in this one domain, limiting sharing, it has mostly not worked.

(TED_SHAPE_Ciências_Políticas_19)

Ao empregar diferentes pacotes lexicais de quantificação para generalizar informações (*e.g. have a lot of complicated effects*), palestrantes TED conseguem ajustar rapidamente detalhes específicos e a complexidade das informações apresentadas ao seu público-alvo. Essa estratégia é essencial para que eles possam transmitir a maior quantidade de informações relevantes dentro do tempo disponível.

Uma padronização notável no que diz respeito à quantificação em TED Talks são os pacotes lexicais *a lot of people* e *a lot of time*; padrões consideravelmente recorrentes em ambos os *corpora* de estudo. Além de reforçar o discurso coloquial e acessível das TED Talks, esses padrões também demonstram como os palestrantes sugerem que o tópico em discussão seria relevante ou de interesse para um amplo grupo de pessoas. Os exemplos a seguir ilustram o uso do pacote lexical *a lot of people* como uma estratégia dos palestrantes para

destacar informações pertinentes que, possivelmente, seriam desconhecidas por uma parcela considerável do público geral (Exemplo 151) e esclarecer concepções comumente disseminadas pelo senso comum (Exemplo 152).

[151] **A lot of people** don't know that Alzheimer's actually has physical symptoms, as well as cognitive symptoms. You lose your sense of balance, you get muscle tremors, and that tends to lead people to being less and less mobile.
(TED_STEM_Medicina_02)

[152] So we had the goal, and now we needed to figure out what are the issues, what is broken. Before we went into schools, **a lot of people** told us that education quality is poor because either the teachers are lazy, they don't come into schools, or they're incapable, they actually don't know how to teach. Well, when we went inside schools, we found something completely different. On most days, most teachers were actually inside schools. And when you spoke with them, you realized they were perfectly capable of teaching elementary classes.
(TED_SHAPE_Educação_37)

Já o pacote lexical *a lot of time* consiste em um padrão recorrente que sinaliza a quantidade de tempo necessária para a execução de ação específica.

[153] Most of the footage I'm showing you is speeded up, because all of this takes **a lot of time**.
(TED_STEM_Biologia_70)

[154] Arab women of my generation have not been very visible in the public eye as they grew up. This explains, I think, to some extent, why you find so few women in politics in the Arab world. The upside of this, however, is that we have spent **a lot of time** developing a social skill behind the scenes, in coffee shops, in living rooms, on the phone, a social skill that is very important to success: networking.
(TED_SHAPE_Gestão_Pessoas_31)

O quantificador *a bit* também se mostrou um componente recorrente entre os pacotes lexicais mais utilizados por palestrantes TED. Entretanto, diferentemente das expressões compostas por *a lot of*, esses padrões tendem a expressar quantificações moderadas do sintagma nominal subsequente. Nos exemplos a seguir, observa-se que os palestrantes recorrem a esse padrão para atenuar a quantidade de informações apresentadas, evitando assim grandes generalizações sobre o tema em questão.

[155] This is not an animation, this is a video taken by the astronaut looking out of the window. It gives you **a bit of a** sense of scale of our two satellites.
(TED_STEM_Astronomia_46)

[156] There perhaps is the CO2 from the person sitting next-door to you. Maybe there is **a little bit of** oxygen from some algae on the beach not far from here. It also connects us in time.
(TED_STEM_Ecologia_10)

[157] I think I have come to understand **a little bit more** about happiness. There are three qualities I want to talk about.
(TED_SHAPE_Artes_11)

Já a recorrência do pacote lexical *a little bit about* é interessante, uma vez que seu uso não está associado à quantificação de entidades (físicas e/ou abstratas), mas sim à tentativa do palestrante em atenuar possíveis preocupações que o público-alvo possa ter quanto à densidade informacional do tópico em discussão (Biber; Conrad; Cortes, 2004; Nesi; Basturkmen, 2006). Tal propósito também se evidencia no discurso das TED Talks, especialmente quando empregada junto a expressões que sinalizam o início de novos tópicos. No exemplo a seguir, o uso do padrão *a little bit about* indica uma preocupação do palestrante em explicitar que não serão apresentadas quantidades excessivas de informações sobre o tema em discussão.

[158] I'm going to now just give you some of the data from that first approach. And in order to do that, I have to explain **a little bit about** how genes work.
(TED_STEM_Agronomia_06)

O Exemplo 159, por sua vez, ilustra como o palestrante recorre ao padrão *a little bit about* para estruturar seu discurso, sinalizando que ele abordará, de maneira introdutória, os tópicos de rentabilidade de poupança e idiomas, para, posteriormente, estabelecer a conexão entre esses temas a partir da discussão de aspectos mais específicos.

[159] Let me tell you **a little bit about** savings rates, **a little bit about** language, and then I'll draw that connection.
(TED_SHAPE_Economia_90)

No *corpus* TED-SHAPE, a recorrência do pacote lexical *over and over again* evidencia uma estratégia discursiva empregada por esses palestrantes para enfatizar a quantidade,

mesmo que imprecisa, de repetições de um determinado acontecimento, conforme pode ser observado no exemplo abaixo.

[160] And so the same kids that you thought were slow six weeks ago, you now would think are gifted. And we're seeing it **over and over again**.

(TED_SHAPE_Educação_35)

Outro aspecto interessante é que palestrantes STEM tendem a utilizar os pacotes lexicais *hundreds of thousands of* e *tens of thousands of* para enfatizar tempo decorrido, enquanto palestrantes SHAPE empregam o padrão *tens of thousands of* para destacar a quantidade de pessoas envolvidas. Tais diferenças podem ser observadas a seguir.

[161] There are wobbles in the Earth's orbit over **hundreds of thousands of** years that change the climate.

(TED_STEM_Ecologia_07)

[162] We find it in every population around the world at all time periods going back **tens of thousands of** years. And we even find it in neanderthals and animals.

(TED_STEM_Genética_2)

[163] So we now have the largest peer-grading pipeline ever devised, where **tens of thousands of** students are grading each other's work, and quite successfully, I have to say.

(TED_SHAPE_Educação_16)

Outro pacote lexical utilizado para expressar, de maneira vaga, quantificações de entidades, lugares e passagem de tempo foi o padrão *the rest of the*. Uma variação interessante em relação ao uso desse pacote lexical foi a maior proporção da combinação *the rest of + world* no *corpus* TED-SHAPE, cerca de 50% das ocorrências.

[164] When you look back on Iraq, those years when we were missing the story, were the years when the society was falling apart, when we were setting the conditions for what would become the rise of ISIS, the ISIS takeover of Mosul and terrorist violence that would spread beyond Iraq's borders to **the rest of the** world.

(TED_SHAPE_Jornalismo_14)

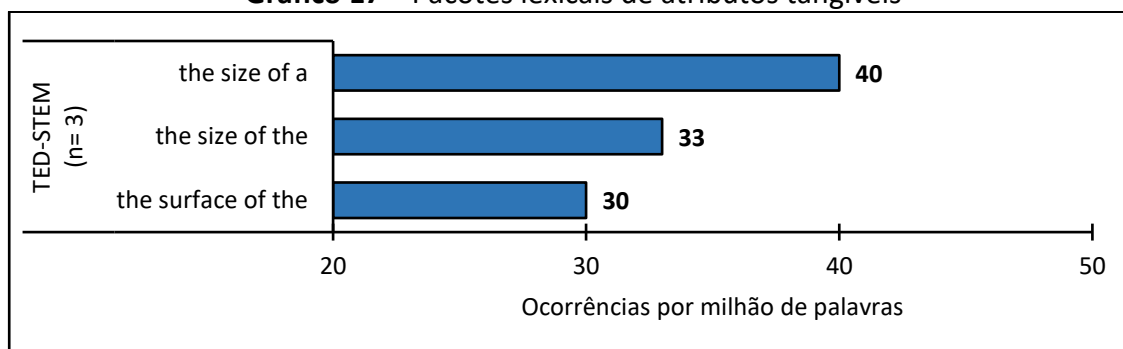
Em comparação, no *corpus* TED-STEM, o pacote lexical *the rest of the* tende a ser empregado junto a uma variedade maior de sintagmas nominais, conforme ilustra o exemplo a seguir.

[165] Agricultural labor productivity increased seven times, 1950 to 2000, whereas **the rest of the** economy increased about 2.5 times.

(TED_STEM_Biotecnologia_09)

A seguir, o Gráfico 17 apresenta as ocorrências dos pacotes lexicais empregados para expressar atributos tangíveis associados às entidades do discurso. Em especial, nota-se que os três pacotes lexicais associados a essa categoria funcional foram recorrentes apenas no *corpus* TED-STEM, evidenciando uma inclinação mais acentuada desses palestrantes em descrever as dimensões e os aspectos físicos dos objetos apresentados ao público-alvo.

Gráfico 17 – Pacotes lexicais de atributos tangíveis



*Nenhum pacote lexical do *corpus* TED-SHAPE foi associado a essa categoria funcional.

Fonte: Elaborado pelo autor

Em relação ao tipo de atributo indicado no decorrer do discurso, é interessante observar a presença dos pacotes lexicais *the size of a* e *the size of the*. A recorrência desses padrões revela uma tendência dos palestrantes STEM em estabelecer comparações de escala e proporção entre fenômenos e objetos em discussão, com o intuito de simplificar a compreensão. Os exemplos abaixo ilustram como esses padrões auxiliam na contextualização nas proporções de diferenças de escala e tamanho, utilizando referências tangíveis e compreensíveis, com as quais seu público-alvo possa ser relacionar facilmente.

[166] So this animal here has got a brain about **the size of a** grain of rice, and yet it can do things that you and I couldn't possibly entertain the idea of doing.

(TED_STEM_Biologia_79)

[167] It's so wide that it takes your plane one full minute to fly past it. That's **the size of the** asteroid that hit the Earth.

(TED_STEM_Astronomia_08)

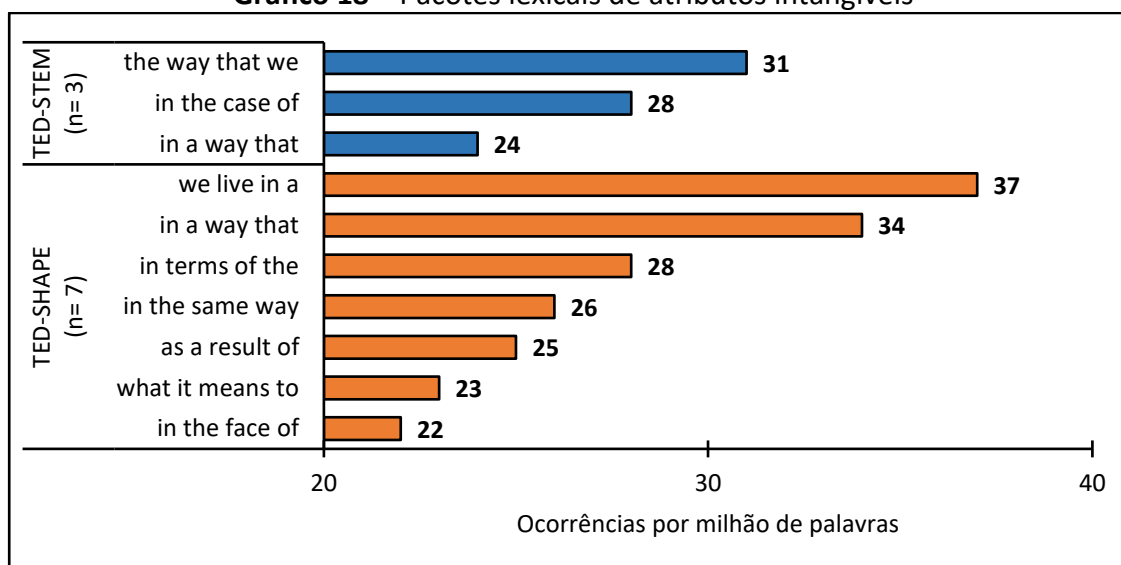
O pacote lexical *the surface of the*, por sua vez, indica que outro atributo comumente destacado no discurso de TED Talks STEM é o aspecto físico das superfícies de objetos, locais e materiais. Palestrantes STEM recorrem a esse padrão para descrever tais características de maneira clara e eficiente, facilitando a compreensão do público-alvo. No exemplo abaixo, nota-se que o padrão *the surface of the* ressalta uma característica física pertinente de uma entidade no discurso (*i.e.* a superfície de uma lua). Embora a mesma entidade esteja sendo visualmente representada, é interessante de se observar a preocupação do palestrante em fornecer informações adicionais que facilitem a compreensão do público-alvo.

[168] So what you're looking at is **the surface of the** moon Europa, which is a thick sheet of ice, probably a hundred kilometers thick.
(TED_STEM_Física_08)

Quanto à maior recorrência de pacotes lexicais relacionados à especificação de atributos no *corpus* TED-STEM, os dados encontrados alinham-se aos resultados apresentados por Durrant (2015); Hyland (2008a) e Simpson-Vlach (2006), que indicam um foco/interesse mais evidente no discurso das *hard sciences* em destacar descrição e proporções físicas de objetos e instrumentos.

Com relação à categoria de atributos intangíveis, o Gráfico 18 exhibe os padrões mais recorrentes entre palestrantes TED. Nele, é notável o maior conjunto e recorrência de padrões encontradas no *corpus* TED-SHAPE. O uso mais frequente de expressões que sinalizam abstrações auxilia os palestrantes das áreas SHAPE a apresentar perspectivas e interpretações mais flexíveis sobre tópicos difíceis de serem completamente delimitados e/ou mensurados.

Gráfico 18 – Pacotes lexicais de atributos intangíveis



Fonte: Elaborado pelo autor

Entre os pacotes lexicais recorrentes na sinalização de atributos intangíveis, o único padrão comum a ambos os *corpora* de estudo foi *in a way that*, que sinaliza aspectos qualitativos de entidades no discurso. Os exemplos a seguir ilustram como esse padrão auxilia os palestrantes a ressaltarem aspectos abstratos e/ou interpretativos em relação ao tópico em discussão, visando facilitar a compreensão do público-alvo.

[169] For these ideas to be feasible, we need to understand whether they can be applied at a vast scale **in a way that** is safe, economic, and socially acceptable.
(TED_STEM_Engenharia_74)

[170] Our species can make stuff, therefore we prospered **in a way that** no other species has.
(TED_SHAPE_Antropologia_14)

Outros padrões que enfatizam a discussão de conteúdos de natureza qualitativa nas TED Talks SHAPE são os pacotes lexicais *in terms of the* e *in the face of*. Estas expressões se destacaram como recorrentes entre palestrantes SHAPE no momento de sinalizar aspectos sob os quais o tópico em discussão está sendo abordado. No exemplo a seguir, o uso do pacote lexical *in terms of the* sinaliza a segmentação do foco da discussão em componentes menores, auxiliando o palestrante a apresentar informações em pequenos blocos e mais gerenciáveis para ao público-alvo.

[171] **In terms of the** political changes, they have introduced village elections. They have increased the security of proprietors. And they have increased the security with long-term land leases.

(TED_SHAPE_Economia_89)

Já no Exemplo 172, o uso do pacote lexical *in the face of* ajuda a realçar a gravidade ou importância da situação que segue, instigando o público-alvo a refletir sobre a complexidade das questões sociais apresentadas (*i.e.* pobreza e fome mundial).

[172] **In the face of** world poverty and hunger, making sure everyone in the world has a legal identity doesn't really seem important, but in reality, it is.

(TED_SHAPE_Desenvolvimento_Internacional_02)

Palestrantes SHAPE também aparentam contextualizar mais o público-alvo ao apresentar relações de causalidade em seus discursos, especialmente por meio do pacote lexical *as a result of*. No exemplo a seguir, seu uso identifica como certos eventos, ações ou condições podem causar ou influenciar diretamente outros elementos no discurso.

[173] Now, the elites basically have been liberated. They're very mobile. You cannot tax them. And basically, they don't fear the people. So, **as a result of** it, you have this very strange situation in which the elites basically got out of the control of the voters.

(TED_SHAPE_Ciências_Políticas_38)

Novamente, o passo retórico “*apresentação de cenário*” (Chang; Huang, 2015) é evidenciado pelo pacote lexical *we live in a*, que se apresenta como um padrão recorrente entre palestrantes da área SHAPE. Um aspecto interessante em relação à utilização deste pacote lexical é que nenhuma de suas ocorrências foi empregada como referência ao conceito de habitação (*e.g. house*), mas sim para uma contextualização abstrata e generalista do mundo ou sociedade na qual tanto os palestrantes quanto seus públicos-alvo estão inseridos, conforme pode ser observado a seguir.

[174] It's unforgivable that **we live in a** society where it's normal or expected for teenagers to grow up hating themselves.

(TED_SHAPE_Estudos_Gênero_04)

Já a delimitação do contexto e escopo das informações a serem apresentadas no discurso, no *corpus* TED-STEM, são identificadas junto ao pacote lexical *in the case of*. No exemplo abaixo, pode-se observar seu uso na identificação do tópico específico (*i.e.* opioides), antes de introduzir e debater como ele se inter-relaciona com a discussão de um tema mais abrangente.

[175] It took me virtually no time at all to discover that many of these facilities are geared towards those battling long-term substance use disorder. **In the case of** opioids, this often involves precisely not weaning the patient off the medication, but transitioning them onto the safer, longer-acting opioids: methadone or buprenorphine for maintenance treatment.

(TED_STEM_Medicina_117)

A categoria de atributos intangíveis também evidenciou que a utilização de comparações de natureza qualitativa entre as entidades do discurso foi mais acentuada entre os palestrantes das áreas SHAPE, especialmente através do pacote lexical *in the same way*, como pode ser observado no exemplo abaixo.

[176] The implications are huge. **In the same way** the internet blew open the doors to an age of information available to everyone, the blockchain will revolutionize trust on a global scale.

(TED_SHAPE_Economia_64)

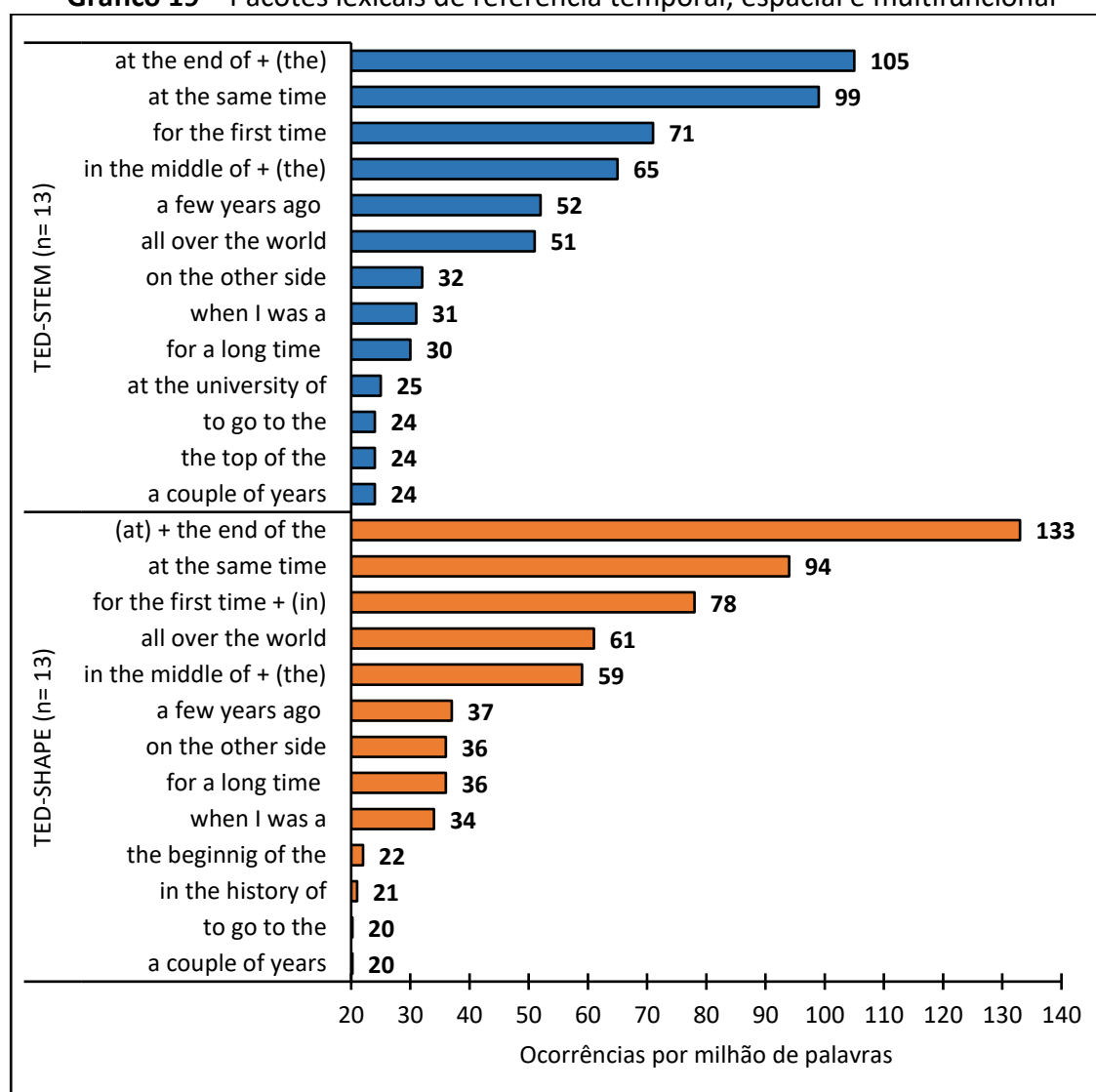
Concluída a análise dos pacotes lexicais referentes à categoria de atributos, a qual contemplou a quantificação, bem como os atributos tangíveis e intangíveis, é oportuno transitar, na seção seguinte, para a exploração dos pacotes lexicais de referência temporal, espacial e multifuncional. Essa categoria é fundamental para a compreensão de como os palestrantes TED orientam seus discursos no tempo e no espaço, além de como empregam estratégias linguísticas multifuncionais para enriquecer e aprimorar sua interação com o público.

4.4.2 Referências de tempo, espaço e multifuncional

Nesta seção, analiso os pacotes lexicais de referência temporal, espacial e multifuncional. Estes elementos linguísticos desempenham papéis críticos na estruturação

dos discursos, orientando o público através das narrativas e permitindo uma navegação mais precisa no espaço discursivo e temporal apresentado pelos palestrantes. Logo, o emprego eficiente dessas expressões pode facilitar a compreensão do público, construir conexões significativas entre ideias e situações, e gerar uma atmosfera de envolvimento e interesse. O Gráfico 19 ilustra o conjunto de pacotes lexicais utilizados para sinalizar referência temporal, espacial e multifuncional em TED Talks das áreas STEM e SHAPE. Ao explorar estes padrões linguísticos, busco entender como os palestrantes manipulam e controlam o espaço discursivo para otimizar a entrega de suas mensagens e maximizar o impacto de suas apresentações.

Gráfico 19 – Pacotes lexicais de referência temporal, espacial e multifuncional



Fonte: Elaborado pelo autor

Com relação às expressões de referência temporal, o pacote lexical mais recorrente em ambos os *corpora* foi *at the same time*. Palestrantes TED frequentemente recorrem a este padrão com a finalidade de indicar relações de simultaneidade entre atividades e/ou eventos apresentados em seus discursos, conforme demonstram os exemplos abaixo.

[177] This flip was everything to global agricultural markets. Luckily now, South America was starting to boom **at the same time** as China's rise, and so therefore, supply and demand were still somewhat balanced.

(TED_STEM_Agronomia_18)

[178] Fake news is not only bad for journalism. It's a threat for democracy and society. Four years ago, unmarked soldiers entered the Crimean Peninsula, and, **at the same time**, Russian media was going crazy with fake news about Ukraine.

(TED_SHAPE_Jornalismo_20)

Observei ainda que *at the same time* também é empregado na sinalização de contradições lógicas no discurso, conforme ilustra o Exemplo 179, o palestrante se vale do padrão *at the same time* com o propósito de ressaltar um contraste entre a crescente quantidade de pacientes na fila e o número de transplantes efetivados.

[179] In the last decade the number of patients waiting for a transplant has doubled. While, **at the same time**, the actual number of transplants has remained almost entirely flat. That really has to do with our aging population.

(TED_STEM_Medicina_04)

Os dados também revelam que palestrantes TED recorrem a certos padrões para realizarem diferentes formas de contextualização histórica do tema. Os pacotes lexicais *for the first time* e *in the history of*, por exemplo, mostraram-se recorrentes quando palestrantes TED buscam enfatizar avanços científicos, tecnológicos e sociais frente ao público-alvo.

[180] So what we've been doing is using the laser-guided spectroscopy from the CAO to map **for the first time** the biodiversity of the Amazon rainforest. Here you see actual data showing different species in different colors.

(TED_STEM_Ecologia_09)

[181] It's the first time **in the history of** Ghana that a woman has been elected head of Student Government at any university.

(TED_SHAPE_Educação_47)

A recorrência desses pacotes lexicais corrobora uma característica apontada por Caliendo (2014), que é a tentativa dos palestrantes TED de estabelecer proximidade com seu público-alvo ao apresentarem conceitos como sendo novos, inovadores e sem precedentes.

Já o padrão *the beginning of the* sugere uma tendência maior dos palestrantes SHAPE em enfatizarem detalhes históricos acerca do tópico exposto. Possivelmente, devido à natureza reiterativa das investigações no campo das *soft sciences*, visto que o conhecimento em disciplinas como artes, economia, história e sociologia tende a ser construído de maneira dialógica com eventos passados (Becher; Trowler, 2001), conforme ilustra o exemplo a seguir.

[182] So I was born in 1966, at **the beginning of the** Biafran-Nigerian Civil War, and the war ended after three years.
(TED_SHAPE_Literatura_11)

É possível inferir, portanto, que palestrantes das áreas STEM demonstram menor preocupação em contextualizar o público-alvo acerca da historicidade dos eventos, em comparação à contraparte SHAPE. Tal particularidade explicaria, ao menos parcialmente, a menor frequência relativa dos pacotes lexicais *the beginning of the* e *in the history of*, no *corpus* TED-STEM.

Já o pacote lexical *for a long time* revela um foco interessante no discurso das TED Talks, que é o destaque de informações que, embora não sejam novidades para os especialistas da área, ainda seriam de relevância para o público leigo (Exemplo 183), ou representam questões ainda não totalmente compreendidas pelo próprio palestrante, como ilustra o Exemplo 184 abaixo.

[183] So how on earth would one of these guys end up in the High Arctic? Well, scientists have known **for a long time** turns out, even before Natalia's discovery, that camels are actually originally American. They started here.
(TED_STEM_Biologia_69)

[184] You know, we had a long 50-year period of dealing with our race problems and legal discrimination, and that was our top priority **for a long time** and it still is important.
(TED_SHAPE_Sociologia_07)

Outras expressões frequentemente empregadas pelos palestrantes TED para estabelecer proximidade com o público-alvo incluem os pacotes lexicais *a couple of years*,

a few years ago e *when I was a*. Tais padrões auxiliam os palestrantes na contextualização de situações pessoais que vivenciaram em momentos específicos de suas vidas.

[185] And so **a couple of years** back, my cofounder and I were teaching university students just like these, only the students we were teaching looked a little bit more like this here.

(TED_STEM_Design_07)

[186] On a beautiful day, just **a few years ago**, my wife and I entered a hospital near our home in Oakland, California for the birth of our first daughter, Maya.

(TED_SHAPE_Arquitetura_36)

[187] **When I was a** student, if you had a degree, you had a job. If you didn't have a job, it's because you didn't want one.

(TED_SHAPE_Educação_38)

No que concerne às referências espaciais, o padrão mais recorrente entre palestrantes TED foi *all over the world*. Conforme evidenciado pelos exemplos a seguir, esse pacote lexical auxilia o palestrante a aludir de maneira generalizada à abrangência ou extensão de um acontecimento e, conseqüentemente, ressaltar a sua relevância frente ao público-alvo.

[188] So building a machine like the LHC requires countries from **all over the world** to pool their resources.

(TED_STEM_Física_21)

[189] In Tijuana, we've met refugees from over 50 countries, speaking 14 different languages. We meet LGBT migrants from **all over the world** who have never been in a country in which they feel safe. We meet women from **all over the world** whose own governments refuse to protect them from brutal domestic violence or repressive social norms.

(TED_SHAPE_Direito_17)

O pacote lexical *on the other side*, por sua vez, foi empregado pelos palestrantes de ambas as áreas para estabelecer uma referência espacial relativa entre os elementos apresentados em seu discurso, conforme é possível observar nos exemplos abaixo.

[190] For example, the CDC in Atlanta, Georgia could send flu vaccine instructions to a DBC **on the other side** of the world, where the flu vaccine is manufactured right on the front lines.

(TED_STEM_Genética_04)

- [191] The temperature difference on a Moon base could be anything up to 200 degrees. On one side of the base, it could be 100 degrees Celsius and, **on the other side**, it could be minus 100 degrees. We need to protect ourselves from that.
(TED_SHAPE_Arquitetura_93)

Uma variação entre os *corpora* de estudo, no entanto, foi a maior recorrência do pacote lexical *the top of the*, por parte de palestrantes STEM, para indicar referências espaciais em seus discursos.

- [192] We were flying at 100 meters above **the top of the** canopy to measure this molecule – incredibly dangerous stuff.
(TED_STEM_Química_13)

No âmbito das referências multifuncionais, essa categoria corresponde aos pacotes lexicais que desempenham funções discursivas de referências temporais, especiais ou dêiticas, a depender do contexto discursivo. Em ambos os *corpora* de estudo, os padrões mais recorrentes foram os encadeamentos *at the end of + (the)* e *in the middle of + (the)*. Algumas dessas ocorrências são apresentadas nos exemplos a seguir.

- [193] There are seven questions. Each question, there's a possible answer between zero and a hundred, and **at the end of the** quiz, you get an overall score between zero and a hundred.
(TED_STEM_Matemática_02)
- [194] Some of the largest vessels ever constructed on planet Earth were built right **at the end of** my street.
(TED_SHAPE_Música_21)
- [195] The Large Magellanic Cloud is right **in the middle of** that image, but despite its name, it's really small.
(TED_STEM_Astronomia_14)
- [196] It's what they believed about America that got them to travel in a bus for eight hours to stand in the sun in Washington **in the middle of** August.
(TED_SHAPE_Gestão_Pessoas_35)

Com a análise dos pacotes lexicais que expressam referências diversas como quantidade, atributos, tempo e espaço concluída, passo agora a discussão macro sobre o que esses dados revelam acerca da natureza do discurso disciplinar nas TED Talks. A partir dos pacotes lexicais identificados, busco apresentar um panorama das principais estratégias

discursivas empregadas pelos palestrantes das áreas STEM e SHAPE, bem como as peculiaridades que caracterizam o discurso das TED Talks.

4.4.3 Expressões de referência: o que os dados nos dados revelam?

A partir da discussão dos dados apresentados, é possível inferir algumas particularidades na utilização de pacotes lexicais de expressões referenciais entre TED Talks das áreas STEM e SHAPE.

- **Palestrantes de ambas as áreas empregam em seus discursos um conjunto amplo de expressões para especificação de atributos, em especial, quantificação.**

Em ambos os *corpora* de estudo foram encontrados um amplo conjunto de expressões padronizadas que auxiliam palestrantes TED a expressar quantificações gerais em seus discursos (*e.g. there's a lot of*) e gerir o fluxo de informação apresentadas ao público-alvo (*e.g. a little bit about*). No entanto, é interessante observar que mesmo entre palestrantes das disciplinas STEM, expressões que denotam valores matemáticos (*e.g. the rate at which*⁶¹) são raramente empregadas. Isso possivelmente se deve ao intuito dos palestrantes em simplificar o discurso para o público-alvo.

Já as expressões que sinalizam atributos tangíveis (*e.g. the size of a*), os dados demonstram claramente que esse tipo de referência ocorre de modo mais convencionalizado entre palestrantes STEM. Em contraste, palestrantes SHAPE exibiram uma maior recorrência de padrões que expressam atributos abstratos (*e.g. in terms of the*), para descreverem entidades em seus discursos.

- **Grande parte dos pacotes lexicais de referências ao tempo/espço e multifuncionais extraídos ocorrem em ambos os *corpora* de estudo.**

As duas áreas se utilizam de um conjunto bastante semelhante de pacotes lexicais para efetuar contextualizações generalizadas de tempo, espaço e multifuncional (*e.g. a few years*

⁶¹ A título de exemplificação, a frequência absoluta da expressão *the rate at which* foi de apenas 11 ocorrências no *corpus* TED-STEM e sua dispersão foi de 8 textos distintos (1% do *corpus*). No *corpus* TED-SHAPE, a mesma expressão ocorreu apenas uma vez.

ago, all over the world e *(at) + the end of the*). Possivelmente, isso ocorre em virtude do fato de que a contextualização histórica do tema e o compartilhamento de experiências pessoais são estratégias comumente empregadas por palestrantes TED para criar uma conexão com o público-alvo (Caliendo, 2014; Scotto di Carlo, 2014b).

A recorrência de padrões que auxiliam os palestrantes a ordenarem de modo cronológico os eventos de seus discursos (*e.g. for the first time*), sinalizarem abrangência de eventos (*e.g. all over the world*) e narrativas pessoais (*e.g. when I was a*) é de suma importância, especialmente para a realização do passo retórico “*apresentação de cenário*” (Chang; Huang, 2015), no qual os palestrantes TED se esforçam para contextualizar o público-alvo acerca da situação que será apresentada em seguida.

- **Pacotes lexicais que expressam imprecisão no discurso não exibiram alta padronização em nenhum dos *corpora* de estudo.**

De forma interessante, pacotes lexicais de imprecisão (*e.g. or something like that*), relativamente comuns em registros orais (Biber; Conrad; Cortes, 2004), não foram identificados em nenhum dos *corpora* de estudo. Embora seja difícil determinar a razão para tal, é possível especular que os palestrantes de TED Talks evitam o uso de expressões que possam indicar inexatidão ou gerar interpretações ambíguas do discurso.

Concluindo a análise dos pacotes lexicais, é evidente que a variação disciplinar no discurso das TED Talks é marcada por diferenças interessantes no emprego de expressões de posicionamento, organização discursiva e expressões referenciais entre as áreas STEM e SHAPE. Os padrões apresentados no decorrer deste capítulo ilustram a diversidade e a especificidade da comunicação em cada campo. Ao mesmo tempo, nota-se também uma similaridade considerável no uso de determinados pacotes lexicais, sugerindo estratégias comuns de engajamento com o público. Avançando para as considerações finais deste estudo, faço uma síntese das descobertas, delineando as implicações para a comunicação interdisciplinar e sugerindo direções para futuras pesquisas.

5 CONSIDERAÇÕES FINAIS E ENCAMINHAMENTOS

Neste capítulo final, busco, inicialmente, revisitar o caminho teórico-metodológico percorrido ao longo deste trabalho, reiterar a discussão em torno dos principais dados encontrados e limitações enfrentadas para, por fim, concluir com os encaminhamentos mais relevantes proporcionados por este estudo.

A presente tese buscou investigar os pacotes lexicais de quatro palavras mais recorrentes entre TED Talks das áreas STEM e SHAPE. A necessidade de investigar esse gênero sob a perspectiva disciplinar é justificada pela sua sub-representação na literatura acadêmica (Caliendo, 2014) e pelo potencial inerente das TED Talks em estreitar a distância entre conhecimentos de natureza científica e um público-leigo através de uma narrativa acessível (Caliendo, 2014; Miranda, A. 2016; Miranda, J. 2018; Scotto di Carlo, 2014a; 2014b).

Deste modo, iniciei a presente tese delineando as principais características do gênero TED Talks a partir da perspectiva do Quadro de Análise Situacional (Biber; Conrad, 2009), dos movimentos/passos retóricos típicos em TED Talks (Chang; Huang, 2015; Liu; Chen, H. 2020b). Em segmento, discuti a concepção de *corpora* eletrônicos como objetos de investigação linguística (Berber Sardinha, 2004; McEnery; Hardie, 2012; Sinclair, 2005; Teubert, 2005; Viana; Zyngier; Barnbrook, 2011, entre outros) e elucidei como esta pesquisa se alinha às abordagens dirigidas (*corpus-driven*) e baseadas (*corpus-based*) em *corpus* (Tognini-Bonelli, 2001). A respeito desta dicotomia, acredito que uma análise de *corpora* que desassocie completamente “teoria” e “dados empíricos” seja um desafio extremamente difícil de se realizar. Portanto, a postura adotada para conduzir esta pesquisa está alinhada com as orientações dos estudos baseados em *corpus* (*corpus-based studies*), conforme sugerido por Viana (2012).

Nesta linha, os pacotes lexicais foram primeiramente identificados por meio da exploração dos *corpora* de estudo (dados → teoria) e, posteriormente, analisados com o apoio de explicações descritivas e funcionais já consolidadas na literatura acadêmica (teoria → dados). Em particular, os estudos conduzidos por Biber (2006a), Biber e Barbieri (2007), Biber, Conrad e Cortes (2004), Cortes (2015); Nesi e Basturkmen (2006) e Hyland (2008a; 2011a) constituíram a base teórica para a análise e discussão dos pacotes lexicais encontrados.

A escolha pela investigação de pacotes lexicais é justificada pela sua natureza como estruturas linguísticas recorrentes que estão associadas a várias funções comunicativas

fundamentais (Biber, 2006a). No contexto das TED Talks, padrões como *I want to talk* ou *to be able to* representam exemplos de estruturas que são rapidamente formuladas e processadas pelos falantes, contribuindo para a fluidez e eficácia na produção do discurso. Através da combinação das abordagens dirigidas e baseadas por *corpus*, foi possível destacar algumas particularidades discursivas associadas à maneira como palestrantes STEM e SHAPE expressam seus posicionamentos, organizam seus discursos e sinalizam referências de múltiplas naturezas durante a construção de seus textos frente ao público-alvo.

Em relação ao material e métodos utilizados para a elaboração desta tese, compilei dois *corpora* de estudo a partir de 1.519 transcrições de TED Talks. Primeiramente, os metadados de cada palestra foram organizados em uma planilha de Excel® para evitar, sobretudo, redundâncias na coleta dos textos e facilitar a busca e localização de metainformações ao longo do estudo (*cf.* Figura 10). A etapa subsequente envolveu o uso da ferramenta *WebBootCat* do Sketch Engine® para a extração automatizada das transcrições a partir das URLs do site oficial TED. No final desse processo, as transcrições obtidas foram organizadas em dois *corpora* de estudo: *corpus* TED-STEM, composto por 796 palestras associadas a 16 tópicos disciplinares, totalizando 1.6 milhão de palavras, e o *corpus* TED-SHAPE, composto por 723 palestras, relacionadas a 18 tópicos disciplinares e totalizando 1.4 milhão de palavras. Ambos os *corpora* de estudos foram apresentados em maior detalhe na Seção 3.3.

Para explorar os *corpora* de estudo, utilizei o *software* AntConc® (Anthony, 2020a). Essa ferramenta permitiu a extração semiautomática de sequências fixas de quatro palavras com frequência relativa superior a 20 ocorrências por milhão de palavras, distribuídas em pelo menos 4% dos textos compilados em cada *corpus* de estudo. Embora “a frequência seja a qualidade máxima dos pacotes lexicais⁶²” (Cortes, 2015, p. 204), as decisões sobre o corte de frequência relativa mínima tendem a ser arbitrárias na literatura acadêmica (Biber, 2006a, 2009; Cortes, 2004, 2013, 2015; Hyland, 2008a, 2012). Por outro lado, a decisão por uma dispersão relativa de 4% permitiu a identificação de padrões que, além de recorrentes, também são empregados por um amplo conjunto de palestrantes de um mesmo registro.

Em resposta à questão de pesquisa: “*Como o conjunto de pacotes lexicais mais recorrentes entre esses domínios disciplinares varia em termos de frequência, dispersão e*

⁶² No original: “*Frequency is the ultimate quality of lexical bundles*”.

estrutura gramatical?”, a análise dos dados revelou inicialmente um conjunto de 158 pacotes lexicais relacionados às TED Talks no domínio disciplinar STEM, e de 128 na contraparte SHAPE. Para mitigar a possibilidade de inflação dos dados, examinei o conjunto de pacotes lexicais extraídos com o intuito de refinar a frequência relativa e dispersão de padrões em contexto de sobreposição completa, subsunção completa e parcial (Chen, Y. 2009; Chen, Y.; Baker, 2010; Grondona, 2015). Nos *corpora* TED-STEM e TED-SHAPE, identifiquei, respectivamente, 38 e 28 padrões suscetíveis de ter suas ocorrências refinadas (*e.g. I’m going to talk | going to talk about*), resultando 19 e 14 em encadeamentos lexicais maiores (*e.g. (I’m) + going to talk about*). Adicionalmente, foram excluídos os pacotes lexicais relacionados ao tópico, isto é, sequências que são dependentes do contexto ou que incorporam nomes próprios (*e.g. in the United States*). Ao final desse processo, os *corpora* de estudo TED-STEM e TED-SHAPE apresentaram, respectivamente, 138 e 114 pacotes lexicais distintos para análise. Dentre estes conjuntos, 79 padrões foram recorrentes em ambos os *corpora* de estudo, indicando que a realização de certas convenções discursivas (*i.e.* desejos, habilidades, referência multifuncional) são realizadas de forma semelhante pelos palestrantes de TED Talks em ambas as áreas. Contudo, o maior conjunto de pacotes lexicais encontrados em TED Talks associadas ao domínio STEM sugere uma padronização relativamente maior em relação à forma como esses palestrantes organizam seus discursos e apresentam informações para seus público-alvo.

Apesar de os pacotes lexicais serem estruturas linguísticas incompletas, eles apresentam fortes correlações gramaticais e a sua classificação com base nas suas composições internas representa um passo crucial para a descrição e identificação das funções discursivas que realizam (Biber, 2006a; Cortes, 2015). Assim, utilizei a taxonomia estrutural proposta por Biber, Conrad e Cortes (2004), na qual os pacotes lexicais podem ser organizados entre aqueles de natureza verbal (Tipo 1 e 2) e aqueles de natureza nominal (Tipo 3). Os dados coletados sugerem que as TED Talks constituem, de fato, um gênero textual híbrido que exhibe um alto índice de uso de pacotes lexicais associados tanto ao discurso acadêmico oral (*i.e. if you look at*) quanto escrito (*i.e. on the other hand*). Contudo, é pertinente salientar que, embora pacotes lexicais de natureza nominal sejam prevalentes em ambos os *corpora* de estudo, as TED Talks associadas ao domínio STEM empregam uma proporção um pouco mais acentuada de pacotes lexicais de natureza verbal em comparação à contraparte SHAPE.

No que se refere a seguinte pergunta de pesquisa: *“Como as funções linguísticas de posicionamento, organização discursiva e expressão referencial podem elucidar diferenças na estruturação do discurso entre palestrantes TED das áreas STEM e SHAPE?”*, o Quadro 9 a seguir apresenta as principais generalizações acerca do discurso de TED Talks associadas aos domínios STEM e SHAPE.

Quadro 9 – Principais particularidades do discurso disciplinar das TED Talks evidenciadas por pacotes lexicais

TED-STEM	Ambos	TED-SHAPE
– Maior conjunto de pacotes lexicais distintos – Maior padronização do verbo <i>show</i> – Enfoque na narração de descobertas científica – Discurso permeado por um “ar de suspense” – Estímulo ao “espírito científico” no público-alvo – Maior direcionamento da atenção e foco do público-alvo – Predominância de padrões que expressam competência técnica (<i>e.g.</i> habilidades) – Prudência ao comentar sobre descobertas recentes – Maior conjunto de padrões que realizam elaboração/clarificação no discurso – Foco na explicação de procedimentos técnicos ou científicos – Identificação de ferramentas ou instrumentos técnicos – Uso de expressões que descrevem atributos tangíveis	– Fomento de credibilidade e autoridade do palestrante – Alta recorrência de padrões que sinalizam intenções/predições – Alta recorrência de padrões que introduzem novos tópicos ao discurso – Alta recorrência de padrões que realizam identificação/foco de entidade no discurso – Alta recorrência de padrões que sinalizam referência temporais, espaciais e dêiticas – Indicação clara dos tópicos e estrutura do discurso (<i>signposting</i>) – Compartilhamento de vivências e eventos pessoais – Emprego de exemplos ilustrativos – Descrição detalhada de entidades, ações e eventos – Foco em descobertas, ideias e conceitos inovadores – Emprego de expressões quantificadoras – Gestão do fluxo de informações apresentadas – Baixa padronização de expressões que sinalizam imprecisão	– Credibilidade evidenciada na exposição dos métodos e procedimentos adotados – Utilização relativamente mais acentuada de pacotes lexicais anguladores – Ênfase em ações de responsabilidade coletiva – Foco no pioneirismo de eventos – Maior conjunto de padrões que expressam atributos intangíveis – Comparação de entidades em termos qualitativos – Contextualização histórica e temporal dos eventos – Destaque para a ocorrência repetida de eventos

Fonte: Elaborado pelo autor

Assim sendo, os palestrantes das TED Talks empregam diferentes estratégias linguísticas para estabelecer sua autoridade e credibilidade, adaptando-se às particularidades e objetivos específicos de seus campos de atuação e estabelecendo uma relação mais próxima e dialógica com seu público-alvo.

Por último, é pertinente abordar as limitações encontradas ao longo da execução deste trabalho, bem como os possíveis encaminhamentos que a presente tese abre para estudos futuros.

No que concerne às diferentes escolhas teórico-metodológicas viáveis, o processo de compilação dos *corpora* de estudo se mostrou uma tarefa relativamente simples e rápida, em virtude da facilidade de acesso e disponibilidade das transcrições. A ferramenta *WebBootCat* do Sketch Engine® também se mostrou de grande valia no processo de compilação e extração semiautomatizada dos textos no site TED. Todavia, acredito que o inventário de disciplinas selecionadas e a associação das TED Talks a uma disciplina específica seja um dos aspectos com maior margem para divergências. Certamente, a investigação desse gênero sob o prisma disciplinar é de extrema relevância para diversas áreas dos estudos linguísticos, contudo, tais decisões não estão imunes a fragilidades, dentre as quais, as principais seriam:

- O grande número de etiquetas disciplinares/tópicos redundantes, em muitos casos;
- A necessidade de decidir/justificar, a partir do conjunto de etiquetas disponíveis no site TED, o que seria ou não considerado como uma disciplina;
- A tarefa de categorizar/justificar as disciplinas escolhidas entre os grupos disciplinares STEM e SHAPE (*e.g.* arquitetura e psicologia são exemplos de disciplinas que transitam facilmente entre ambas as áreas);
- Dificuldades no âmbito da representatividade disciplinar (algumas áreas possuem quantidades significativamente maiores de palestras do que outras);
- A inevitabilidade de algumas decisões terem sido tomadas de forma subjetiva;

Frente às fragilidades expostas acima, é necessário reforçar que a escolha dos tópicos e etiquetagem das TED Talks disponibilizadas são processos realizados internamente e estão além do controle do pesquisador. Quanto à escolha pelas disciplinas e organização interna dos *corpora* STEM e SHAPE, reitero que são condicionadas ao contexto desta tese e estão sujeitas a interpretações subjetivas.

No que se refere aos encaminhamentos, a presente tese sinaliza para diversos tipos de investigações sobre a linguagem formulaica em discursos voltados à popularização do conhecimento. Especificamente, algumas sugestões que considero relevantes são:

Disponibilização pública dos metadados e dos corpora de estudo compilados. Após a conclusão desta tese, pretendo disponibilizar publicamente os metadados e os arquivos de textos coletados para a compilação dos *corpora* de estudo. O propósito dessa iniciativa é facilitar o acesso e contribuir com o avanço de futuras pesquisas que tenham interesse no discurso das TED Talks.

Comparação do quadro de análise situacional com outros registros. Conforme sugerido por Biber e Conrad (2009), a análise situacional de um registro é melhor evidenciada quando comparada a outro registro. No caso das TED Talks, resultados interessantes poderiam emergir da comparação com a situação de produção linguística de outros gêneros de popularização do conhecimento, como os vlogs de divulgação científica *Big Think*® e o 99U Talks®.

Aplicações pedagógicas. No contexto do ensino de línguas para fins específicos, LinFE, os resultados obtidos no presente estudo podem ser utilizados na criação de materiais e na elaboração de atividades destinadas à preparação de estudantes para apresentações orais em contextos acadêmicos, tais como seminários e congressos.

Pacotes lexicais com slot produtivo. Investigações de pacotes lexicais com um (ou mais) *slot* variável (e.g. *the * of the*) em TED Talks também se configura como uma alternativa relevante para o entendimento da produção discursiva neste registro. Conforme argumentado por Gray e Biber (2015) e Reppen e Olson (2020), a identificação de *slots* variáveis em pacotes lexicais pode revelar muito sobre a variação e produtividade linguística desse tipo de sequência formulaica.

Ao analisar as TED Talks sob uma ótica disciplinar, a presente pesquisa buscou expandir a compreensão sobre quais estratégias discursivas estão mais vinculadas a diferentes campos do saber técnico e acadêmico no momento de recontextualização e simplificação do conteúdo informacional para o público não especializado. Os resultados obtidos podem, por exemplo, pavimentar caminhos para que acadêmicos e palestrantes iniciantes possam produzir seus textos de maneira mais adequada a contextos de acadêmicos orais. Finalmente, é válido

reforçar aqui, que, independentemente da natureza disciplinar, qualquer TED Talk objetiva a propagação de ideias valiosas⁶³.

⁶³ *Ideas worth spreading.*

REFERÊNCIAS

- ÄDEL, A.; ERMAN, B. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, v. 31, n. 2, p. 81–92, 2012.
- ALTENBERG, B. On the phraseology of spoken English: The evidence of recurrent word combinations. Em: COWIE, A. P. (Ed.). *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press, 1998. p. 101–122.
- ANDERSON, C. *TED Talks: O guia oficial do TED para falar em público*. Tradução: Donaldson Garschagen; Tradução: Renata Guerra. 1. ed. Rio de Janeiro: Intrínseca, 2016.
- ANTHONY, L. *AntConc* (Version 4.2.0) [Computer Software]. Waseda University, Tokyo, 2020a. Disponível em <https://www.laurenceanthony.net/software/antconc/>. Acesso em 04 abr. 2021.
- ANTHONY, L. *AntGram* (Version 1.2.0) [Computer Software]. Waseda University, Tokyo, 2020b. Disponível em <https://www.laurenceanthony.net/software/antgram/>. Acesso em 04 abr. 2021.
- ASTON, G. Applied Corpus Linguistics and the learning experience. Em: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (Eds.). *Perspectives on Corpus Linguistics*. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2011. v. 48, p. 1–16.
- ATKINS, S.; CLEAR, J.; OSTLER, N. Corpus Design Criteria. *Literary and Linguistic Computing*, v. 7, n. 1, p. 1–16, 1992.
- BAKER, P. Social involvement in corpus studies. Em: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (Eds.). *Perspectives on Corpus Linguistics*. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2011. v. 48, p. 17–27.
- BEACCO, J.-C. et al. Science in media and social discourse: new channels of communication, new linguistic forms. *Discourse Studies*, v. 4, n. 3, p. 277–300, 2002.
- BECHER, T. Disciplinary Discourse. *Studies in Higher Education*, v. 12, n. 3, p. 261–274, 1987.
- BECHER, T. The Significance of Disciplinary Differences. *Studies in Higher Education*, v. 19, n. 2, p. 151–161, 1994.
- BECHER, T.; TROWLER, P. R. *Academic Tribes and Territories*. 2. ed. Buckingham: SRHE, 2001.
- BERBER SARDINHA, T. *Linguística de Corpus*. 1. ed. Barueri: Manole, 2004.
- BERBER SARDINHA, T. Corpus Linguistics in South America. Em: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (Eds.). *Perspectives on Corpus Linguistics*. Amsterdam: John Benjamins Publishing Company, 2011. v. 48, p. 29–45.

BÉRTOLI, P.; SHEPHERD, T. Escrita Acadêmica: Um estudo exploratório de quadrigramas. *The ESpecialist*, v. 36, n. 2, p. 241–262, 2015.

BESTGEN, Y. Comparing Lexical Bundles across Corpora of Different Sizes: The Zipfian Problem. *Journal of Quantitative Linguistics*, v. 27, n. 3, p. 272–290, 2020.

BIBER, D. Representativeness in corpus design. *Literary and Linguistic Computing*, v. 8, n. 4, p. 243–257, 1993.

BIBER, D. et al. *Longman Grammar of Spoken and Written English*. 1. ed. Essex: Longman, 1999.

BIBER, D. *University language: a corpus-based study of spoken and written registers*. 1. ed. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2006a. v. 23.

BIBER, D. Stance in spoken and written university registers. *Journal of English for Academic Purposes*, v. 5, n. 2, p. 97–116, 2006b.

BIBER, D. A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics*, v. 14, n. 3, p. 275–311, 2009.

BIBER, D.; BARBIERI, F. Lexical bundles in university spoken and written registers. *English for Specific Purposes*, v. 26, n. 3, p. 263–286, 2007.

BIBER, D.; CONRAD, S. *Register, Genre and Style*. 2. ed. Cambridge: Cambridge University Press, 2009.

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

BIBER, D.; REPPEN, R. Introduction. Em: BIBER, D.; REPPEN, R. (Eds.). *The Cambridge Handbook of English Corpus Linguistics*. 1. ed. Cambridge: Cambridge University Press, 2015. v. 1, p. 1–8.

BOCORNHY, A. E. P. et al. A produção de artigos da área das ciências da saúde com o auxílio de key lexical bundles: um estudo direcionado por corpus. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada*, v. 37, n. 1, p. 1–37, 2021.

BREZINA, V.; WEILL-TESSIER, P.; MCENERY, T. *#LancsBox*. Lancaster University, 2020. Disponível em: <http://corpora.lancs.ac.uk/lancsbox>. Acesso em 04 abr. 2021.

BYBEE, R. W. *The Case for STEM Education: Challenges and Opportunities*. 1. ed. Arlington: NSTA Press, 2013.

BYRD, P.; COXHEAD, A. On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*, v. 5, n. 5, p. 31–64, 2010.

- CALIENDO, G. The popularization of science in web-based genres. Em: BONGO, G.; CALIENDO, G. (Eds.). *The Language of Popularization- Die Sprache der Popularisierung*. 1. ed. Bern: Peter Lang, 2014. p. 111–138.
- CALIENDO, G.; COMPAGNONE, A. Expressing epistemic stance in University lectures and TED talks: a contrastive corpus-based analysis. *Lingue Linguaggi*, v. 11, n. 11, p. 105–122, 2014.
- CALSAMIGLIA, H. Popularization discourse. *Discourse Studies*, v. 5, n. 2, p. 139–146, 2003.
- CAO, F. A comparative study of lexical bundles across paradigms and disciplines. *Corpora*, v. 16, n. 1, p. 97–128, 2021.
- CETTOLO, M.; GIRARDI, C.; FEDERICO, M. WIT 3 : Web Inventory of Transcribed and Translated Talks. (M. Cettolo et al., Eds.) *Preceedings of the 16th European Association for Machine Translation*. Trento: European Association for Machine Translation, 2012.
- CHANG, Y. J.; HUANG, H. T. Exploring TED talks as a pedagogical resource for oral presentations: A corpus-based move analysis. *English Teaching and Learning*, v. 39, n. 4, p. 29–62, 2015.
- CHEN, L. An Investigation of Lexical Bundles in ESP Textbooks and Electrical Engineering Introductory Textbooks. Em: WOOD, D. (Ed.). *Perspective on Formulaic Language: Acquisition and Communication*. 1. ed. New York: Continuum, 2010. p. 107–127.
- CHEN, Y. H. *Lexical Bundles Across Learner Writing Development*. Tese de doutorado—Lancaster: Lancaster University, 1 dez. 2009.
- CHEN, Y. H.; BAKER, P. Lexical Bundles in L1 and L2 Academic Writing. *Language Learning and Technology*, v. 14, n. 2, p. 30–49, 2010.
- COMPAGNONE, A. The reconceptualization of academic discourse as a professional practice in the digital age: A critical genre analysis of TED Talks. *Hermes*, n. 54, p. 49–69, 2015.
- CONRAD, S.; BIBER, D. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, v. 20, p. 56–71, 2004.
- CORTES, V. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, v. 23, n. 4, p. 397–423, 2004.
- CORTES, V. Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, v. 17, n. 4, p. 391–406, 2006.
- CORTES, V. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, v. 12, n. 1, p. 33–43, 2013.

CORTES, V. Situating lexical bundles in the formulaic language spectrum: Origins and functional analysis developments. Em: CORTES, V.; CSOMAY, E. (Eds.). *Corpus-based research in Applied Linguistics: Studies in Honor of Doug Biber*. 1. ed. Amsterdam / Philadelphia: John Benjamins Publishing, 2015. p. 197–216.

CORTES, V.; CSOMAY, E. Positioning lexical bundles in university lectures. Em: CAMPOY, M. C.; LUZÓN, M. J. (Eds.). *Spoken corpora in applied linguistics*. 1. ed. Bern: Peter Lang, 2007. p. 57–76.

COXHEAD, A.; BYRD, P. Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, v. 16, n. 3, p. 129–147, 2007.

CRAWFORD CAMICIOTTOLI, B. Interactive discourse structuring in L2 guest lectures: Some insights from a comparative corpus-based study. *Journal of English for Academic Purposes*, v. 3, n. 1, p. 39–54, 2004.

CRAWFORD CAMICIOTTOLI, B. Audience-oriented relevance markers in business studies lectures. Em: DEL LUNGO CAMICIOTTI, G.; TOGNINI BONELLI, E. (Eds.). *Academic Discourse: New insights into evaluation*. 1. ed. Lausanne: Peter Lang, 2014. v. 15, p. 81–97.

CSOMAY, E. Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied Linguistics*, v. 34, n. 3, p. 369–388, 2013.

D'AVANZO, S. Speaker identity vs speaker diversity: the case of ted talks corpus. Em: BALIRANO, G.; NISCO, M. C. (Eds.). *Languaging Diversity : Identities, Genres, Discourses*. 1. ed. Newcastle: Cambridge Scholars Publishing, 2015. v. 1, p. 279–296.

DAVIES, M. Synchronic and diachronic uses of corpora. Em: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (Eds.). *Perspectives on Corpus Linguistics*. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2011. v. 48, p. 63–80.

DURRANT, P. Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*, v. 38, n. 2, p. 165–193, 2015.

DURRANT, P.; MATHEWS-AYDINLI, J. A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, v. 30, n. 1, p. 58–72, 2011.

DUTRA, D. P.; BERBER SARDINHA, T. Pacotes lexicais em corpora de aprendizes. Em: IBAÑOS, A. M. T. et al. (Eds.). *Pesquisas e Perspectivas em Linguística de Corpus*. 1. ed. Campinas: Mercado de Letras, 2014. p. 57–80.

DUTRA, D. P.; ORFANO, B. M.; BERBER SARDINHA, T. Stance bundles in learner corpora. Em: ALUISIO, S. M.; TAGNIN, S. E. O. (Eds.). *New language technologies and linguistic research: a two-way road*. 1. ed. Newcastle: Cambridge Scholars Publishing, 2014. p. 2–15.

EGBERT, J. Corpus Design and Representativeness. Em: BERBER SARDINHA, T.; VEIRANO PINTO, M. (Eds.). *Multi-Dimensional Analysis: Research Methods and Current Issues*. 1. ed. Londres / Nova York: Bloomsbury Publishing, 2019. p. 27–42.

ELLIS, N. C.; SIMPSON-VLACH, R.; MAYNARD, C. Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL QUARTERLY*, v. 42, n. 3, p. 375–396, 2008.

ERMAN, B.; WARREN, B. The idiom principle and the open choice principle. *Text*, v. 20, n. 1, p. 29–62, 2000.

FLOWERDEW, J.; MILLER, L. The teaching of academic listening comprehension and the question of authenticity. *English for Specific Purposes*, v. 16, n. 1, p. 27–46, 1997.

FLOWERDEW, L. The argument for using English specialized corpora to understand academic and professional language. Em: CONNOR, U.; UPTON, T. A. (Eds.). *Discourse in the Professions: Perspectives from corpus linguistics*. 1. ed. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2004. v. 16, p. 11–36.

FREITAS, A. L. P. *Proficiência escrita em inglês especializado: estudo de corpus de abstracts em medicina, nutrição e farmácia*. Tese de Doutorado—Porto Alegre, Brasil: Universidade Federal do Rio Grande do Sul, 2016.

GRANGER, S.; PAQUOT, M. Disentangling the phraseological web. Em: GRANGER, S.; MEUNIER, F. (Eds.). *Phraseology: An interdisciplinary perspective*. 1. ed. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2008. v. 1, p. 27–49.

GRAY, B. Lexical Bundles. Em: BAKER, P.; EGBERT, J. (Eds.). *Triangulating Methodological Approaches in Corpus Linguistic Research*. Nova York / Londres: Routledge, 2016. p. 33–56.

GRAY, B.; BIBER, D. Phraseology. Em: BIBER, D.; REPPEN, R. (Eds.). *The Cambridge Handbook of English Corpus Linguistics*. 1. ed. Cambridge: Cambridge University Press, 2015. p. 125–145.

GRONDONA, C. B. *Eliminação de pacotes lexicais relacionados ao tópico e de pacotes lexicais em contexto de sobreposição: uma proposta metodológica para os estudos da linguística de corpus*. Dissertação de Mestrado—Belo Horizonte: Universidade Federal de Minas Gerais, 2015.

HASEBE, Y. Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks. *Procedia - Social and Behavioral Sciences*, v. 198, p. 174–182, 2015.

HOEY, M. *Lexical priming: A new theory of words and language*. 1. ed. London / New York: Taylor & Francis, 2005.

HUNSTON, S. *Corpora in Applied Linguistics*. 6. ed. Cambridge: Cambridge University Press, 2002.

HYLAND, K. *Disciplinary Discourses: Social Interactions in Academic Writing*. 1. ed. Michigan: The University of Michigan Press, 2004a.

HYLAND, K. A convincing argument: Corpus analysis and academic persuasion. Em: CONNOR, U.; UPTON, T. A. (Eds.). *Discourse in the Professions: Perspectives from corpus linguistics*. 1. ed. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2004b. v. 16, p. 87–112.

HYLAND, K. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, v. 7, n. 2, p. 173–192, 2005.

HYLAND, K. Disciplinary Differences: Language Variation in Academic Discourses. Em: HYLAND, K.; BONDI, M. (Eds.). *Academic Discourse Across Disciplines*. 1. ed. Bern: Peter Lang, 2006. p. 17–45.

HYLAND, K. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, v. 27, n. 1, p. 4–21, 2008a.

HYLAND, K. Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, v. 18, n. 1, p. 41–62, 2008b.

HYLAND, K. *Academic Discourse: English in a Global Context*. 1. ed. New York: Continuum, 2009.

HYLAND, K. Constructing proximity: Relating to readers in popular and professional science. *Journal of English for Academic Purposes*, v. 9, n. 2, p. 116–127, 2010.

HYLAND, K. Academic discourse. Em: HYLAND, K.; PALTRIDGE, B. (Eds.). *The Continuum Companion to Discourse Analysis*. 1. ed. New York: Continuum, 2011a. p. 171–184.

HYLAND, K. Looking through corpora into writing practices. Em: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (Eds.). *Perspectives on Corpus Linguistics*. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2011b. v. 48, p. 99–113.

HYLAND, K. Bundles in academic discourse. *Annual Review of Applied Linguistics*, v. 32, p. 150–169, 2012.

HYLAND, K.; JIANG, F. Is academic writing becoming more informal? *English for Specific Purposes*, v. 45, n. 1, p. 40–51, 2017.

HYLAND, K.; JIANG, F. Academic lexical bundles: How are they changing? *International Journal of Corpus Linguistics*, v. 23, n. 4, p. 383–407, 2018.

HYLAND, K.; JIANG, F. *Academic Discourse and Global Publishing Disciplinary Persuasion in Changing Times*. 1. ed. London / New York: Routledge, 2019.

HYLAND, K.; TSE, P. Is There an “Academic Vocabulary”? *TESOL Quarterly*, v. 41, n. 2, p. 235–253, 2007.

JOHANSSON, S. Some aspects of the development of corpus linguistics in the 1970s and 1980s. Em: LUDELING, A.; KYTO, M. (Eds.). *Corpus Linguistics: An International Handbook*. Berlin: Walter de Gruyter, 2008. p. 33–52.

KASHIHA, H.; HENG, C. S. An exploration of lexical bundles in academic lectures: Examples from hard and soft sciences. *Journal of Asia TEFL*, v. 10, n. 4, p. 133–161, 2013.

KASHIHA, H.; HENG, C. S. Discourse functions of formulaic sequences in academic speech across two disciplines. *GEMA Online Journal of Language Studies*, v. 14, n. 2, p. 15–27, 2014.

KEDROWICZ, A. A.; TAYLOR, J. L. Shifting rhetorical norms and electronic eloquence: TED talks as formal presentations. *Journal of Business and Technical Communication*, v. 30, n. 3, p. 352–377, 2016.

KENNEDY, G. *An Introduction to Corpus Linguistics*. 1. ed. London / New York: Longman, 1998.

KILGARRIFF, A. et al. The Sketch Engine. *Information Technology*, v. 04, n. 08, p. 105–116, 2004.

LEECH, G. Corpora and theories of linguistic performance. Em: SVARTVIK, J. (Ed.). *Direction in Corpus Linguistics: Proceedings of Nobel Symposium 82*. Berlin: Mouton de Gruyter, 1992. p. 105–122.

LEECH, G. New resources, or just better old ones? The Holy Grail of representativeness. Em: HUNDT, M.; NESSELHAUF, N.; BIEWER, C. (Eds.). *Corpus Linguistics and the Web*. 1. ed. Amsterdam/New York: Rodopi, 2007. p. 133–149.

LEECH, G. Principles and applications of Corpus Linguistics. Em: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (Eds.). *Perspectives on Corpus Linguistics*. 1. ed. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2011. p. 155–170.

LIU, C. Y.; CHEN, H. J. H. Analyzing the functions of lexical bundles in undergraduate academic lectures for pedagogical use. *English for Specific Purposes*, v. 58, p. 122–137, 2020a.

LIU, C. Y.; CHEN, H. J. H. Functional variation of lexical bundles in academic lectures and TED talks. *Register Studies*, v. 2, n. 2, p. 176–208, 2020b.

LUDEWIG, J. TED talks as an emergent genre. *CLCWeb - Comparative Literature and Culture*, v. 19, n. 1, 2017.

MATTE, M. L.; GOULART, L. Lexical Bundles across levels of Proficiency in Portuguese as a Second Language. *Letras de Hoje*, v. 55, n. 4, p. 477–495, 2020.

MATTIELLO, E. The popularisation of science via TED talks. *International Journal of Language Studies*, v. 11, n. 4, p. 77–106, 2017.

MATTIELLO, E. A corpus-based analysis of scientific TED Talks: Explaining cancer-related topics to non-experts. *Discourse, Context and Media*, v. 28, p. 60–68, 2019.

MCENERY, T.; HARDIE, A. *Corpus Linguistics: Method, Theory and Practice*. 1. ed. Cambridge: Cambridge University Press, 2012.

MCENERY, T.; WILSON, A. *Corpus Linguistics: an introduction*. 2. ed. Edinburgh: Edinburgh University Press, 2001.

MCENERY, T.; XIAO, R.; TONO, Y. *Corpus-based language studies: An advanced resource book*. 1. ed. New York: Routledge; Taylor & Francis, 2006.

MIRANDA, A. A. P. B. *Palestras TED: Um novo gênero do discurso?* Dissertação de Mestrado—São Paulo, Brasil: Pontifícia Universidade Católica de São Paulo, 2016.

MIRANDA, J. A. S. *TED TALKS: A Genre Analysis*. Dissertação de Mestrado—Florianópolis, Brasil: Universidade Federal de Santa Catarina, 2018.

MOTTA-ROTH, D.; SCHERER, A. S. Popularização da ciência: a interdiscursividade entre ciência, pedagogia e jornalismo. *Bakhtiniana: Revista de Estudos do Discurso*, v. 11, n. 2, p. 164–189, 2016.

NEELY, E.; CORTES, V. A little bit about: analyzing and teaching lexical bundles in academic lectures. *Language Value*, v. 1, n. 1, p. 17–38, 2009.

NESI, H.; ALSOP, S. Stories and scenarios: Lecturers' use of fantastic hypothetical events. *Journal of English for Academic Purposes*, v. 53, 2021.

NESI, H.; BASTURKMEN, H. Lexical bundles and discourse signaling in academic lecturers. *International Journal of Corpus Linguistics*, v. 11, n. 3, p. 283–304, 2006.

NESI, H.; THOMPSON, P. *The British Academic Spoken English Corpus Manual*. Warwick / Reading: [s.n.], 2006.

POOS, D.; SIMPSON, R. Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. Em: REPPEN, R.; FITZMAURICE, S. M.; BIBER, D. (Eds.). *Using corpora to explore linguistic variation*. 1. ed. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2002. p. 3–24.

REPPEN, R.; OLSON, S. B. Lexical bundles across disciplines: A look at consistency and variability. Em: RÖMER, U.; CORTES, V.; FRIGNAL, E. (Eds.). *Advances in Corpus-based Research on Academic Writing: Effects of discipline, register, and writer expertise*. 1. ed. Amsterdam / Philadelphia: John Benjamins Publishing, 2020. p. 282–304.

ROWLEY-JOLIVET, E. Image as Text. Aspects of the shared visual language of scientific conference participants. *ASp*, n. 27–30, p. 133–154, 2000.

ROWLEY-JOLIVET, E. Visual discourse in scientific conference papers: A genre-based study. *English for Specific Purposes*, v. 21, n. 1, p. 19–40, 2002a.

ROWLEY-JOLIVET, E. Science in the Making: Scientific Conference Presentations and the Construction of Facts. Em: VENTOLA, E.; SHALOM, C.; THOMPSON, S. (Eds.). *The Language of Conferencing*. Berlin: Peter Lang, 2002b. p. 95–125.

ROWLEY-JOLIVET, E. Different Visions, Different Visuals: A Social Semiotic Analysis of Field-Specific Visual Composition in Scientific Conference Presentations. *Visual Communication*, v. 3, n. 2, p. 145–175, 2004.

SALAZAR, D. *Lexical Bundles in Native and Non-native Writing: Applying a corpus-based study to language teaching*. 1. ed. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2014.

SCHMITT, N.; CARTER, R. Formulaic sequences in action. Em: SCHMITT, N. (Ed.). *Formulaic sequences: Acquisition, processing and use*. 1. ed. Amsterdam: John Benjamins, 2004. p. 1–22.

SCOTT, M. The technological aspect of Corpus Linguistics. Em: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (Eds.). *Perspectives on Corpus Linguistics*. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2011. v. 48, p. 213–219.

SCOTT, M. *WordSmith Tools version 8*. Stroud: Lexical Analysis Software, 2022. Disponível em: <https://www.lexically.net/wordsmith/>. Acesso em 15 abr. 2022.

SCOTTO DI CARLO, G. New trends in knowledge dissemination: TED Talks. *Acta Scientiarum Language and Culture*, v. 36, n. 2, p. 121–130, 2014a.

SCOTTO DI CARLO, G. The role of proximity in online popularizations: The case of TED talks. *Discourse Studies*, v. 16, n. 5, p. 591–606, 2014b.

SCOTTO DI CARLO, G. Stance in TED Talks: Strategic use of subjective adjectives in online popularization. *Ibérica*, n. 29, p. 201–222, 2015.

SHEPHERD, T. M. G. O estatuto da Linguística de corpus: metodologia ou área da Linguística? *Matraga*, v. 16, n. 24, p. 150–172, 2009.

SHEPHERD, T. M. G.; ZYNGIER, S.; VIANA, V. Feixes lexicais e visões de mundo: Um estudo sobre corpus. *Matraga*, v. 19, p. 125–140, 2006.

SHIRAZIZADEH, M.; AMIRFAZLIAN, R. Lexical bundles in theses, articles and textbooks of applied linguistics: Investigating intradisciplinary uniformity and variation. *Journal of English for Academic Purposes*, v. 49, p. 1–13, 2021.

SIMPSON, R. C. Stylistic features of academic speech: The role of formulaic expressions. Em: CONNOR, U.; UPTON, T. A. (Eds.). *Discourse in the Professions: Perspectives from corpus linguistics*. 1. ed. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2004. p. 37–64.

SIMPSON, R. C.; LEE, D. Y. W.; LEICHER, S. *MICASE Manual: The Michigan Corpus of Academic Spoken English*. Michigan: [s.n.]. 2002.

SIMPSON-VLACH, R. Academic speech across disciplines: lexical and phraseological distinctions. Em: HYLAND, K.; BONDI, M. (Eds.). *Academic Discourse Across Disciplines*. 1. ed. Bern: Peter Lang, 2006. p. 295–316.

SIMPSON-VLACH, R.; ELLIS, N. C. An academic formulas list: New methods in phraseology research. *Applied Linguistics*, v. 31, n. 4, p. 487–512, set. 2010.

SINCLAIR, J. M. *Corpus, Concordance, Collocation*. 1. ed. Oxford: Oxford University Press, 1991.

SINCLAIR, J. M. Corpus and text-basic principles. Em: WYNNE, M. (Ed.). *Developing linguistic corpora: A guide to good practice*. 1. ed. Barnsley: Oxbow Books, 2005. p. 1–16.

SIYANOVA-CHANTURIA, A. On the “holistic” nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, v. 11, n. 2, p. 285–301, 1 out. 2015.

SUGIMOTO, C. R. et al. Scientists Popularizing Science: Characteristics and Impact of TED Talk Presenters. *PLoS ONE*, v. 8, n. 4, p. 1–8, 2013.

SWALES, J. A critical view on the use of corpora. Em: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (Eds.). *Perspectives on Corpus Linguistics*. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2011. v. 48, p. 222–227.

SWALES, J. M. *Genre Analysis: English in academic and research settings*. 14. ed. Cambridge: Cambridge University Press, 1990.

TAGNIN, S. E. O. Glossário de Linguística de Corpus. Em: VIANA, V.; TAGNIN, S. E. O. (Eds.). *Corpora no ensino de línguas estrangeiras*. 1. ed. São Paulo: HUB Editorial, 2011. p. 357–362.

TED. *Our mission: Spread ideas, foster community and create impact*. 2022. Disponível em <https://www.ted.com/about/our-organization>. Acesso em 10 de abril de 2022.

TEUBERT, W. My version of corpus linguistics. *International Journal of Corpus Linguistics*, v. 10, n. 1, p. 1–13, 2005.

TOGNINI-BONELLI, E. *Corpus Linguistics at Work*. 2. ed. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2001.

TROWLER, P. Disciplines and Academic Practices. Em: TROWLER, P.; SAUNDERS, M.; BAMBER, V. (Eds.). *Tribes and territories in the 21st century: rethinking the significance of disciplines in higher education*. 1. ed. Londres / Nova York: Routledge, 2012, p. 30–38.

TROWLER, P. Academic Tribes and Territories: the theoretical trajectory. *ÖZG*, v. 25, n. 3, p. 17–26, 2014.

VIANA, V. Linguística de Corpus: conceitos, técnicas e análises. Em: VIANA, V.; TAGNIN, S. E. O. (Eds.). *Corpora no ensino de línguas estrangeiras*. 1. ed. São Paulo: HUB Editorial, 2011. v. 1, p. 25–95.

VIANA, V. *Disciplinary Variation in Academic Writing: A Corpus Study of PhD theses in English Language and Literature*. Tese de Doutorado—Belfast: Queen’s University Belfast, 2012.

VIANA, V.; FAUSTO, F.; ZYNGIER, S. Literatura através do computador: Feixes lexicais e cânone. *Revista Philologus*, v. 13, n. 39, p. 92–105, 2007.

VIANA, V.; O’BOYLE, A. *Corpus Linguistics for English for Academic Purposes*. 1. ed. New York: Routledge, 2022.

VIANA, V.; ZYNGIER, S.; BARNBROOK, G. *Perspectives on Corpus Linguistics*. 1. ed. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2011.

WANG, Y. Lexical bundles in spoken academic ELF. *International Journal of Corpus Linguistics*, v. 22, n. 2, p. 187–211, 2017.

WOOD, D. *Fundamentals of Formulaic Language: An introduction*. 1. ed. London: Bloomsbury Publishing, 2015.

WRAY, A. *Formulaic Language and the Lexicon*. 1. ed. Cambridge: Cambridge University Press, 2002.

WRAY, A.; PERKINS, M. R. The functions of formulaic language: An integrated model. *Language and Communication*, v. 20, n. 1, p. 1–28, 2000.

WRIGHT, H. R. Lexical bundles in stand-alone literature reviews: Sections, frequencies, and functions. *English for Specific Purposes*, v. 54, p. 1–14, 2019.

APÊNDICE A – Pacotes lexicais extraídos do *corpus* TED-STEM

(continua)

#	Pacote Lexical	FR	% Textos	Categoria Funcional	#	Pacote Lexical	FR	% Textos	Categoria Funcional
01	to be able to + (do)	191	22,6%	Posicionamento	26	when it comes to	55	8%	Organizador Discursivo
02	(and) + it turns out that	132	17,3%	Organizador Discursivo	27	you can see that	54	8,1%	Posicionamento
03	at the end of + (the)	105	16,8%	Expressão de Referência	28	a few years ago	52	9,4%	Expressão de Referência
04	if you look at + (the)	100	13,6%	Posicionamento	29	all over the world	51	7,6%	Expressão de Referência
05	at the same time	99	14,5%	Expressão de Referência	30	so this is a	50	8,9%	Organizador Discursivo
06	one of the most	92	14,5%	Organizador Discursivo	31	there's a lot of	50	7,9%	Expressão de Referência
07	(this) + is one of the	82	13,6%	Organizador Discursivo	32	I want to talk + (about)	49	7%	Organizador Discursivo
08	one of the things + (that)	82	12,1%	Organizador Discursivo	33	(I'm) + going to talk about	48	8%	Organizador Discursivo
09	I want you to	80	10,7%	Posicionamento	34	let me give you + (an example)	48	8,5%	Organizador Discursivo
10	and you can see	75	11,1%	Posicionamento	35	on the other hand	48	7,7%	Organizador Discursivo
11	if you want to	74	11,5%	Posicionamento	36	a little bit more	47	7,5%	Expressão de Referência
12	for the first time	71	10,5%	Expressão de Referência	37	let me show you	45	7,4%	Posicionamento
13	a little bit of	69	11%	Expressão de Referência	38	if you think about + (it)	44	7,6%	Organizador Discursivo
14	(I'm) + going to show you	67	9,8%	Posicionamento	39	and this is the	42	7,7%	Organizador Discursivo
15	in the middle of + (the)	65	10,6%	Expressão de Referência	40	we were able to	42	6,9%	Posicionamento
16	and one of the	64	10,7%	Organizador Discursivo	41	have a lot of	41	7,2%	Expressão de Referência
17	the rest of the	63	10,2%	Expressão de Referência	42	let me tell you	41	7%	Organizador Discursivo
18	you can see the	63	10,4%	Posicionamento	43	so this is the	41	7%	Organizador Discursivo
19	a lot of people	61	10%	Expressão de Referência	44	that we need to	41	6,4%	Posicionamento
20	is going to be	61	9,7%	Posicionamento	45	(I'm) + going to tell you	40	6,9%	Organizador Discursivo
21	and this is a	58	10,5%	Organizador Discursivo	46	(what) + we need to do	40	6,9%	Posicionamento
22	and I want to	55	9%	Posicionamento	47	the size of a	40	6,4%	Expressão de Referência
23	and I'm going to	55	9%	Posicionamento	48	to talk to you + (about)	40	6,9%	Organizador Discursivo
24	as you can see	55	8,7%	Posicionamento	49	going to be a	38	6,4%	Posicionamento
25	what I want to + (do)	55	9,3%	Posicionamento	50	so I'm going to	38	6%	Posicionamento

(continua)

#	Pacote Lexical	FR	% Textos	Categoria Funcional	#	Pacote Lexical	FR	% Textos	Categoria Funcional
51	to tell you about	38	6,5%	Organizador Discursivo	76	not going to be	30	5,2%	Posicionamento
52	you don't have to	37	6,2%	Posicionamento	77	so I want to	30	4,7%	Posicionamento
53	I would like to	36	6,2%	Organizador Discursivo	78	the surface of the	30	4,9%	Expressão de Referência
54	to figure out how	36	5,7%	Posicionamento	79	this is what we	30	5,2%	Organizador Discursivo
55	what I'm going to	36	5,7%	Posicionamento	80	to look at the	30	5,2%	Organizador Discursivo
56	and this is what	35	6,1%	Organizador Discursivo	81	I don't want to	29	5%	Posicionamento
57	are going to be	35	6,1%	Posicionamento	82	to give you a	29	4,6%	Organizador Discursivo
58	I want to show + (you)	35	5%	Posicionamento	83	to make sure that	29	4,9%	Posicionamento
59	I want to tell + (you)	35	5,9%	Organizador Discursivo	84	if you have a	28	5,2%	Organizador Discursivo
60	a little bit about	33	6,1%	Expressão de Referência	85	in the case of	28	5,1%	Expressão de Referência
61	and a lot of	33	6,1%	Expressão de Referência	86	to do is to	28	4,9%	Organizador Discursivo
62	it's going to be	33	5,9%	Posicionamento	87	when you look at	28	5%	Posicionamento
63	the size of the	33	5%	Expressão de Referência	88	and we need to	27	4,5%	Posicionamento
64	on the other side	32	5,2%	Expressão de Referência	89	might be able to	27	4,9%	Posicionamento
65	so you can see	32	5,4%	Posicionamento	90	so how do we	27	5,1%	Organizador Discursivo
66	take a look at	32	5,1%	Posicionamento	91	so that we can	27	4,9%	Posicionamento
67	this is a very	32	6,1%	Organizador Discursivo	92	this is going to	27	4,6%	Posicionamento
68	a lot of the	31	6%	Expressão de Referência	93	this is not a	27	5,2%	Organizador Discursivo
69	hundreds of thousands of	31	5,5%	Expressão de Referência	94	we need to be	27	5%	Posicionamento
70	I can tell you	31	5,4%	Posicionamento	95	you can see here	27	4,9%	Posicionamento
71	the way that we	31	4,5%	Expressão de Referência	96	you can think of	27	4,4%	Organizador Discursivo
72	when I was a	31	5,5%	Expressão de Referência	97	a lot of time	26	5,1%	Expressão de Referência
73	all of a sudden	30	4,5%	Organizador Discursivo	98	and I think that	26	4,6%	Posicionamento
74	for a long time	30	5,1%	Expressão de Referência	99	going to have to	26	4%	Posicionamento
75	in the first place	30	5,1%	Organizador Discursivo	100	will be able to	26	4,2%	Posicionamento

(conclusão)

#	Pacote Lexical	FR	% Textos	Categoria Funcional	#	Pacote Lexical	FR	% Textos	Categoria Funcional
101	and by the way	25	4,6%	Organizador Discursivo	126	that we have to	22	4,2%	Posicionamento
102	at the university of	25	4,7%	Expressão de Referência	127	those of you who	22	4%	Organizador Discursivo
103	If we want to	25	4,5%	Posicionamento	128	to do with the	22	4%	Organizador Discursivo
104	tens of thousands of	25	4,6%	Expressão de Referência	129	to tell you that	22	4,1%	Posicionamento
105	this is a picture	25	4%	Organizador Discursivo	130	well it turns out	22	4%	Organizador Discursivo
106	to share with you	25	4,5%	Organizador Discursivo	131	what can we do	22	4,4%	Organizador Discursivo
107	trying to figure out	25	4,7%	Posicionamento	132	be able to see	21	4%	Posicionamento
108	turns out to be	25	4,2%	Organizador Discursivo	133	I think this is	21	4%	Posicionamento
109	we are going to	25	4%	Posicionamento	134	the good news is	21	4%	Organizador Discursivo
110	a bit of a	24	4,4%	Expressão de Referência	135	this is the first	21	4,2%	Organizador Discursivo
111	a couple of years	24	4,4%	Expressão de Referência	136	what we do is	21	4%	Organizador Discursivo
112	and we're going to	24	4,5%	Posicionamento	137	one of the first	20	4,1%	Organizador Discursivo
113	are we going to	24	4,4%	Organizador Discursivo	138	one of the reasons	20	4,1%	Organizador Discursivo
114	in a way that	24	4,6%	Expressão de Referência					
115	some of the most	24	4,2%	Organizador Discursivo					
116	the top of the	24	4,2%	Expressão de Referência					
117	there are a lot + (of)	24	4,5%	Expressão de Referência					
118	to go to the	24	4,2%	Expressão de Referência					
119	what do we do	24	4,1%	Organizador Discursivo					
120	what was going on	24	4,1%	Organizador Discursivo					
121	but it turns out	23	4,4%	Organizador Discursivo					
122	so we have to	23	4,1%	Posicionamento					
123	to come up with	23	4,2%	Posicionamento					
124	you can see it	23	4,2%	Posicionamento					
125	I'm not going to	22	4,1%	Posicionamento					

APÊNDICE B – Pacotes lexicais extraídos do *corpus* TED-SHAPE

(continua)

#	Pacote Lexical	FR	% Textos	Categoria Funcional	#	Pacote Lexical	FR	% Textos	Categoria Funcional
01	(at) + the end of the	133	19,5%	Expressão de Referência	26	let me give you + (an example)	43	7,6%	Organizador Discursivo
02	at the same time	94	13,6%	Expressão de Referência	27	we were able to	41	5,8%	Posicionamento
03	to be able to	91	13,1%	Posicionamento	28	and this is the	39	6%	Organizador Discursivo
04	one of the most	88	14,5%	Organizador Discursivo	29	it turns out that	39	6,2%	Organizador Discursivo
05	for the first time + (in)	78	12,8%	Expressão de Referência	30	that we need to	38	6,3%	Posicionamento
06	if you want to	78	10,9%	Posicionamento	31	a few years ago	37	6,9%	Expressão de Referência
07	if you look at + (the)	77	11,7%	Posicionamento	32	and this is a	37	6%	Organizador Discursivo
08	a lot of people	71	11,2%	Expressão de Referência	33	we live in a	37	6,5%	Expressão de Referência
09	(this) + is going to be	69	10,7%	Posicionamento	34	a little bit about	36	5,1%	Expressão de Referência
10	(this) + is one of the	68	12,0%	Organizador Discursivo	35	for a long time	36	6,2%	Expressão de Referência
11	when it comes to	66	10,2%	Organizador Discursivo	36	on the other side	36	5,5%	Expressão de Referência
12	all over the world	61	9,4%	Expressão de Referência	37	I was going to	35	5,6%	Posicionamento
13	in the middle of + (the)	59	9,9%	Expressão de Referência	38	one of the first	35	6,3%	Organizador Discursivo
14	I want to talk + (about)	56	8,9%	Organizador Discursivo	39	(I'm) + going to talk about	34	5,9%	Organizador Discursivo
15	one of the things + (that)	56	8,4%	Organizador Discursivo	40	all of a sudden	34	4%	Organizador Discursivo
16	if you think about + (it)	54	8,1%	Organizador Discursivo	41	and one of the	34	6,5%	Organizador Discursivo
17	(we're) + going to have to	51	6%	Posicionamento	42	how many of you	34	4%	Organizador Discursivo
18	a little bit of	51	8,3%	Expressão de Referência	43	I don't want to	34	5,5%	Posicionamento
19	I want you to	51	6,6%	Posicionamento	44	in a way that	34	5,9%	Expressão de Referência
20	the rest of the	51	7,7%	Expressão de Referência	45	let me tell you	34	6,2%	Organizador Discursivo
21	to make sure that	45	6,9%	Posicionamento	46	on the other hand	34	5,2%	Organizador Discursivo
22	as you can see	44	7,3%	Posicionamento	47	when I was a	34	5,8%	Expressão de Referência
23	and I want to	43	6,9%	Posicionamento	48	over and over again	33	5,1%	Expressão de Referência
24	are we going to	43	6,2%	Organizador Discursivo	49	and I'm going to	31	5,3%	Posicionamento
25	I would like to	43	6,6%	Organizador Discursivo	50	I want to tell + (you)	31	4,9%	Organizador Discursivo

(continua)

#	Pacote Lexical	FR	% Textos	Categoria Funcional	#	Pacote Lexical	FR	% Textos	Categoria Funcional
51	in the first place	31	5,3%	Organizador Discursivo	76	If we want to	25	4,7%	Posicionamento
52	so that we can	31	5%	Posicionamento	77	not going to be	25	4%	Posicionamento
53	that we have to	31	5,8%	Posicionamento	78	was one of the	25	5%	Organizador Discursivo
54	we are going to	31	4,4%	Posicionamento	79	and a lot of	24	4,2%	Expressão de Referência
55	and this is what	29	5,2%	Organizador Discursivo	80	going to be a	24	4,2%	Posicionamento
56	are going to be	29	5,1%	Posicionamento	81	going to tell you	24	4%	Organizador Discursivo
57	it's going to be	29	4,5%	Posicionamento	82	have a lot of	24	4,4%	Expressão de Referência
58	to share with you	29	5%	Organizador Discursivo	83	I can tell you	24	4,2%	Posicionamento
59	and I think that	28	4,5%	Posicionamento	84	tens of thousands of	24	4,4%	Expressão de Referência
60	and we need to	28	5,2%	Posicionamento	85	we have to do	24	4,4%	Posicionamento
61	I'm not going to	28	4,8%	Posicionamento	86	what I want to	24	4,2%	Posicionamento
62	in terms of the	28	4,4%	Expressão de Referência	87	and by the way	23	4,1%	Organizador Discursivo
63	and you can see	27	4,2%	Posicionamento	88	I don't know what	23	4,2%	Posicionamento
64	we need to do	27	4,7%	Posicionamento	89	I'm going to do	23	4,2%	Posicionamento
65	you can see the	27	4,1%	Posicionamento	90	to give you a	23	4,5%	Organizador Discursivo
66	a lot of the	26	4,5%	Expressão de Referência	91	what it means to	23	4%	Expressão de Referência
67	and we have to	26	4,7%	Posicionamento	92	a lot of time	22	4,1%	Expressão de Referência
68	I just want to	26	4,5%	Posicionamento	93	I don't know if	22	4,2%	Posicionamento
69	in the same way	26	4%	Expressão de Referência	94	in the face of	22	4%	Expressão de Referência
70	there's a lot of	26	4,1%	Expressão de Referência	95	some of the most	22	4%	Organizador Discursivo
71	this is not a	26	4,7%	Organizador Discursivo	96	the beginning of the	22	4%	Expressão de Referência
72	to talk to you	26	4,8%	Organizador Discursivo	97	the problem is that	22	4%	Organizador Discursivo
73	you don't have to	26	4,5%	Posicionamento	98	there are a lot + (of)	22	4,1%	Expressão de Referência
74	as a result of	25	4%	Expressão de Referência	99	to go back to	22	4%	Organizador Discursivo
75	for all of us	25	4,8%	Organizador Discursivo	100	was going to be	22	4%	Posicionamento

(conclusão)

#	Pacote Lexical	FR	% Textos	Categoria Funcional	#	Pacote Lexical	FR	% Textos	Categoria Funcional
101	we have to be	22	4%	Posicionamento	111	so I'm going to	20	4,1%	Posicionamento
102	what we need to	22	4%	Posicionamento	112	so this is the	20	4%	Organizador Discursivo
103	a little bit more	21	4,1%	Expressão de Referência	113	there are so many	20	4%	Expressão de Referência
104	I think this is	21	4%	Posicionamento	114	to go to the	20	4%	Expressão de Referência
105	in the history of	21	4,1%	Expressão de Referência					
106	this is a very	21	4%	Organizador Discursivo					
107	those of you who	21	4,9%	Organizador Discursivo					
108	to talk about the	21	4%	Organizador Discursivo					
109	a couple of years	20	4%	Expressão de Referência					
110	one of the reasons	20	4%	Organizador Discursivo					