

## A produção de artigos da área das ciências da saúde com o auxílio de *key lexical bundles*: um estudo direcionado por corpus

*Writing scientific articles with the support of key lexical bundles: a corpus-driven study in the area of health sciences<sup>1</sup>*

Ana Eliza Pereira Bocorny<sup>2</sup>

Rozane Rebechi<sup>3</sup>

Randi Reppen<sup>4</sup>

Maria Claudia Nunes Delfino<sup>5</sup>

Vivian Meurer Lameira<sup>6</sup>

1. Agradecimento: Este trabalho foi conduzido durante o período de concessão da bolsa de Professor Visitante no Exterior na Universidade do Norte do Arizona (EUA) e financiado pelo Programa Institucional de Internacionalização da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES/PRINT/UFRGS - Edital N° 003/2019, no âmbito do Ministério da Educação do Brasil.

2. Universidade Federal do Rio Grande do Sul – UFRGS. Rio Grande do Sul – Brasil. <https://orcid.org/0000-0002-0515-9630>. E-mail: [ana.bocorny@gmail.com](mailto:ana.bocorny@gmail.com).

3. Universidade Federal do Rio Grande do Sul – UFRGS. Rio Grande do Sul – Brasil. <https://orcid.org/0000-0002-1878-7548>. E-mail: [rozanereb@gmail.com](mailto:rozanereb@gmail.com).

4. Northern Arizona University. Arizona – EUA. <https://orcid.org/0000-0001-5657-9195>. E-mail: [randi.reppen@nau.edu](mailto:randi.reppen@nau.edu).

5. Pontifícia Universidade Católica de São Paulo - PUC-SP. São Paulo – Brasil. <https://orcid.org/0000-0003-1976-8725>. E-mail: [claudia@fatecpq.com.br](mailto:claudia@fatecpq.com.br).

6. Universidade Federal do Rio Grande do Sul – UFRGS. Rio Grande do Sul – Brasil. <https://orcid.org/0000-0002-5429-6252>. E-mail: [vilameira@gmail.com](mailto:vilameira@gmail.com).



This content is licensed under a Creative Commons Attribution License, which permits unrestricted use and distribution, provided the original author and source are credited.

## RESUMO

*O inglês é a língua franca da academia. Publicar nesse idioma aumenta a disseminação do conhecimento científico e as possibilidades de citação. Frente ao exposto, buscamos contribuir para o aumento do impacto da produção científica brasileira, propondo recursos pedagógicos baseados em padrões linguísticos obtidos com uma abordagem direcionada por corpus. Por meio de uma análise empírica de um corpus de artigos acadêmicos publicados em periódicos indexados, e utilizando os pressupostos da Linguística de Corpus, objetivamos levantar key lexical bundles na introdução de artigos em inglês da área da saúde, classificar esses bundles de acordo com sua função comunicativa e sugerir aplicações pedagógicas para os dados coletados que visem à redação de artigos acadêmicos. Os dados, levantados a partir da ferramenta de análise textual Sketch Engine mostram que (i) há concentração de key lexical bundles em certos movimentos e passos da seção estudada e (ii) é possível agrupar unidades com mesma função comunicativa identificadas em um mesmo passo de forma a construir uma estrutura lexical (lexical frame) que pode ser usada na construção de recursos pedagógicos.*

**Palavras-chave:** *Linguística de Corpus; key lexical bundles; recursos pedagógicos direcionados por corpus; escrita acadêmica.*

## ABSTRACT

*English is the lingua franca in academic settings. Publishing in this language enhances the spread of scientific knowledge and citations. Hence, we intend to contribute to increase the impact of Brazilian scientific production by proposing pedagogical resources based on linguistic patterns obtained with a corpus-driven approach. Through an empirical analysis of a corpus of academic articles published in indexed journals, and using Corpus Linguistics principles, we aim to identify key lexical bundles from the introduction section of articles in the area of health sciences, classify these bundles according to their communicative function and suggest pedagogical applications for the collected data aimed at writing academic articles. The data, collected with a textual analysis tool (Sketch Engine), show that (i) key lexical bundles concentrate in certain movements and steps of the section under study; and (ii) units can be grouped according to the communicative functions identified in the same step and, hence, a lexical frame is built to be used in the construction of pedagogical resources.*

**Keywords:** *corpus linguistics; key lexical bundles; corpus-driven teaching resources; academic writing.*

## 1. Introdução

O ensino superior mudou consideravelmente nas últimas décadas (cf. Mok, 2015; Blessinger & Anchan, 2015; Altbach *et al.*, 2019). Uma dessas mudanças está relacionada ao fato de um grande número de instituições de ensino superior em todo o mundo estar buscando internacionalizar suas atividades (cf. Baumvol, 2018; Robson & Wihlborg, 2018). Sendo o inglês a língua franca da academia (cf. Jenkins, 2009; Tardy, 2004; Ammon, 2001), publicar nesse idioma facilita a disseminação do conhecimento científico produzido no país e aumenta as chances de o trabalho ser citado por outros autores, assim como também as oportunidades de colaboração científica internacional (cf. Meneghini & Packer, 2007, *apud* Baumvol, 2018). Apesar de haver movimentos em prol da aceitação de uma linguagem que visa ao sucesso da comunicação, independentemente do uso da gramática normativa (cf. Crystal, 2003), a utilização da linguagem considerada convencional pela comunidade discursiva da área ainda é determinante para o aceite de artigos submetidos a periódicos de circulação internacional.

Tendo em vista o contexto apresentado, os objetivos deste estudo são: (i) identificar *key lexical bundles*<sup>7</sup>, ou sequências contínuas e recorrentes de palavras consideradas características de uma coleção de textos – neste caso da seção *Introdução* de artigos acadêmicos publicados em inglês em periódicos internacionais da área das ciências da saúde –, quando comparadas a um corpus de referência; (ii) classificar os *key lexical bundles* de acordo com sua função comunicativa nos movimentos retóricos da seção *Introdução* dos artigos; e (iii) sugerir como os dados coletados podem informar a construção de recursos pedagógicos de livre acesso direcionados por corpus (*corpus-driven*) para auxiliar pesquisadores brasileiros na redação de artigos acadêmicos em inglês nas áreas selecionadas. Acredita-se que a utilização dos referidos recursos pedagógicos possa levar ao aprimoramento dos artigos produzidos em inglês pelos pesquisadores brasileiros e a um consequente aumento do impacto desses textos.

---

7. Optamos por manter o termo *key lexical bundle* (KLB) em inglês por ainda não haver um equivalente consagrado em português.

O presente artigo está assim organizado: na seção 2 são apresentados os pressupostos teóricos relativos à Linguística de Corpus e aos estudos sobre gêneros do discurso que dizem respeito a este estudo. A seção 3 descreve os procedimentos metodológicos usados para a extração dos *key lexical bundles*. Na seção 4 os resultados obtidos são relatados e discutidos. Por fim, a seção 5 traz as conclusões e propostas de trabalhos futuros.

## 2. Revisão de literatura

O resultado prático que se pretende atingir com este estudo deriva do encontro e do entrelaçamento de pressupostos teóricos oriundos de duas áreas do conhecimento: (i) os princípios da Linguística de Corpus e (ii) os estudos sobre gêneros do discurso.

### 2.1. Linguística de Corpus

A Linguística de Corpus parte de uma perspectiva de descrição da linguagem autêntica, quer ocorra no âmbito geral ou especializado. A visão da língua como um sistema probabilístico é um dos fundamentos principais da Linguística de Corpus (cf. Berber Sardinha, 2004). Os traços linguísticos (lexicais, estruturais, pragmáticos e discursivos) não ocorrem todos com a mesma regularidade (cf. Berber Sardinha, 2004). Por esse motivo, a variação dos traços não é aleatória; pelo contrário, existe “um mapeamento regular entre a frequência maior ou menor de um traço e um contexto de ocorrência” (Berber Sardinha 2000, p. 351). Logo, defender que os traços não são aleatórios significa dizer que a linguagem é formulaica, i.e., padronizada. Como afirma Berber Sardinha (2000, p. 351), “A padronização se evidencia pela recorrência, isto é, uma colocação, coligação ou estrutura que se repete significativamente mostra sinais de ser, na verdade, um padrão lexical ou léxico-gramatical”.

#### 2.1.1. Pesquisa baseada em corpus *versus* pesquisa direcionada pelo corpus

Biber (2012) diferencia estudos baseados em corpus (*corpus-based studies*) de estudos direcionados por corpus (*corpus-driven studies*).

O primeiro tem como principal objetivo de pesquisa “analisar os padrões sistemáticos de variação e uso das características linguísticas pré-definidas”<sup>8</sup> (Biber, 2012, p. 162)<sup>9</sup>, enquanto o segundo busca uma perspectiva de pesquisa em que “as próprias construções linguísticas emergem da análise do corpus”<sup>10</sup> (Biber, 2012, p. 162). Assim, a abordagem direcionada por corpus difere da pesquisa baseada em corpus na medida em que faz “o mínimo de afirmações a priori em relação às características linguísticas empregadas na análise de um corpus”<sup>11</sup> (Biber, 2012, p. 162). A perspectiva adotada neste estudo é *corpus-driven*, pois não partimos de categorias *a priori*, mas sim da identificação do objeto de estudo no corpus com base na sua frequência de ocorrência, para posteriores análise e classificação.

### 2.1.2. Linguagem formulaica

A identificação de padrões lexicais, entendidos aqui como combinações de palavras que apresentam certa estabilidade e frequência (cf. Hyland, 2012), tem grande importância pedagógica no âmbito do ensino das linguagens especializadas. Para capacitar pesquisadores menos proficientes a produzir textos conforme os padrões considerados convencionais pelos especialistas das áreas nas quais se inserem, é preciso identificar os padrões recorrentes utilizados em diferentes disciplinas.

Textos especializados comunicam informações específicas por meio de elementos linguísticos (termos, unidades terminológicas, unidades fraseológicas, *lexical bundles*) e não linguísticos (gráficos, tabelas, imagens). Segundo o princípio da idiomaticidade de Sinclair (1991), a comunicação acontece, em grande parte, por meio de unidades formulaicas e não com a seleção aleatória de palavras individuais. As relações que essas sequências formulaicas estabelecem entre si e com outros elementos do texto se constituem como fios que se entrelaçam em uma trama de sentido, resultando em uma linguagem que pode ser

8. Todas as traduções neste artigo são de nossa autoria.

9. No original: “to analyze the systematic patterns of variation and use for those pre-defined linguistic features” (Biber, 2012, p. 162).

10. No original: “the linguistic constructs themselves emerge from analysis of a corpus” (Biber, 2012, p. 162).

11. No original: “minimal a priori assumptions regarding the linguistic features that should be employed for the corpus analysis” (Biber, 2012, p. 162).

considerada mais fluente e idiomática (cf. Hyland, 2012; Paquot & Granger, 2012).

Sequências formulaicas têm sido amplamente estudadas no âmbito do discurso acadêmico por meio da Linguística de Corpus (cf. Biber & Conrad, 1999; Biber *et al.*, 2004; Biber & Barbieri, 2007; Biber, 2009; Staples *et al.*, 2013; Gray & Biber, 2013). As sequências formulaicas recorrentes e contínuas de três ou mais palavras (por exemplo, / *the aim of this paper is*) mais frequentes em dado registro são chamadas de *lexical bundles* (Biber *et al.*, 1999; Biber *et al.*, 2004), enquanto as sequências formulaicas recorrentes descontínuas, isto é, com *slots* variáveis (por exemplo, / *don't \* to, it is \* to*), são denominadas *formulaic frames* (Biber, 2009) ou *lexical frames* (Gray & Biber, 2013).

Sobre os *lexical bundles*, Biber e Barbieri (2007) destacam três características que os diferenciam de outras expressões formulaicas: (i) são extremamente comuns, (ii) não são expressões idiomáticas, já que podem ser compreendidos a partir do significado das palavras que os compõem, e (iii) geralmente não representam uma unidade estrutural completa. Gray (2016) destaca a importância de tais unidades como componentes básicos da construção do discurso. A autora também ressalta a importância dos *lexical bundles* como auxílio aos usuários da língua na realização de funções discursivas específicas.

### 2.1.3. Key lexical bundles

Para a Linguística de Corpus, a palavra-chave (*keyword*) é “uma palavra estatisticamente significativa que caracteriza um documento, texto ou corpus” (Rayson, 2012, p. 1). Em outros termos, trata-se de um elemento que ocorre estatisticamente com mais frequência em um corpus de estudo (isto é, aquele que se pretende investigar de forma mais detalhada) do que em um corpus de referência (ou seja, aquele usado como termo de comparação).

Para os propósitos desta pesquisa, o termo *key lexical bundle* (doravante, *KLB*) será utilizado para designar o nosso objeto de estudo. Como os *lexical bundles*, os *KLBs* são sequências contínuas e recorrentes de três ou mais palavras bastante frequentes, que usualmente não representam estruturas completas. O que diferencia os *KLBs* dos

*lexical bundles* é o fato de os primeiros serem unidades identificadas como típicas de uma coleção de textos de um domínio, gênero e/ou seção específicos, quando comparados a um corpus de referência, o que leva essas unidades a desempenharem funções discursivas específicas na construção do discurso especializado de determinada disciplina ou área de especialidade.

## 2.2. *Os gêneros do discurso e a análise de movimentos retóricos*

Em sua primeira proposta de análise dos movimentos retóricos nas introduções de artigos acadêmicos, Swales (1981) apresenta o modelo *CARS (Create a Research Space)*, segundo o qual a organização retórica é composta de movimentos (*moves*), que expressam funções comunicativas. No caso das introduções de artigos de pesquisa há três movimentos obrigatórios: (i) estabelecendo o território; (ii) estabelecendo um nicho; e (iii) ocupando o nicho. Tais movimentos retóricos, por sua vez, são divididos em unidades menores, chamadas passos (*steps*), que podem ser obrigatórios ou opcionais. O modelo é revisado pelo autor em trabalhos posteriores (Swales, 1990; 2004). Dentre as contribuições de Swales (1981, 1990, 2004) aos estudos de gênero, a mais relevante para esta pesquisa é, justamente, a sua investigação sobre a macroestrutura da introdução do gênero artigo acadêmico por meio da análise dos padrões organizacionais e retóricos.

Os estudos de Swales (1981, 1990, 2004) influenciaram várias pesquisas sobre gêneros acadêmicos em diversas disciplinas (Nwogu, 1997, na medicina; Kanoksilapatham, 2005, na bioquímica; Cortes, 2013, em contexto multidisciplinar), tendo sido o seu modelo adaptado de acordo com os propósitos de cada um dos trabalhos. Nwogu (1997) apresentou um estudo sobre a organização retórica de trinta artigos de pesquisa em inglês da área da medicina. A análise detalhada de uma amostra de quinze artigos do corpus inicialmente compilado revelou onze movimentos retóricos típicos dos artigos da disciplina estudada. Kanoksilapatham (2005), por sua vez, realizou a análise de movimentos retóricos de 60 artigos da área da bioquímica, propondo um modelo de organização para os artigos dessa disciplina.

Swales (2004) ressalta que os movimentos retóricos são unidades discursivas, e não unidades formais, que desempenham uma função comunicativa que os membros da comunidade discursiva pretendem expressar. Para tanto, são usados elementos linguísticos adequados à expressão do propósito comunicativo e reconhecíveis pelos membros dessa comunidade. O estudo de Cortes (2013) sobre movimentos retóricos de gêneros acadêmicos e linguagem formulaica concentra-se na relação entre *lexical bundles* e movimentos retóricos em algumas seções de artigos acadêmicos de diversas disciplinas. Cortes (2013) inova, portanto, propondo uma abordagem baseada em corpus na qual relaciona movimentos retóricos e *lexical bundles* como elementos que, combinados, constroem o discurso. Ela identifica *lexical bundles* em introduções de artigos de pesquisa como uma etapa inicial para a análise dessas expressões nas diferentes seções dos referidos artigos. Assim, a partir de um corpus de estudo de um milhão de palavras de introduções de artigos de treze disciplinas diferentes, Cortes (2013) propõe a classificação gramatical e funcional dos *lexical bundles* e a posterior realização da análise de movimentos retóricos.

No presente trabalho, a partir dos resultados de Cortes (2013), relacionamos elementos formulaicos e recorrentes do discurso aos movimentos retóricos da seção *Introdução* de artigos acadêmicos. Diferentemente de Cortes (2013), contudo, temos como objeto de estudo os *key lexical bundles*, identificados em um corpus de aproximadamente um milhão de palavras provenientes de textos de apenas uma área do conhecimento. Assim, propomos um estudo direcionado por corpus (*corpus-driven*) que identifica e faz uma classificação funcional de *key lexical bundles*, relacionando tais estruturas recorrentes a movimentos retóricos em apenas uma área do conhecimento, buscando identificar o que é típico da seção, gênero e domínio escolhidos, e não o que é generalizável. Nossos dados são extraídos de um corpus de estudo de aproximadamente um milhão de palavras de textos da seção *Introdução* de artigos acadêmicos no âmbito das ciências da saúde, e são comparados a um corpus de referência composto por artigos de outras áreas do conhecimento.

Em função da proximidade das áreas, optamos por usar a adaptação do modelo de Swales (1990) feita por Kanoksilapatham (2005), construída para artigos da área da bioquímica, como referência para a



classificação dos *key lexical bundles* identificados em nosso corpus de estudo, composto de artigos das áreas da saúde. Partindo do modelo de Swales (1990), Kanoksilapatham (2005) identifica a frequência de ocorrência dos movimentos retóricos nos artigos da área da bioquímica a partir de um ponto de corte de 60%, ou seja, “para ser reconhecido como um movimento retórico convencional, o movimento deve ocorrer em 60% das seções apropriadas do corpus.”<sup>12</sup> (Kanoksilapatham 2005, p. 272). O Quadro 1 mostra a frequência de ocorrência dos movimentos retóricos identificados por Kanoksilapatham (2005) na seção *Introdução* dos artigos da área da bioquímica. Como pode ser observado, os movimentos retóricos 1 e 3 estavam presentes em 100% dos artigos analisados. O movimento retórico 2 se mostrou menos frequente, ocorrendo em 66% dos artigos.

**Quadro 1** – Frequência de ocorrência dos movimentos retóricos identificados por Kanoksilapatham (2005) na seção *Introdução* dos artigos da área da bioquímica.

<b>Movimento retórico (doravante MR) Passo (doravante P)</b>	<b>Frequência de ocorrência (%)</b>
<b>MR1:</b> Apresentando a relevância do campo <b>P1:</b> Defendendo a centralidade do tópico	100%
<b>MR1:</b> Apresentando a relevância do campo <b>P2:</b> Fazendo generalizações	
<b>MR1:</b> Apresentando a relevância do campo <b>P3:</b> Revisando pesquisas prévias	
<b>MR2:</b> Preparando para a presente pesquisa <b>P1:</b> Indicando lacunas	66%
<b>MR2:</b> Preparando para a presente pesquisa <b>P2:</b> Levantando questões	
<b>MR3:</b> Introduzindo o presente estudo <b>P1:</b> Estabelecendo objetivo(s)	100%
<b>MR3:</b> Introduzindo o presente estudo <b>P2:</b> Descrevendo procedimentos	
<b>MR3:</b> Introduzindo o presente estudo <b>P3:</b> Apresentando resultados	

Fonte: Elaborado pelas autoras.

12. No original: “to be recognized as a conventional move, a move must occur in 60% of the appropriate sections in the corpus” (Kanoksilapatham, 2005, p. 272).

### 3. Corpus e procedimentos metodológicos

Nesta seção, detalharemos a compilação do corpus de estudo e de referência, assim como as análises quantitativa e qualitativa. A análise quantitativa será feita por meio de ferramentas computacionais e a qualitativa por meio da análise manual dos textos.

#### 3.1. Constituição do corpus de estudo e do corpus de referência

Os dados para esta pesquisa foram coletados a partir de um corpus de estudo, com aproximadamente um milhão de palavras, composto de textos de seção *Introdução* dos artigos acadêmicos estudados. Escritos em língua inglesa, todos os artigos são da área das ciências da saúde, publicados entre 2003 e 2019 em periódicos internacionais de livre acesso, revisados por pares. Também foi utilizado um corpus de referência de aproximadamente cinco milhões de palavras composto de textos de seção *Introdução* dos artigos acadêmicos de diferentes áreas do conhecimento, exceto da área das ciências da saúde. Os artigos que compõem o corpus de referência, publicados entre 2003 e 2019 em periódicos internacionais de livre acesso, também foram revisados por pares.

#### 3.2. Ferramenta e critérios para compilação dos corpora

A compilação dos corpora de estudo e de referência foi grandemente facilitada pelo uso da ferramenta *AntCorGen* (Anthony, 2019), que permite a compilação automática de textos das diferentes seções de artigos de diversas disciplinas e áreas do conhecimento. A ferramenta rastreia (*crawls*) os textos da plataforma *PLoS ONE*<sup>13</sup>, conforme os parâmetros selecionados pelo pesquisador.

Tendo em vista os objetivos desta pesquisa, os critérios considerados para a compilação do corpus de estudo foram: (i) o registro e o gênero dos textos; (ii) o meio, o idioma e o ano (ou período) da

13. Disponível em: <https://journals.plos.org/plosone/>. Acesso em: 20 out. 2019.

publicação; (iii) os domínios aos quais os textos estavam relacionados; e (iv) as seções dos artigos. O Quadro 2 apresenta as características de *design* do corpus de estudo utilizado nesta pesquisa.

**Quadro 2** – Critérios de compilação do corpus de estudo.

<b>Registro*</b>	Acadêmico
<b>Gênero</b>	Artigos
<b>Meio de publicação</b>	Periódicos revisados por pares e de acesso aberto da plataforma <i>PLoS ONE</i>
<b>Língua de publicação</b>	Inglês
<b>Ano (ou período) de publicação</b>	de 2003 a 2019
<b>Domínio</b>	Ciências da saúde
<b>Seção do artigo</b>	Introdução

Fonte: Elaborado pelas autoras.

A Tabela 1 mostra o número de *tokens*, *types*, frases e textos, bem como a média de palavras por seção do corpus do estudo. Cabe mencionar que todos os textos do corpus de estudo tinham mais de 100 palavras, sendo 585 a média de palavras em cada texto.

**Tabela 1** – Corpus de estudo em números.

Domínio	Corpus de estudo	Número total de palavras ( <i>tokens</i> **)	Número total de palavras sem repetição ( <i>types</i> )	Frases	Textos	Média de palavras por seção
Ciências da saúde	Introdução	1.000.003	53.569	40.008	1.709	585

Fonte: Elaborada pelas autoras.

\*. Biber *et al.* (2007, p. 8) definem *registro* (quando diferenciado de *gênero*) como a linguagem associada a uma área do conhecimento ou a um domínio, como o registro jurídico ou o registro acadêmico. O termo *gênero*, por sua vez, quando contrastado com *registro*, é usado para referir-se a um tipo de mensagem com uma estrutura interna convencionalizada, como em um artigo da biologia ou um memorando de negócios.

\*\* . A ferramenta *Sketch Engine* denomina *token* a menor unidade existente em um *corpus*. Assim, diferentes formas das palavras e sinais de pontuação são contabilizados como *tokens* distintos. Palavras unidas por apóstrofo e hífen são contabilizadas separadamente.

Os critérios de compilação do corpus de referência foram semelhantes àqueles usados para compilar o corpus do estudo, exceto pelo fato de aquele conter artigos acadêmicos de diversos domínios (biologia e ciências da vida, ciências da computação e da informação, ciências da terra, ecologia e ciências do meio-ambiente, engenharia e tecnologia, física e ciências sociais), e não das áreas das ciências da saúde. Os artigos acadêmicos compilados foram revisados por pares e publicados entre 2003 e 2019 em revistas internacionais de acesso aberto. A Tabela 2 indica o número de *tokens*, *types*, frases e textos, bem como a média de palavras por seção, do corpus de referência. O corpus resultante da compilação descrita foi denominado *Corpus of Discipline and Section-Specific Academic English* (CODISAE).

**Tabela 2** – Corpus de referência em números.

Domínio	Corpus de referência	Número total de palavras ( <i>tokens</i> )	Número total de palavras sem repetição ( <i>types</i> )	Frases	Textos	Média de palavras por seção
Artigos de diversas disciplinas, mas não das ciências da saúde	<i>Introdução</i>	5.000.008	127.135	193.272	6.067	824

Fonte: Elaborada pelas autoras.

### 3.3. Critérios para extração e classificação dos KLBs

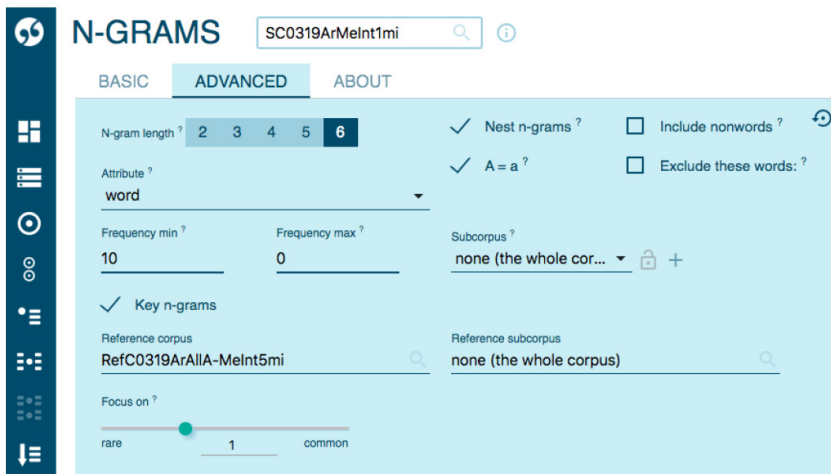
Como sugere Cortes (2013), pontos de corte diferentes foram estabelecidos para *KLBs* de 4, 5 e 6 palavras, de forma que o número de unidades extraídas de cada grupo (4, 5 e 6 palavras) fosse mais equilibrado, já que quanto maior a unidade, menor o número de suas repetições. Por exemplo, se fosse usado um ponto de corte de 40 ocorrências por milhão de palavras para *KLBs* de 6 palavras, apenas 3 unidades com essas características seriam extraídas, o que reduziria consideravelmente o volume de dados a serem observados.

Assim, três grupos de *KLBs* foram extraídos do corpus de estudo: sequências de quatro, cinco e seis palavras. Os critérios para a extração foram os seguintes: (i) as sequências de palavras deveriam ocorrer

pelo menos 40 vezes em um milhão de palavras para as sequências de 4 palavras, 20 vezes para as sequências de 5 palavras e 10 vezes para sequências de 6 palavras; e (ii) tais sequências de palavras deveriam constar em dez textos ou mais.

A ferramenta *Sketch Engine* (Kilgarriff *et al.*, 2004) foi usada para a extração dos *KLBs* do corpus de estudo. Feito o *upload* dos corpora de estudo e de referência na ferramenta, e considerando-se os critérios para identificação dos *KLBs* apresentados, foi realizada uma extração de forma a identificar os *KLBs* típicos da seção *Introdução* dos artigos da área das ciências da saúde em comparação com o corpus de referência usado neste estudo. A Figura 1 exemplifica os critérios para a identificação de *KLBs* na ferramenta *Sketch Engine* (Kilgarriff *et al.*, 2004) para as sequências de 6 palavras.

**Figura 1** – Critérios para a identificação de *KLBs* na ferramenta *Sketch Engine*



Fonte: Sketch Engine.

A classificação dos *KLBs* foi um processo manual e iniciou com a organização de todas as unidades extraídas do corpus de estudo em uma tabela (cf. Anexo 4). Com as unidades organizadas, iniciamos o processo de categorização, onde as unidades que tinham estrutura semelhante (*the \* of this study was to \**), alguma palavra lexical comum (*aim, study*) ou palavras lexicais que expressassem a mesma função

comunicativa (*investigate, evaluate*) foram marcadas com a mesma cor. A Tabela 3 mostra alguns exemplos de uma dessas categorias, onde exemplos de *KLBs* de 5 palavras foram sombreadas, indicando que pertenciam a um mesmo grupo.

**Tabela 3** – Frequências absoluta e normalizada (por milhão de palavras) dos *KLBs* de 5 palavras nos corpora de estudo e de referência

<i>KLB</i> (5 palavras)	Corpus de estudo		Corpus de referência		IC*
	Frequência absoluta	Frequência Normalizada	Frequência absoluta	Frequência Normalizada	
<i>by the world health organization</i>	25	19.9	18	2.9	5.4
<i>the present study aimed to</i>	20	15.0	17	2.7	4.6
<i>it has been reported that</i>	28	22.3	30	4.8	4.0
<i>study was to evaluate the</i>	23	18.3	24	3.8	4.0
<i>purpose of this study was</i>	34	27.0	43	6.8	3.6
<i>study was to investigate the</i>	26	20.7	32	5.1	3.6
<i>this study was to investigate</i>	23	18.3	29	4.6	3.4
<i>the aim of this study</i>	99	78.7	161	25.6	3.0
<i>the purpose of this study</i>	41	32.6	66	10.5	2.9
<i>aim of this study was</i>	85	67.6	145	23.1	2.8

\*IC, ou índice de chavicidade (*keyness score*), representa o quanto determinada unidade é mais frequente no corpus de estudo em relação ao corpus de referência, sendo, por isso, considerada típica (*key*) de determinada coleção de textos.

Fonte: Elaborado pelas autoras.

A identificação das funções comunicativas exercidas por vários dos *KLBs* nos movimentos retóricos se deu pela observação das unidades organizadas em categorias (cf. Anexo 5) e pela observação das unidades em seu contexto de uso, tendo-se como parâmetro o modelo apresentado por Kanoksilapatham (2005). Tal procedimento foi realizado por dois dos autores do presente trabalho. Nos casos de discrepância na categorização das unidades e na identificação da função comunicativa das mesmas, os contextos de ocorrência das unidades foram analisados novamente até que se chegasse a um consenso. Vale ressaltar, ainda, que o procedimento de observação e identificação das funções comunicativas dos *KLBs* foi realizado a partir dos dados revelados pelo corpus e não com base em categorias *a priori*.

## 4. Resultados e discussão

A extração dos *KLBs* conforme os critérios apresentados em 3.1 e os procedimentos descritos em 3.2 resultou em um total de 107 unidades (38 com 4 palavras, 34 com 5 palavras e 35 com 6 palavras). As cinco unidades com maior índice de chavicidade de cada grupo são mostradas na Tabela 4. A tabela com a totalidade das unidades encontra-se no Anexo 5.

**Tabela 4** – Cinco *KLBs* com maior índice de chavicidade com 4, 5 e 6 palavras da seção *Introdução*

<i>KLBs</i> de 4 palavras		<i>KLBs</i> de 5 palavras		<i>KLBs</i> de 6 palavras		
Frequência mínima: 40	IC	Frequência mínima: 20	IC	Frequência mínima: 10	IC	
1	<i>for the treatment of</i>	10.1	<i>by the world health organization</i>	5.4	<i>this study was to compare the</i>	4.2
2	<i>is the most common</i>	5.0	<i>the present study aimed to</i>	4.6	<i>the purpose of this study was</i>	3.8
3	<i>the world health organization</i>	4.0	<i>it has been reported that</i>	4.0	<i>of this study was to compare</i>	3.7
4	<i>has been associated with</i>	3.6	<i>study was to evaluate the</i>	4.0	<i>purpose of this study was to</i>	3.5
5	<i>to be associated with</i>	3.1	<i>purpose of this study was</i>	3.6	<i>this study was to evaluate the</i>	3.5

Fonte: Elaborada pelas autoras.

Na Tabela 4, o valor indicado na coluna IC representa o índice de chavicidade (*keyness score*) de cada *KLB*, ou seja, o quanto aquela unidade é mais frequente no corpus de estudo em relação ao corpus de referência, sendo, por isso, considerada típica (*key*) de determinada coleção de textos. Na ferramenta *Sketch Engine*, tal valor é calculado a partir da seguinte fórmula:

$$\frac{fpm_{rmfocus} + N}{fpm_{rmref} + N}$$

Em que  $fpm_{rmfocus}$  é a frequência normalizada (por milhão de palavras) da sequência de palavras no corpus de estudo,  $fpm_{rmref}$  é a frequência normalizada (por milhão de palavras) da sequência de palavras no corpus de referência, e  $N$  é o parâmetro de suavização ( $N = 1$  é o valor predefinido).

No caso do primeiro *KLB* de 6 palavras (*this study was to compare the*) da Tabela 4, por exemplo, a fórmula para o cálculo do índice de chavicidade (*keyness score*) é:  $(8.7 + 1) / (1.3 + 1) = 4.2$ . Assim, nessa tabela, os primeiros *KLBs* são as unidades mais prototípicas da seção *Introdução* do corpus de artigos da área das ciências da saúde, em relação ao corpus de referência usado no estudo.

A análise dos *KLBs* extraídos da seção *Introdução* levou às seguintes observações: (i) a totalidade dos *KLBs*, apesar de estatisticamente típicos dos artigos da área das ciências da saúde (conforme indicado pelo índice de chavicidade), também ocorre em artigos de outras áreas, como pode ser observado nas tabelas dos anexos 2, 3 e 4 (e, como indica a frequência absoluta dos *KLBs* no corpus de referência, nunca igual a zero); (ii) dos 107 *KLBs* de 4, 5 e 6 palavras extraídos do corpus de estudo, apenas quatro (*for the treatment of, the world health organization, by the world health organization e centers for disease control and prevention*) apresentam em sua composição elementos lexicais típicos da área das ciências da saúde (*treatment, health e disease*) (grifo nosso). Por hipótese, e considerando que os critérios de extração de *KLBs* buscam as unidades mais típicas do domínio em estudo, seria possível imaginar que haveria um maior número de *KLBs* do tipo (ii). Também poderíamos imaginar a existência de *KLBs* exclusivos do domínio, ou seja, com frequência absoluta igual a zero no corpus de referência. Entretanto, ainda que as palavras *treatment, health e disease* vinculem esses *KLBs* às áreas da saúde, os dados numéricos extraídos do corpus de estudo e do corpus de referência mostram que essas unidades também ocorrem em artigos de outras áreas do conhecimento, embora com frequência muito menor. Como pode ser observado na Figura 2, o *KLB for the treatment of*, por exemplo, tem uma frequência normalizada (por milhão de palavras) de 46.1 no corpus de estudo e de 3.7 no corpus de referência ou seja, ainda que ele ocorra com maior frequência no domínio em estudo, ele não é exclusivo do domínio.



**Figura 2** – A ocorrência dos *KLBS* de 4p com maior chavicidade nos corpora de estudo e de referência.

**N-GRAMS** SC0319ArMeInt1mi ⓘ

! 4-grams word (lowercase) (Items: NaN, total frequency: NaN)

Word	SC0319ArMeInt1mi		RefC0319ArAlIA-MeInt(2.0)5mi		Score
	Frequency	Relative freq.	Frequency	Relative freq.	
1 for the treatment of	58	46.1	23	3.7	10.1
2 is the most common	53	42.1	48	7.6	5
3 the world health organization	96	76.3	115	18.3	4
4 has been associated with	46	36.6	59	9.4	3.6

Fonte: Sketch Engine.

A partir das observações descritas, os *KLBS* extraídos neste estudo foram categorizados em dois grupos: (i) típicos da disciplina, área ou domínio, ou seja, aqueles que ocorrem com maior frequência em um corpus de estudo em relação a um corpus de referência, sendo, por isso, considerados próprios de determinada coleção de textos (por exemplo, *the present study aimed to, this study was to compare the*); e (ii) específicos de uma disciplina, área ou domínio, isto é, aqueles que, além da tipicidade descrita no item anterior, também apresentam em sua composição elementos lexicais que os vinculam ainda mais à área de especialidade à qual estão relacionados (por exemplo, *for the **treatment** of, the world **health** organization, by the world **health** organization e centers for **disease** control and prevention*).

#### 4.1. Classificação e distribuição dos *KLBS* conforme sua função comunicativa nos movimentos retóricos e passos da seção Introdução

O processo de classificação dos *KLBS* se deu em duas etapas. Inicialmente, as unidades extraídas com 4, 5 e 6 palavras foram organizadas na Tabela 3 (cf. tabela completa no Anexo 5). Em seguida, foram agrupadas e classificadas de acordo com sua função comunicativa nos movimentos retóricos (MR) e nos passos (P) da seção *Introdução*, tendo-se como parâmetro o modelo de Kanoksilapatham (2005), conforme mostra o Quadro 3 (cf. quadro completo no Anexo 1).

**Quadro 3** – Classificação dos *KLBs* de 4, 5 e 6 palavras extraídos da seção *Introdução*, conforme o modelo proposto por Kanoksilapatham (2005).

<b>Movimento retórico</b> (doravante MR) <b>Passo</b> (doravante P)	<b><i>KLB</i></b> <b>(4 palavras)</b>	<b><i>KLB</i></b> <b>(5 palavras)</b>	<b><i>KLB</i></b> <b>(6 palavras)</b>
<b>MR1: Apresentando a relevância do campo</b> <b>P1: Defendendo a centralidade do tópico</b>	<i>is the most common is one of the</i>	<i>one of the most common is one of the most</i>	<i>is one of the most common is one of the most important</i>
<b>MR1: Apresentando a relevância do campo</b> <b>P2: Fazendo generalizações</b>	<i>to the development of has been shown to</i>	<i>has been shown to be it has been shown that</i>	
<b>MR1: Apresentando a relevância do campo</b> <b>P3: Revisando pesquisas prévias</b>	<i>been shown to be have been shown to</i>	<i>it has been reported that it has been shown that</i>	
<b>MR2: Preparando para a presente pesquisa</b> <b>P1: Indicando lacunas</b>	<i>the best of our best of our knowledge</i>	<i>there is a lack of to the best of our</i>	<i>to the best of our knowledge</i>
<b>MR2: Preparando para a presente pesquisa</b> <b>P2: Levantando questões</b>			
<b>MR3: Introduzindo o presente estudo</b> <b>P1: Estabelecendo objetivo(s)</b>	<i>aim of this study of this study was</i>	<i>the present study aimed to study was to evaluate the</i>	<i>this study was to compare the the purpose of this study was</i>
<b>MR3: Introduzindo o presente estudo</b> <b>P2: Descrevendo procedimentos</b>	<i>in the present study</i>		
<b>MR3: Introduzindo o presente estudo</b> <b>P3: Apresentando resultados</b>	<i>in the present study</i>		

Fonte: Elaborado pelas autoras.

Os *KLBs* de 4, 5 e 6 palavras (4p, 5p e 6p) com o maior índice de chavicidade de cada movimento retórico e passo foram agrupados na Tabela 6, indicando-se qual *KLB* é o mais típico de cada movimento e passo. No Quadro 3, também foi inserido o número total de *KLBs* com a mesma função comunicativa em cada passo dos movimentos retóricos, mostrando-se sua distribuição ao longo da seção estudada.

**Quadro 4** – Distribuição de *KLBS* de 4, 5 e 6 palavras em cada movimento retórico e passo<sup>14</sup> da seção *Introdução* dos artigos estudados.

<b>Movimento retórico</b> (doravante, MR)	<b>Passo</b> (doravante, P)	<b><i>KLBS</i> com maior IC</b>	<b><i>KLBS</i> com a mesma função comunicativa</b>
<b>MR1: Apresentando a relevância do campo</b>	<b>P1: Defendendo a centralidade do tópico</b>	4p - <i>is the most common</i>	3
		5p - <i>is one of the most</i>	5
		6p - <i>is one of the most important</i>	2
	<b>P2: Fazendo generalizações</b>	4p - <i>to the development of</i>	6
		5p - <i>has been shown to be</i>	4
		6p - -	0
	<b>P3: Revisando pesquisas prévias</b>	4p - <i>have been shown to</i>	5
		5p - <i>it has been reported that</i>	4
		6p - -	0
<b>MR2: Preparando para a presente pesquisa</b>	<b>P1: Indicando lacunas</b>	4p - <i>the best of our</i>	6
		5p - <i>there is a lack of</i>	5
		6p - <i>to the best of our knowledge</i>	1
	<b>P2: Levantando questões</b>	4p - -	0
		5p - -	0
		6p - -	0
<b>MR3: Introduzindo o presente estudo</b>	<b>P1: Estabelecendo objetivos</b>	4p - <i>aim of this study</i>	13
		5p - <i>the present study aimed to</i>	18
		6p - <i>this study was to compare the</i>	30
	<b>P2: Descrevendo procedimentos</b>	4p - <i>in the present study</i>	1
		5p - -	0
		6p - -	0
	<b>P3: Apresentando resultados</b>	4p - <i>in the present study</i>	1
		5p - -	0
		6p - -	0

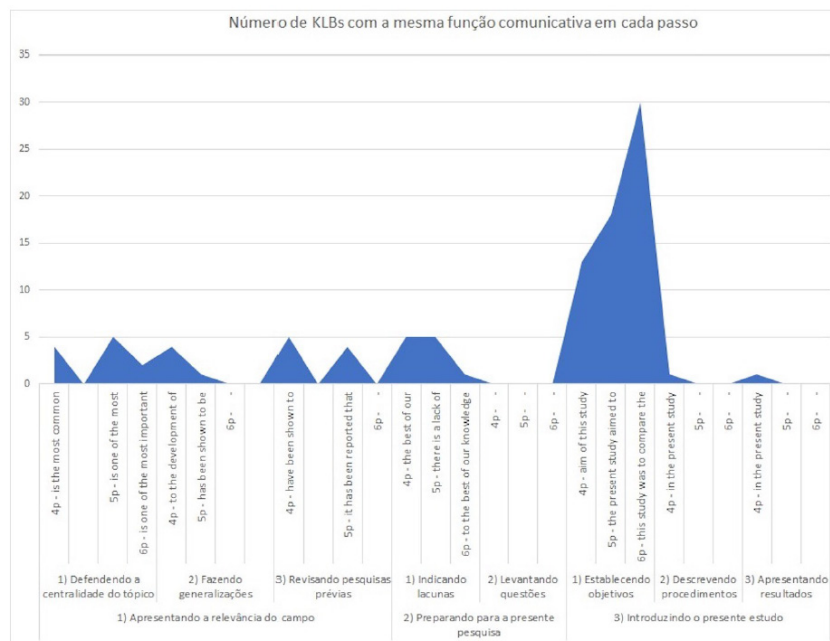
Fonte: Elaborado pelas autoras.

14. Os títulos dos movimentos retóricos e dos passos em português são traduções da tabela original em Kanoksilapatham (2005, p. 290).

A observação dos dados apresentados no Quadro 4 mostra bastante irregularidade na distribuição dos *KLBs* ao longo dos movimentos retóricos e dos passos da seção estudada. A maior incidência de *KLBs* de 4, 5 e 6 palavras acontece no movimento retórico 3 (introduzindo o presente estudo), passo 1 (estabelecendo objetivo[s]), no qual se encontram 61 dos 107 *KLBs* extraídos. Também chama a atenção o fato de que, no passo 1 (estabelecendo objetivo[s]), há um número bem maior de *KLBs* com 6 palavras (30) do que unidades com 4 e 5 palavras, que contam com 13 e 18 ocorrências, respectivamente. A incidência de *KLBs* nos movimentos 1 (apresentando a relevância do campo) e 2 (preparando para a presente pesquisa) diminui bastante, havendo uma média de 2,7 ocorrências por passo. Por fim, nos passos 2 (descrevendo os procedimentos), 3 (apresentando os resultados) do movimento retórico 3 (introduzindo o presente estudo) e no passo 2 (levantando questões) do movimento retórico 2 (preparando para a presente pesquisa), há uma incidência mínima (ou inexistência) de *KLBs*, sendo apenas duas unidades identificadas em nove passos. Chama a atenção o fato de que no passo 2 (levantando questões) do movimento retórico 2 (preparando para a presente pesquisa), não há ocorrências de *KLBs*, indicando a possível inexistência do referido passo. É interessante notar que os resultados obtidos por Kanoksilapatham (2005) também indicam que nem todos os artigos da área da bioquímica apresentam o movimento retórico 2 em sua estrutura (cf. quadro 1).

A observação do Gráfico 1 deixa mais evidente a irregularidade na distribuição dos *KLBs* ao longo dos movimentos retóricos e dos passos da seção estudada, bem como o grande volume de unidades com seis palavras usadas para apresentar os objetivos do estudo.

**Gráfico 1** – Distribuição dos *KLBS* nos movimentos retóricos e nos passos da seção *Introdução* dos artigos da área das ciências da saúde.



Fonte: Elaborado pelas autoras.

Podemos inferir que a irregularidade observada indica, por exemplo, que os passos com menor incidência de *KLBS* são menos frequentes nos artigos da área das ciências da saúde em comparação com os artigos da bioquímica (cf. Kanoksilapatham, 2005), ou que a linguagem dos referidos passos é menos formulaica, o que justificaria a menor incidência de *KLBS*, ou, ainda, que um determinado passo não ocorre na seção dos artigos de determinado domínio (cf. quadro 4).

Também merece reflexão o número bem maior de *KLBS* com seis palavras em um único passo da seção estudada, quando comparados ao número menor de unidades com 4 e 5 palavras. Cortes (2013) menciona como um dos principais achados em seu estudo, que relaciona *lexical bundles* e movimentos retóricos em introduções de artigos, o fato de terem sido identificados *lexical bundles* extremamente longos, como *such as the rest of the paper is organized as follows* ou *the purpose of the present study was to*, fenômeno que, segundo a autora, ainda não havia sido relatado em estudos do tipo (cf. Cortes, 2013).

## 4.2. *A construção de recursos pedagógicos a partir dos dados coletados*

Os resultados obtidos a partir da análise linguística orientada pelos princípios da Linguística de Corpus quando vinculados ao entendimento propiciado pelos estudos de gênero, em especial à análise de movimentos retóricos, dão margem a várias aplicações pedagógicas. Como afirma Cotos (2018, p. 6), “A crescente sinergia entre a análise de movimentos retóricos e outras tradições epistemologicamente diferentes abre caminhos para teorias robustas, metodologias de pesquisa mistas e aplicações tecnológicas e pedagógicas mais fundamentadas.”<sup>15</sup>

Um exemplo de aplicação derivada dessa sinergia é o *Manchester Phrasebank*<sup>16</sup> (cf. Morley, 2014, 2017), descrito por seu desenvolvedor como um recurso para aqueles que almejam produzir textos acadêmicos. Segundo Morley (2014), o *Manchester Phrasebank* tem o objetivo de “fornecer as ‘porcas e parafusos’ fraseológicos da escrita acadêmica organizada de acordo com as seções principais de um trabalho de pesquisa ou dissertação”<sup>17</sup> (Morley, 2014, p. 2). O recurso é um exemplo prático da utilização de dados linguísticos, no caso as expressões multipalavras, na construção de um recurso pedagógico usado para o ensino de gêneros acadêmicos. Entretanto, como ressalta seu criador, o *Manchester Phrasebank* não é específico de determinada disciplina (Morley, 2014), ou seja, não trata apenas de textos provenientes de uma área de especialidade. Tampouco é específico de determinado gênero, isto é, as expressões multipalavras encontradas nesse recurso são extraídas de um corpus composto por mais de um gênero acadêmico (teses e artigos de pesquisa). Por fim, o recurso não é *corpus-driven*, o que significa dizer que as expressões multipalavras disponibilizadas nesse material não foram exclusivamente extraídas de corpora.

Outra aplicação que exemplifica a relação entre dados linguísticos obtidos a partir de corpora e movimentos retóricos é o AWSuM - *Aca-*

15. No original: “The growing synergy between move analysis and other epistemologically different traditions is paving the way for enriched theories, blended research methodologies, and more grounded pedagogies and technological applications.” (Cotos, 2018, p. 6).

16. Disponível em: <<http://www.phrasebank.manchester.ac.uk>>. Acesso em: 29 nov. 2019.

17. No original: “to provide the phraseological ‘nuts and bolts’ of academic writing organised according to the main sections of a research paper or dissertation.” (Morley, 2014, p. 2).

*demio Word Suggestion Machine* (Mizumoto *et al.*, 2017). A ferramenta on-line de suporte à escrita de artigos acadêmicos foi desenvolvida utilizando-se o que os autores chamam de *Bundle–Move Connection Approach*, ou seja, a abordagem sugerida por Cortes (2013), segundo a qual os *lexical bundles* são relacionados aos movimentos retóricos onde ocorrem.

O recurso pedagógico proposto a partir dos dados desta pesquisa assemelha-se ao *Manchester Phrasebank* e ao *Academic Word Suggestion Machine* na medida em que propõe, como ponto de partida, a apresentação da estrutura retórica de gêneros estudados, que, como pontuam Biber *et al.* (2007, p. 40), são “especialmente importantes nos contextos educacionais e de treinamento para auxiliar os aprendizes a compreender e produzir textos de gênero com o qual não estão familiarizados”<sup>18</sup>.

Quanto às diferenças, dois aspectos distanciam o recurso proposto neste estudo do *Manchester Phrasebank*: (i) ser direcionado por corpus (*corpus-driven*); e (ii) ser específico das diferentes áreas do conhecimento (*discipline-specific*). Neste estudo, todas as informações linguísticas apresentadas são provenientes dos corpora compilados para o estudo, ou seja, os *KLBS* utilizados em nosso estudo e que irão futuramente informar a construção do recurso são típicos das disciplinas e das áreas do conhecimento-alvo.

Por fim, pretende-se que a informação linguística seja apresentada de forma econômica, amigável<sup>19</sup> e intuitiva para o usuário, para que, durante o processo de produção textual, os *KLBS* possam ser facilmente incorporados aos textos em produção. Como sugerem Simpson-Vlach e Ellis (2010, p. 510), “uma classificação linguística funcional, assim como a organização de construções de acordo com as necessidades e os propósitos acadêmicos, é essencial para transformar uma lista em algo que pode informar o currículo ou materiais desenvolvidos para testagem”<sup>20</sup>. Por essa razão, em nossa proposta de aplicação, os *KLBS*

18. No original: “*particularly valuable in educational and training contexts to help novices learn to understand and produce a genre that is new to them.*” (Biber *et al.*, 2007, p. 40).

19. O termo *amigável*, neste trabalho, diz respeito ao sistema ou à interface com o qual o usuário interage com facilidade.

20. No original: “*functional linguistic classification and the organization of constructions according to academic needs and purposes is essential in turning a list into something*

não são disponibilizados em forma de listas, como no *Manchester Phrasebank* e no *Academic Word Suggestion Machine*, mas sim a partir de perguntas-guia (por exemplo, *What was the aim of this study?*), que, ao serem respondidas pelo usuário, buscam os *KLBs* com maior chavacidade de cada movimento retórico e passo para comporem o texto-alvo.

Dois caminhos foram pensados para permitir a incorporação do dado linguístico obtido neste estudo à construção de um recurso que possa auxiliar o processo de escrita acadêmica realizado em sala de aula ou individualmente, sem supervisão e suporte do professor. O primeiro apresenta um padrão a partir do *KLB* com maior chavacidade extraído do movimento retórico-alvo, enquanto o segundo reúne todos os *KLBs* com mesma função comunicativa em um único *lexical frame*. Cada um desses caminhos é descrito nas seções que seguem.

#### 4.2.1. Padrão a partir do *KLB* com maior chavacidade

A observação das linhas de concordância dos *KLBs* com maior chavacidade de cada seção dos artigos estudados mostra certo padrão tanto para a esquerda quanto para a direita, como indica a Figura 3:

**Figura 3** – Linhas de concordância do *KLB aim of this study*.

26	doc#461	restioned by operational aspects related to patient flow. </> The	aim of this study	was therefore to assess the diagnostic accuracy of a rapid RT-PCR a
27	doc#479	ies that might reverse these alarming trends. </> Therefore, the	aim of this study	was to determine the prevalence of vitamin D deficiency among preg
28	doc#486	ical overrepresentation of rupture, such as women [24]. </> The	aim of this study	was to characterize the morphology of rAAAs in a population-based c
29	doc#493	nd pharmacological interventions in these women [14]. </> The	aims of this study	was to investigate plasma levels of sCD163 and sTWEAK in pregnan
30	doc#501	ffective and safe use of medications in clinical practice. </> The	aim of this study	was to examine the validity and reliability of the original MRCI to crea
31	doc#503	sed compared to that of the general population. </> Hence, the	aim of this study	was to assess risk factors for mortality, as well as to provide an overv
32	doc#535	icidualization and uNK cells are less investigated. </> Thus, the	aim of this study	was to investigate the effect of testosterone and flutamide on the emt
33	doc#577	IP3 protein was an additional aim in our research work. </> The	aim of this study	was to compare the expression levels of TNFAIP3 gene in blood sam
34	doc#582	XPJ treatments and early rehabilitation [13,14]. </> The primary	aim of this study	therefore was to compare patient-reported COPD pain intensity durin
35	doc#612	ollow-up in patients with positive endocervical margins. </> The	aim of this study	was to determine the demographic and clinicopathological predictors
36	doc#615	stic imaging, include haemoglobin, melanin and lipids. </> The	aim of this study	was to elucidate the intrinsic optoacoustic properties of DOX and eva
37	doc#623	never not received sufficient research attention. </> Hence, the	aim of this study	was to document women's self-narrated experiences of induced abor
38	doc#642	A for NGS from LBC samples has not been introduced. </> The	aims of this study	is to clarify a suitable method for extracting DNA to perform NGS from
39	doc#647	direction in interpreting ambiguous cues [18]. </> Thus, another	aim of this study	is exploring the longitudinal effect of angry rumination on hostile attrit
40	doc#658	*hypothesis in exercise-stimulated AVP production. </> [34]The	aim of this study	was therefore first to investigate the course of copeptin before, during
41	doc#668	l becoming the new gold standard in RVO diagnostics. </> The	aim of this study	was to compare the size of the FAZ in FA vs SCL and DCL in OCTA.
42	doc#679	working conditions on merchant vessels [1, 2]. </> The primary	aim of this study	was to compare stress and strain amongst a sample of merchant sea
43	doc#688	at germ cell development are different from mice. </> Thus, the	aim of this study	was to look at the expression of genes and proteins involved in the m

Fonte: Sketch Engine.

*that might usefully inform curriculum or language testing materials.*" (Simpson-Vlach & Ellis, 2010, p. 510).



No caso do *KLB aim of this study* percebe-se, à direita, a recorrência de elementos como *was/is* seguidos de *to* + verbo (*compare/investigate/determine*) e, à esquerda, *Thus/Hence/Therefore*, seguidos de *primary/main/first/second*.

Pensando no uso pedagógico dos dados obtidos, é possível imaginar a construção de um padrão a ser apresentado e praticado pelos alunos em sala de aula, ou disponibilizado on-line para uso daqueles interessados em produzir artigos acadêmicos sem o suporte de um professor. A partir da identificação do *KLB* com maior chavidade de cada passo dos movimentos retóricos, são observadas as linhas de concordância na busca de elementos recorrentes que permitam identificar um padrão linguístico à direita e à esquerda do *KLB*. A pergunta-guia serve para auxiliar na busca do conteúdo relativo ao estudo em desenvolvimento por aquele que está escrevendo o texto acadêmico. O Quadro 5 apresenta um exemplo de como a informação relativa ao *KLB aim of this study* pode ser organizada e disponibilizada para os usuários.

**Quadro 5** – Padrão linguístico a partir do *KLB* de 4 palavras com maior chavidade para o movimento retórico 3 (introduzindo o presente estudo), passo 1 (estabelecendo objetivo[s]).

Pergunta-guia	Padrão					Exemplo de resposta à pergunta-guia	
<i>What was the aim of this study?</i>	<i>Thus,</i>	<i>the</i>	(-)	<b>AIM OF THIS STUDY</b>	<i>was to</i>	<i>compare</i>	...the effect and mechanism of hyperandrogenemia and flutamide (a non-steroidal anti-androgen) on the embryo implantation and pregnancy during mid-pregnancy.
	<i>Hence,</i>		<i>primary</i>		<i>is to</i>	<i>investigate</i>	
	<i>Therefore,</i>					<i>determine</i>	
<b>Exemplo de possível texto resultante</b>							
<i>Thus, the primary AIM OF THIS STUDY was to investigate the effect and mechanism of hyperandrogenemia and flutamide (a non-steroidal anti-androgen) on the embryo implantation and pregnancy during mid-pregnancy.</i>							

Fonte: Elaborado pelas autoras.

#### 4.2.2. O agrupamento de todos os *KLBS* de um passo, com mesma função comunicativa, em um único *lexical frame*

Outra maneira de organizar os dados linguísticos obtidos no estudo de forma a torná-los mais amigáveis aos usuários é agrupar os *KLBS* de um mesmo passo que tenham a mesma função comunicativa. O Quadro 6 ilustra o referido agrupamento realizado com os *KLBS* da seção *Introdução*, movimento retórico 3 (introduzindo o presente estudo), passo 1 (estabelecendo objetivo[s]), a partir da tabela intitulada *Classificação dos KLBS* (vide Anexo 1).

**Quadro 6** – Agrupamento de *KLBS* com a mesma função comunicativa da seção *Introdução*, movimento retórico 3 (introduzindo o presente estudo), passo 1 (estabelecendo objetivo[s]).

	<i>purpose</i>	<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>		
<i>the</i>	<i>purpose</i>	<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>			
			<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>	<i>compare</i>	
		<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>	<i>compare</i>	
			<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>	<i>evaluate</i>	
<i>the</i>	<i>aim</i>	<i>of</i>	<i>our</i>	<i>study</i>	<i>was</i>			
		<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>	<i>investigate</i>	
			<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>	<i>investigate</i>	<i>the</i>
	<i>aim</i>	<i>of</i>	<i>our</i>	<i>study</i>	<i>was</i>	<i>to</i>		
<i>the</i>	<i>aim</i>	<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>			
		<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>	<i>evaluate</i>	
	<i>objective</i>	<i>of</i>	<i>the present</i>	<i>study</i>	<i>was</i>			
<i>the</i>	<i>objective</i>	<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>			
<i>the</i>	<i>aim</i>	<i>of</i>	<i>this</i>	<i>study</i>	<i>is</i>			
<i>the</i>	<i>aim</i>	<i>of</i>	<i>the present</i>	<i>study</i>				
<i>the</i>	<i>objective</i>	<i>of</i>	<i>the present</i>	<i>study</i>				
		<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>	<i>assess</i>	
			<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>	<i>determine</i>	<i>the</i>
	<i>objective</i>	<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>		
		<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>	<i>examine</i>	
		<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>	<i>determine</i>	
		<i>of</i>	<i>the present</i>	<i>study</i>	<i>was</i>	<i>to</i>		
	<i>aim</i>	<i>of</i>	<i>this</i>	<i>study</i>	<i>is</i>	<i>to</i>		
<i>the</i>	<i>objective</i>	<i>of</i>	<i>this</i>	<i>study</i>	<i>is</i>			
	<i>objective</i>	<i>of</i>	<i>this</i>	<i>study</i>	<i>is</i>	<i>to</i>		
	<i>aim</i>	<i>of</i>	<i>the present</i>	<i>study</i>	<i>was</i>			
	<i>aim</i>	<i>of</i>	<i>this</i>	<i>study</i>	<i>was</i>	<i>to</i>		

Fonte: Elaborado pelas autoras.

Nota-se que os elementos que constituem as unidades de certa forma ‘orbitam’ ao redor do substantivo *study*. O Quadro 7 mostra o *lexical frame* resultante do agrupamento de 27 *KLBs* de 6 palavras do movimento retórico 3 (introduzindo o presente estudo), passo 1 (estabelecendo objetivo[s]).

**Quadro 7** – *Lexical frame* resultante do agrupamento de 27 *KLBs* de 6 palavras do movimento retórico 3 (introduzindo o presente estudo), passo 1 (estabelecendo objetivo[s]).

Pergunta guia: <i>What was/is the aim of this study?</i>								
<i>the</i>	*	<i>of</i>	*	<i>study</i>	*	<i>to</i>	*	<i>the</i>
	<i>purpose</i>		<i>this</i>		<i>was</i>		<i>evaluate</i>	
	<i>aim</i>		<i>the present</i>		<i>is</i>		<i>compare</i>	
	<i>objective</i>		<i>our</i>				<i>investigate</i>	
							<i>assess</i>	
							<i>determine</i>	
							<i>examine</i>	

Fonte: Elaborado pelas autoras.

A organização dos *KLBs*, apresentada no quadro 6, remete à descrição feita por Biber (2009) quando ele afirma que “a maioria dos padrões na escrita acadêmica são estruturas formulaicas que consistem em palavras gramaticais invariáveis com um *slot* variável intercalado que é preenchido por palavras lexicais.”<sup>21</sup> (Biber, 2009, p. 275) e resulta no que Gray e Biber (2013) denominam *lexical frames*, ou seja, “seqüências descontínuas nas quais as palavras formam uma ‘estrutura’ em torno de um *slot* variável (por exemplo, / *don’t \* to, it is \* to*)”<sup>22</sup> (Gray & Biber, 2013, p. 109)<sup>23</sup>.

21. No original: “most patterns in academic writing are formulaic frames consisting of invariable function words with an intervening variable slot that is filled by content words.”

22. No original: “discontinuous sequences in which words form a ‘frame’ surrounding a variable slot (e.g. / *don’t \* to, it is \* to*)”.

23. Ainda que as definições de Biber (2009) e de Gray e Biber (2013) mencionem “um (grifo nosso) *slot* variável”, os autores indicam a possibilidade de múltiplos *slots* nos *lexical frames*.

Nota-se que o *lexical frame* identificado apresenta: (i) palavras gramaticais fixas (*the, of, to e the*), (ii) quatro *slots* variáveis, sendo dois preenchidos por palavras lexicais (*purpose, aim, objective e evaluate, compare, investigate*) e dois por palavras gramaticais (*this, the present, our e was, is*), e (iii) uma palavra lexical fixa (*study*) ao redor da qual os outros elementos do *frame* orbitam. Ambas as formas apresentadas nos Quadros 5 e 7 para disponibilizar os dados linguísticos de maneira mais amigável e compreensível para o usuário nos parecem válidas, ainda que derivem de trajetórias diferentes.

Em tempo, é importante lembrar que, como já mencionado por Cortes (2004) e Hyland (2008), é comum a existência de sobreposição de elementos lexicais em *lexical bundles* com tamanhos diferentes (por exemplo, de 4, 5 e 6 palavras), como é o caso em nosso trabalho. Como ressalta Grondona (2015), em alguns estudos, para evitar que os resultados sejam inflados por essa sobreposição, as unidades menores (3, 4 e 5 palavras) contidas nos *lexical bundles* mais longos (6 e 7 palavras, por exemplo) são excluídas.

Nesta pesquisa, quanto maior o bloco de discurso recorrente identificado no corpus, maior será sua relevância pedagógica, pois permitirá àquele que pretende produzir um texto acadêmico reconhecido como autêntico visualizar uma porção maior do discurso. Por essa razão, cumpridos todos os requisitos mínimos de frequência e dispersão, todos os *KLBs* extraídos em nosso estudo são mantidos, ainda que sejam unidades menores contidas em unidades maiores.

## 5. Conclusão

Os objetivos do estudo relatado neste artigo foram: (i) identificar *KLBs* na introdução de artigos acadêmicos publicados em inglês em periódicos internacionais da área das ciências da saúde, (ii) classificar os *KLBs* de acordo com sua função comunicativa nos movimentos retóricos da seção *Introdução* dos artigos; (iii) sugerir como os dados coletados podem informar a construção de recursos pedagógicos de livre acesso direcionados por corpus (*corpus-driven*) para auxiliar pesquisadores brasileiros na redação de artigos acadêmicos em inglês nas áreas selecionadas.

Quanto à classificação dos *KLBs* de acordo com sua função comunicativa nos movimentos retóricos da seção *Introdução* dos artigos estudados, chamou a atenção a acentuada irregularidade na distribuição das unidades ao longo dos movimentos retóricos e dos passos. Tais informações são pedagogicamente relevantes, pois indicam em quais passos dos movimentos retóricos se faz maior ou menor uso da linguagem formulaica.

Também chamou a atenção a grande ocorrência de *KLBs* com seis palavras usadas para apresentar os objetivos do estudo. *Lexical bundles* de seis ou mais palavras não costumam ser investigados no âmbito da Linguística de Corpus. A organização desses *KLBs* longos resultou na identificação de sequências descontínuas nas quais as palavras formam uma ‘estrutura’ em torno de *slots* variáveis.

Por fim, a reflexão sobre a forma como tais dados linguísticos podem ser utilizados pedagogicamente nos levou a considerar dois caminhos: enquanto o primeiro sugere um padrão a partir do *KLB* com maior chavidade do movimento retórico e do passo-alvo, o segundo propõe reunir todos os *KLBs* com mesma função comunicativa de um mesmo passo em um único *lexical frame*. Uma limitação do segundo caminho sugerido reside no fato de ele só poder ser efetivado quando houver vários *KLBs* com a mesma função comunicativa em um mesmo passo.

De uma forma geral, pode-se dizer que o conhecimento dos *KLBs* típicos de uma seção de um gênero em um domínio específico pode auxiliar na construção do texto especializado na medida em que: (i) permite o acesso a elementos formulaicos que se constituem como componentes básicos da construção do discurso, e (ii) propicia que o texto produzido seja reconhecido como autêntico pelos membros da comunidade discursiva da área em questão.

Espera-se que a utilização de recursos pedagógicos construídos com dados obtidos na presente investigação possa contribuir com o aprimoramento da formação de pesquisadores de alto nível para atuação em áreas estratégicas e com um consequente aumento no impacto dos artigos produzidos em inglês por brasileiros nas áreas eleitas. Pesquisas futuras buscarão ampliar, a outras áreas do conhecimento, outros gêneros acadêmicos e outras seções dos gêneros estudados, a descrição linguística e a construção de recursos pedagógicos propostos neste trabalho.

## Referências

- ALTBACH, Philip; REISBERG, Liz; RUMBLEY, Laura. 2019. *Trends in global higher education: Tracking an academic revolution*. Boston: BRILL.
- AMMON, Ulrich (ed.). 2001. *The dominance of English as a language of science: Effects on other languages and language communities* (Vol. 84). Nova Iorque: Walter de Gruyter.
- ANTHONY, Laurence. 2019. *AntCorGen* (Version 1.1.2) [Computer Software]. Tokyo, Japan: Waseda University. Disponível em: <https://www.laurenceanthony.net/software>. Acesso em: 25 out. 2019.
- BAUMVOL, Laura. 2018. *Language practices for knowledge production and dissemination: the case of Brazil* (Doctoral dissertation, Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil).
- BERBER SARDINHA, Tony. 2000. Lingüística de Corpus: histórico e problemática. *DELTA*, 16(2), 323-367. Disponível em: [http://www.scielo.br/scielo.php?script=sci\\_arttext&pid=S0102-44502000000200005&lng=en&nrm=iso](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005&lng=en&nrm=iso). Acesso em: 27 nov. 2019.
- \_\_\_\_\_. 2004. *Lingüística de corpus*. Barueri: Editora Manole Ltda.
- BIBER, Douglas. 2009. A corpus-driven approach to formulaic language in English. *International journal of corpus linguistics*, 14(3), 275-311.
- \_\_\_\_\_. 2012. Corpus-based and corpus-driven analyses of language variation and use. In Heine, Bernd; Narrog, Heiko. (Eds.). *The Oxford handbook of linguistic analysis*. Oxford: Oxford University Press.
- BIBER, Douglas; BARBIERI, Federica. 2007. Lexical bundles in university spoken and written registers. *English for specific purposes*, 26(3), 263-286.
- BIBER, Douglas; CONNOR, Ulla; UPTON, Thomas. 2007. *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdã: John Benjamins Publishing.
- BIBER, Douglas; CONRAD, Susan; CORTES, Viviana. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3), 371-405.
- BIBER, Douglas; CONRAD, Susan. 1999. Lexical bundles in conversation and academic prose. *Language and Computers*, 26, 181-190.
- BIBER, Douglas; JOHANSSON, Stig; LEECH, Geoffrey; CONRAD, Susan & FINEGAN, Edward. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson.

- BLESSINGER, Patrick; ANCHAN, John (eds.). 2015. *Democratizing higher education: International comparative perspectives*. Nova Iorque: Routledge.
- CORTES, Viviana. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for specific purposes*, 23(4), 397-423.
- \_\_\_\_\_. 2013. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for academic purposes*, 12(1), 33-43.
- COTOS, Elena. 2018. Move analysis. In: Chapelle, Carol (ed.). *The Encyclopedia of Applied Linguistics* (pp. 1-8). Hoboken, NJ: John Wiley & Sons.
- CRYSTAL, David. 2003. *English As a Global Language*. Cambridge, UK: Cambridge University Press.
- GRAY, Bethany. 2016. Lexical bundles. *Triangulating methodological approaches in corpus linguistic research*. London: Routledge, pp. 33-56.
- GRAY, Bethany; BIBER, Douglas. 2013. Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, 18(1), 109-136.
- GRONDONA, Carolina. 2015. *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, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil.
- HYLAND, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, 27(1), 4-21.
- \_\_\_\_\_. 2012. *Disciplinary identities: Individuality and community in academic discourse*. Munique: Ernst Klett Sprachen.
- JENKINS, Jennifer. 2009. English as a lingua franca: Interpretations and attitudes. *World Englishes*, 28(2), 200-207.
- KANOKSILAPATHAM, Budsaba. 2005. Rhetorical structure of biochemistry research articles. *English for specific purposes*, 24(3), 269-292.
- KILGARRIFF, Adam; RYCHLY, Pavel; SMRZ, Pavel; TUGWELL, David. 2004. The Sketch Engine. In: Williams, Geoffrey; Vessier, Sandra (eds.). *Proceedings of the 11th EURALEX International Congress* (pp. 105-116), Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.
- MIZUMOTO, Atsushi; HAMATANI, Sawako; IMAO, Yasuhiro. 2017. Applying the bundle–move connection approach to the development

- of an online writing support tool for research articles. *Language Learning*, 67(4), 885-921.
- MOK, Ka Ho. 2015. Higher education transformations for global competitiveness: Policy responses, social consequences and impact on the academic profession in Asia. *Higher Education Policy*, 28(1), 1-15.
- MORLEY, John. 2014. *Academic phrasebank*. Manchester: University of Manchester. [on-line]. Disponível em: <http://www.phrasebank.manchester.ac.uk/> [Acesso em 19 set. 2019].
- \_\_\_\_\_. 2017. *The academic phrasebank: An academic writing resource for students and researchers*. Manchester: University of Manchester.
- NWOGU, Kevin N. 1997. The medical research paper: Structure and functions. *English for specific purposes*, 16(2), 119-138.
- PAQUOT, Magali; GRANGER, Sylviane. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130-149.
- RAYSON, Paul. 2012. Corpus analysis of key words. In: Carol A. Chapelle (ed.). *The encyclopedia of applied linguistics*. Hoboken: Blackwell Publishing, pp. 320-326.
- ROBSON, Sue; WIHLBORG, Monne. 2019. Internationalisation of higher education: Impacts, challenges and future possibilities. *European Educational Research Journal*, 18(2), 127-134.
- SIMPSON-VLACH, Rita; ELLIS, Nick. 2010. An academic formulas list: New methods in phraseology research. *Applied linguistics*, 31(4), 487-512.
- SINCLAIR, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- STAPLES, Shelley; EGBERT, Jesse; BIBER, Douglas. 2013. Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for academic purposes*, 12(3), 214-225.
- SWALES, John. 1981. *Aspects of article introductions*. Birmingham: University of Aston.
- \_\_\_\_\_. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- \_\_\_\_\_. 2004. *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.
- TARDY, Christine. 2004. The role of English in scientific communication: lingua franca or Tyrannosaurus rex? *Journal of English for academic purposes*, 3(3), 247-269.

Recebido em: 21/12/2019  
Aprovado em: 09/09/2020



## ANEXOS

**Anexo 1** – Classificação dos *KLBs* de 4, 5 e 6 palavras extraídos da seção *Introdução*, conforme o modelo proposto por Kanoksilapatham (2005).

Movimento Retórico (doravante, MR) Passo (doravante, P)	<i>KLB</i> (4 palavras)	<i>KLB</i> (5 palavras)	<i>KLB</i> (6 palavras)
MR1: Apresentando a relevância do campo P1: Defendendo a centralidade do tópico	<i>is the most common is one of the one of the most</i>	<i>one of the most common is one of the most plays an important role in one of the most important an important role in the</i>	<i>is one of the most common is one of the most important</i>
MR1: Apresentando a relevância do campo P2: Fazendo generalizações	<i>to the development of has been shown to have been shown to it has been shown are more likely to on the other hand</i>	<i>has been shown to be it has been shown that have been shown to be it has been suggested that</i>	
MR1: Apresentando a relevância do campo P3: Revisando pesquisas prévias	<i>been shown to be have been shown to it has been shown studies have shown that in addition to the</i>	<i>it has been reported that it has been shown that have been shown to be it has been suggested that</i>	
MR2: Preparando para a presente pesquisa P1: Indicando lacunas	<i>the best of our best of our knowledge to the best of in the absence of it is necessary to little is known about</i>	<i>there is a lack of to the best of our little is known about the the best of our knowledge a better understanding of the</i>	<i>to the best of our knowledge</i>
MR2: Preparando para a presente pesquisa P2: Levantando questões			
MR3: Introduzindo o presente estudo P1: Estabelecendo objetivo(s)	<i>aim of this study of this study was purpose of this study the aim of this this study was to the purpose of this the aim of the the objective of this objective of this study the present study was this study is to of this study is of the present study</i>	<i>the present study aimed to study was to evaluate the purpose of this study was study was to investigate the this study was to investigate the aim of this study the purpose of this study aim of this study was of our study was to of this study was to objective of this study was the objective of this study of the present study was aim of this study is the aim of the present aim of the present study the present study was to of this study is to</i>	<i>this study was to compare the the purpose of this study was of this study was to compare purpose of this study was to this study was to evaluate the the aim of our study was of this study was to investigate this study was to investigate the aim of our study was to the aim of this study was of this study was to evaluate aim of the present study was aim of this study was to objective of the present study was the objective of this study was the aim of this study is the aim of the present study the objective of the present study of this study was to assess this study was to determine the objective of this study was to of this study was to examine of this study was to determine of the present study was to aim of this study is to the objective of this study is objective of this study is to goal of this study was to objectives of this study were to the objectives of this study were</i>
MR3: Introduzindo o presente estudo P2: Descrevendo procedimentos	<i>in the present study</i>		
MR3: Introduzindo o presente estudo P3: Apresentando resultados	<i>in the present study</i>		

## Anexo 2 – Frequências absoluta e normalizada (por milhão de palavras) dos *KL*Bs de 4 palavras nos corpora de estudo e de referência.

<i>KL</i> B (4 palavras)	Corpus de estudo		Corpus de referência		IC
	Frequência absoluta	Frequência Normalizada	Frequência absoluta	Frequência Normalizada	
<i>for the treatment of</i>	58	46.1	23	3.7	10.1
<i>is the most common</i>	53	42.1	48	7.6	5.0
<i>the world health organization</i>	96	76.3	115	18.3	4.0
<i>has been associated with</i>	46	36.6	59	9.4	3.6
<i>to be associated with</i>	51	40.5	78	12.4	3.1
<i>aim of this study</i>	114	90.6	199	31.7	2.8
<i>of this study was</i>	176	139.9	323	51.4	2.7
<i>purpose of this study</i>	43	34.2	76	12.1	2.7
<i>the aim of this</i>	131	104.1	241	38.4	2.7
<i>this study was to</i>	166	132.0	313	49.8	2.6
<i>the purpose of this</i>	51	40.5	108	17.2	2.3
<i>the aim of the</i>	49	38.9	105	16.7	2.3
<i>the objective of this</i>	60	47.7	131	20.9	2.2
<i>objective of this study</i>	60	47.7	137	21.8	2.1
<i>to the development of</i>	42	33.4	102	16.2	2.0
<i>the present study was</i>	45	35.8	112	17.8	2.0
<i>has been shown to</i>	106	84.3	281	44.7	1.9
<i>in the united states</i>	118	93.8	336	53.5	1.7
<i>is one of the</i>	204	162.2	646	102.8	1.6
<i>been shown to be</i>	61	48.5	201	32.0	1.5
<i>this study is to</i>	47	37.4	156	24.8	1.5
<i>the best of our</i>	45	35.8	151	24.0	1.5
<i>best of our knowledge</i>	45	35.8	151	24.0	1.5
<i>in the present study</i>	82	65.2	279	44.4	1.5
<i>of this study is</i>	50	39.7	170	27.1	1.5
<i>to the best of</i>	46	36.6	157	25.0	1.4
<i>of the present study</i>	52	41.3	184	29.3	1.4
<i>in the absence of</i>	55	43.7	202	32.2	1.3
<i>have been shown to</i>	74	58.8	277	44.1	1.3
<i>it has been shown</i>	41	32.6	160	25.5	1.3
<i>studies have shown that</i>	70	55.6	275	43.8	1.3
<i>it is necessary to</i>	45	35.8	184	29.3	1.2
<i>little is known about</i>	46	36.6	200	31.8	1.1
<i>one of the most</i>	178	141.5	777	123.7	1.1
<i>in addition to the</i>	43	34.2	189	30.1	1.1
<i>it is important to</i>	74	58.8	345	54.9	1.1
<i>are more likely to</i>	47	37.4	223	35.5	1.1
<i>on the other hand</i>	128	101.7	614	97.7	1.0

Fonte: Elaborado pelas autoras.

### Anexo 3 – Frequências absoluta e normalizada (por milhão de palavras) dos *KLBs* de 5 palavras nos corpora de estudo e de referência.

<i>KLB</i> (5 palavras)	Corpus de estudo		Corpus de referência		IC
	Frequência absoluta	Frequência Normalizada	Frequência absoluta	Frequência Normalizada	
<i>by the world health organization</i>	25	19.9	18	2.9	5.4
<i>the present study aimed to</i>	20	15.0	17	2.7	4.6
<i>it has been reported that</i>	28	22.3	30	4.8	4.0
<i>study was to evaluate the</i>	23	18.3	24	3.8	4.0
<i>purpose of this study was</i>	34	27.0	43	6.8	3.6
<i>study was to investigate the</i>	26	20.7	32	5.1	3.6
<i>this study was to investigate</i>	23	18.3	29	4.6	3.4
<i>the aim of this study</i>	99	78.7	161	25.6	3.0
<i>the purpose of this study</i>	41	32.6	66	10.5	2.9
<i>aim of this study was</i>	85	67.6	145	23.1	2.8
<i>of our study was to</i>	22	17.5	37	5.9	2.7
<i>of this study was to</i>	162	128.8	304	48.4	2.6
<i>one of the most common</i>	29	23.1	58	9.2	2.4
<i>objective of this study was</i>	43	34.2	88	14.0	2.3
<i>the objective of this study</i>	44	35.0	91	14.5	2.3
<i>of the present study was</i>	39	31.0	83	13.2	2.3
<i>aim of this study is</i>	23	18.3	48	7.6	2.2
<i>the aim of the present</i>	31	24.6	66	10.5	2.2
<i>aim of the present study</i>	28	22.3	60	9.6	2.2
<i>the present study was to</i>	37	29.4	81	12.9	2.2
<i>there is a lack of</i>	20	15.9	43	6.8	2.2
<i>has been shown to be</i>	36	28.6	99	15.8	1.8
<i>is one of the most</i>	99	78.7	281	44.7	1.7
<i>of this study is to</i>	45	35.8	148	23.6	1.5
<i>to the best of our</i>	45	35.8	150	23.9	1.5
<i>little is known about the</i>	31	24.6	103	16.4	1.5
<i>the best of our knowledge</i>	45	35.8	151	24.0	1.5
<i>plays an important role in</i>	25	19.9	98	15.6	1.3
<i>it has been shown that</i>	33	26.2	132	21.0	1.2
<i>have been shown to be</i>	21	16.7	84	13.4	1.2
<i>a better understanding of the</i>	24	19.1	103	16.4	1.2
<i>it has been suggested that</i>	22	17.5	102	16.2	1.1
<i>one of the most important</i>	36	28.6	170	27.1	1.1
<i>an important role in the</i>	21	16.7	103	16.4	1.0

Fonte: Elaborado pelas autoras.

#### Anexo 4 – Frequências absoluta e normalizada (por milhão de palavras) dos *KLBs* de 6 palavras nos corpora de estudo e de referência.

<i>KLB</i> (6 palavras)	Corpus de estudo		Corpus de referência		IC
	Frequência Absoluta	Frequência Normalizada	Frequência Absoluta	Frequência Normalizada	
<i>this study was to compare the</i>	11	8.7	8	1.3	4.2
<i>the purpose of this study was</i>	32	25.4	38	6.0	3.8
<i>of this study was to compare</i>	13	10.3	13	2.1	3.7
<i>purpose of this study was to</i>	32	25.4	41	6.5	3.5
<i>this study was to evaluate the</i>	13	10.3	14	2.2	3.5
<i>the aim of our study was</i>	10	7.9	10	1.6	3.5
<i>of this study was to investigate</i>	23	18.3	29	4.6	3.4
<i>is one of the most common</i>	25	19.9	32	5.1	3.4
<i>this study was to investigate the</i>	16	12.7	20	3.2	3.3
<i>aim of our study was to</i>	11	8.7	13	2.1	3.2
<i>the aim of this study was</i>	77	61.2	118	18.8	3.1
<i>of this study was to evaluate</i>	18	14.3	25	4.0	3.1
<i>associated with an increased risk of</i>	11	8.7	14	2.2	3.0
<i>aim of the present study was</i>	24	19.1	37	5.9	2.9
<i>aim of this study was to</i>	79	62.8	136	21.6	2.8
<i>centers for disease control and prevention</i>	12	9.5	18	2.9	2.7
<i>objective of the present study was</i>	10	7.9	16	2.5	2.5
<i>the objective of this study was</i>	32	25.4	61	9.7	2.5
<i>the aim of this study is</i>	20	15.9	37	5.9	2.5
<i>the aim of the present study</i>	27	21.5	52	8.3	2.4
<i>the objective of the present study</i>	10	7.9	17	2.7	2.4
<i>of this study was to assess</i>	11	8.7	20	3.2	2.3
<i>this study was to determine the</i>	10	7.9	18	2.9	2.3
<i>objective of this study was to</i>	39	31.0	83	13.2	2.3
<i>of this study was to examine</i>	10	7.9	19	3.0	2.2
<i>of this study was to determine</i>	14	11.1	28	4.5	2.2
<i>of the present study was to</i>	36	28.6	79	12.6	2.2
<i>aim of this study is to</i>	20	15.9	48	7.6	2.0
<i>the objective of this study is</i>	10	7.9	27	4.3	1.7
<i>objective of this study is to</i>	15	11.9	43	6.8	1.6
<i>is one of the most important</i>	20	15.9	65	10.3	1.5
<i>to the best of our knowledge</i>	45	35.8	150	23.9	1.5
<i>goal of this study was to</i>	10	7.9	38	6.0	1.3
<i>objectives of this study were to</i>	13	10.3	51	8.1	1.2
<i>the objectives of this study were</i>	11	8.7	54	8.6	1.0

Fonte: Elaborado pelas autoras.

**Anexo 5** – Resultado da extração de *KLBs* com 4, 5 e 6 palavras da seção *Introdução*, realizada com base nos critérios apresentados em 3.1 e nos procedimentos descritos em 3.2.

<i>KLBs</i> de 4 palavras		<i>KLBs</i> de 5 palavras		<i>KLBs</i> de 6 palavras		
Frequência mínima: 40	IC	Frequência mínima: 20	IC	Frequência mínima: 10	IC	
1	<i>for the treatment of</i>	10.1	<i>by the world health organization</i>	5.4	<i>this study was to compare the</i>	4.2
2	<i>is the most common</i>	5.0	<i>the present study aimed to</i>	4.6	<i>the purpose of this study was</i>	3.8
3	<i>the world health organization</i>	4.0	<i>it has been reported that</i>	4.0	<i>of this study was to compare</i>	3.7
4	<i>has been associated with</i>	3.6	<i>study was to evaluate the</i>	4.0	<i>purpose of this study was to</i>	3.5
5	<i>to be associated with</i>	3.1	<i>purpose of this study was</i>	3.6	<i>this study was to evaluate the</i>	3.5
6	<i>aim of this study</i>	2.8	<i>study was to investigate the</i>	3.6	<i>the aim of our study was</i>	3.5
7	<i>of this study was</i>	2.7	<i>this study was to investigate</i>	3.4	<i>of this study was to investigate</i>	3.4
8	<i>purpose of this study</i>	2.7	<i>the aim of this study</i>	3.0	<i>is one of the most common</i>	3.4
9	<i>the aim of this</i>	2.7	<i>the purpose of this study</i>	2.9	<i>this study was to investigate the</i>	3.3
10	<i>this study was to</i>	2.6	<i>aim of this study was</i>	2.8	<i>aim of our study was to</i>	3.2
11	<i>the purpose of this</i>	2.3	<i>of our study was to</i>	2.7	<i>the aim of this study was</i>	3.1
12	<i>the aim of the</i>	2.3	<i>of this study was to</i>	2.6	<i>of this study was to evaluate</i>	3.1
13	<i>the objective of this</i>	2.2	<i>one of the most common</i>	2.4	<i>associated with an increased risk of</i>	3.0
14	<i>objective of this study</i>	2.1	<i>objective of this study was</i>	2.3	<i>aim of the present study was</i>	2.9
15	<i>to the development of</i>	2.0	<i>the objective of this study</i>	2.3	<i>aim of this study was to</i>	2.8
16	<i>the present study was</i>	2.0	<i>of the present study was</i>	2.3	<i>centers for disease control and prevention</i>	2.7
17	<i>has been shown to</i>	1.9	<i>aim of this study is</i>	2.2	<i>objective of the present study was</i>	2.5
18	<i>in the united states</i>	1.7	<i>the aim of the present</i>	2.2	<i>the objective of this study was</i>	2.5
19	<i>is one of the</i>	1.6	<i>aim of the present study</i>	2.2	<i>the aim of this study is</i>	2.5
20	<i>been shown to be</i>	1.5	<i>the present study was to</i>	2.2	<i>the aim of the present study</i>	2.4
21	<i>this study is to</i>	1.5	<i>there is a lack of</i>	2.2	<i>the objective of the present study</i>	2.4
22	<i>the best of our</i>	1.5	<i>has been shown to be</i>	1.8	<i>of this study was to assess</i>	2.3
23	<i>best of our knowledge</i>	1.5	<i>is one of the most</i>	1.7	<i>this study was to determine the</i>	2.3
24	<i>in the present study</i>	1.5	<i>of this study is to</i>	1.5	<i>objective of this study was to</i>	2.3
25	<i>of this study is</i>	1.5	<i>to the best of our</i>	1.5	<i>of this study was to examine</i>	2.2
26	<i>to the best of</i>	1.4	<i>little is known about the</i>	1.5	<i>of this study was to determine</i>	2.2
27	<i>of the present study</i>	1.4	<i>the best of our knowledge</i>	1.5	<i>of the present study was to</i>	2.2
28	<i>in the absence of</i>	1.3	<i>plays an important role in</i>	1.3	<i>aim of this study is to</i>	2.0
29	<i>have been shown to</i>	1.3	<i>it has been shown that</i>	1.2	<i>the objective of this study is</i>	1.7
30	<i>it has been shown</i>	1.3	<i>have been shown to be</i>	1.2	<i>objective of this study is to</i>	1.6
31	<i>studies have shown that</i>	1.3	<i>a better understanding of the</i>	1.2	<i>is one of the most important</i>	1.5
32	<i>it is necessary to</i>	1.2	<i>it has been suggested that</i>	1.1	<i>to the best of our knowledge</i>	1.5
33	<i>little is known about</i>	1.1	<i>one of the most important</i>	1.1	<i>goal of this study was to</i>	1.3
34	<i>one of the most</i>	1.1	<i>an important role in the</i>	1.0	<i>objectives of this study were to</i>	1.2
35	<i>in addition to the</i>	1.1			<i>the objectives of this study were</i>	1.0
36	<i>it is important to</i>	1.1				
37	<i>are more likely to</i>	1.1				
38	<i>on the other hand</i>	1.0				

Fonte: Elaborado pelas autoras.