UNIVERSIDADE FEDERAL DE MINAS GERAIS

FACULDADE DE LETRAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

CAROLINA BOHÓRQUEZ GRONDONA

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

BELO HORIZONTE
FACULDADE DE LETRAS DA UFMG
2015

Carolina Bohórquez Grondona

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 apresentada ao Programa de Pós-Graduação em Estudos Linguísticos da Faculdade de Letras da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Linguística Aplicada.

Área de Concentração: Linguística Aplicada
Linha de Pesquisa: Ensino/Aprendizagem de Línguas Estrangeiras

Orientadora: Profa. Dra. Deise Prina Dutra

Co-orientador: Prof. Dr. Crysttian Arantes Paixão

Belo Horizonte
Faculdade de Letras da UFMG
2015

Para meus pais.

AGRADECIMENTOS

Em primeiro lugar, agradeço aos meus pais, cuja parceria exemplar e dedicação amorosa a sua família proveram a mim e a minhas irmãs todos os instrumentos necessários para que pudéssemos realizar os nossos sonhos. Agradeço pela educação libertadora que nos permitiu cultivar o espírito emancipador e seguir qualquer que fosse o caminho escolhido.

À professora Deise Prina Dutra, que por meio de sua prática como professora na graduação inspirou-me a continuar a minha formação e que diariamente me ensina o que é ser uma pesquisadora comprometida. Agradeço pela orientação cuidadosa para a realização deste trabalho, pela confiança, incentivo e carinho.

Ao professor Crysttian Arantes Paixão, por propiciar que o universo da estatística e da programação se aproximassem do meio do qual faço parte como linguista e por permitir que eu pudesse enxergar a importância do diálogo entre essas áreas. Agradeço pela disponibilidade e orientação criteriosa.

Ao meu grupo de pesquisa, pelas discussões enriquecedoras e espírito colaborativo.

Ao Rodrigo Araújo, pelas tardes de estudo de R.

Aos meus colegas de pós-graduação e amigos, pelo apoio extremamente necessário.

Ao Bruno, por fazer-se presente na caminhada para a realização dos meus sonhos e por acreditar em mim.

À minha sobrinha Manuela, pela renovação constante da alegria em nossa família.

Às minhas irmãs, pelo amor e incentivo.

À FAPEMIG pelo apoio financeiro.

RESUMO

O objetivo geral deste trabalho foi de investigar a correlação entre o uso de pacotes lexicais e o nível de proficiência linguística na escrita acadêmica de inglês, utilizando dois *subcorpora* do *International Corpus of Learner of English Version 2* (ICLEv2), a saber, parte do *subcorpus* de aprendizes de inglês de língua materna chinesa – Ch-ICLE – e o *subcorpus* de aprendizes de inglês de língua materna holandesa – Dt-ICLE. O ICLE é composto, em sua maior parte, por redações argumentativas de inglês de estudantes universitários. Os *subcorpora* escolhidos são representantes dos níveis de proficiência menor e maior, respectivamente, neste trabalho, com base nos índices do *Common European Framework of Reference for Languages* (CEFR). Uma vez que há resultados divergentes na literatura quanto à correlação investigada, o presente estudo busca checar se a eliminação de *pacotes lexicais relacionados ao tópico* e de *pacotes lexicais em contexto de sobreposição* pode ser um fator que influencie a percepção da correlação entre nível de proficiência e uso pacotes lexicais na escrita acadêmica de inglês. Para alcançar os objetivos descritos, esta pesquisa propõe uma metodologia automatizada para a eliminação dos tipos de pacotes lexicais supracitados, desenvolvida na linguagem R, e compara o número de pacotes lexicais produzidos por cada *subcorpora* antes e depois das eliminações. O teste estatístico qui-quadrado foi utilizado para atestar as diferenças encontradas. Os resultados evidenciam que houve, na maioria das vezes, uma porcentagem de diminuição maior de pacotes lexicais dos tipos que deveriam ser eliminados no *corpus* dos aprendizes de inglês menos proficientes, o Ch-ICLE, indicando que uma correlação entre menor produção de pacotes lexicais e menor nível de proficiência pode ser produtiva. Não foi possível, porém, comprovar que a eliminação de *pacotes lexicais relacionados ao tópico* e *de pacotes lexicais em contexto de sobreposição* é um fator definitivo para a correlação investigada, uma vez que a metodologia automatizada não foi capaz de eliminar esses pacotes em sua totalidade.

Palavras-chave: Eliminação automatizada de pacotes lexicais. *Pacotes lexicais relacionados ao tópico*. *Pacotes lexicais em contexto de sobreposição*. Escrita acadêmica. Linguística de Corpus.

ABSTRACT

The main goal of this research was to investigate the correlation between the use of lexical bundles and the level of proficiency of learners of English when writing academic texts. Two subcorpora that compose the International Corpus of Learner of English Version 2 (ICLEv2), which is mainly built of argumentative essays, were examined. Part of the Chinese subcorpus – Ch-ICLE – and the Dutch subcorpus – Dt-ICLE – were chosen for examination. These subcorpora respectively represent the lowest and the highest levels of proficiency in this study based on the Common European Framework of Reference for Languages (CEFR). Researchers have found opposing results concerning lexical bundles depending on whether they were eliminated on the lines of being related to a topic, or as part of a longer unit, or maintained. Therefore, this study investigates if the elimination of *topic-related lexical bundles* and *overlapping lexical bundles* has a significant influence on the correlation between the level of proficiency of learners of English and their reliance on lexical bundles when writing academic texts. In order to achieve the objectives of the research, this study proposes an automated methodology developed in the R language to eliminate the types of lexical bundles mentioned. Afterwards, the number of lexical bundles produced by each of the subcorpora prior to and after the application of the methodology was contrasted. The chi-square test was used to attest the differences found. Results show that there was, most of the times, a higher percentage of elimination of *topic-related lexical bundles* and *overlapping lexical bundles* within the least proficient group, Ch-ICLE, indicating that a correlation between lower proficiency level and a lesser level of lexical bundles use may be productive. It was not possible however to attest that the elimination of *topic-related lexical bundles* and *overlapping lexical bundles* is a definite factor for the investigated correlation, since the automated methodology did not eliminated those types of bundles in its totality.

KayWords: Automated elimination of lexical bundles. *Topic-related lexical bundles*. *Overlapping lexical bundles*. Academic writing. Corpus Linguistics.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo, é apresentada, primeiramente, a motivação para a elaboração desta pesquisa. Em seguida, na seção intitulada justificativa, delimita-se o objeto de estudo e sua relevância para as pesquisas da área da Linguística Aplicada. Nessa mesma seção, as definições de termos e referências a estudos complementares a este trabalho, necessários para a melhor compreensão da pesquisa, são mencionados. Por fim, são apresentados os objetivos do estudo a serem alcançados, as perguntas de pesquisa, e uma breve apresentação da estrutura da dissertação.

1.1 Motivação

Na vida estudantil universitária destacam-se as competências linguísticas da escrita acadêmica[1]. Esse tipo de escrita vem sendo investigado extensamente, sob diversas perspectivas, permitindo que pesquisadores possam descrevê-lo e, dessa maneira, afluir para a optimização de seu ensino. A Linguística de Corpus (LC) tem contribuído robustamente para esse fim, possibilitando que padrões da escrita acadêmica sejam melhor extraídos e examinados.

Um dos olhares investigativos da LC para a escrita acadêmica foca-se em seu caráter formulaico. É possível verificar quais porções de textos acadêmicos repetem-se consistentemente. Essas porções ou pedaços de texto, denominados pacotes lexicais[2], carregam forma e sentido, mesmo que incompletos, e exercem funções na redação de um texto. Diferentes pesquisadores buscam categorizar os pacotes lexicais em taxonomias pragmático-funcionais, permitindo, em suma, que estudos possam investigar quais tipos de pacotes são mais ou menos usados por aqueles que escrevem o texto.

Diante do grande potencial trazido pelos estudos sobre pacotes lexicais para o exame do gênero acadêmico, o grupo de pesquisa do qual faço parte, mesmo antes da minha entrada no

[1] Os termos *escrita acadêmica, discurso acadêmico e gênero acadêmico,* adotados neste estudo, são considerados amplos, uma vez que englobam diferentes graus de escrita acadêmica como textos argumentativos elaborados por aprendizes de inglês da graduação do curso de letras de diferentes universidades brasileiras e internacionais, dissertações de mestrado e teses de doutorado de nativos e não nativos, e de artigos científicos produzidos por acadêmicos experientes, entre outros. Enfatiza-se, porém, a importância de considerar dados dessa natureza nas pesquisas da área da Linguística de Corpus, evitando, desse modo, generalizações equivocadas acerca de gêneros textuais distintos, que compõem a escrita acadêmica, por exemplo. Este trabalho investigou principalmente o gênero redação argumentativa. Esse gênero será definido no capítulo de metodologia.

[2] O termo pacote lexical será definido ao longo desta dissertação e, em detalhes, na seção 2.3.

Programa de Pós-Graduação em Estudos Linguísticos, procurava contrastar o uso de pacotes lexicais na escrita acadêmica de aprendizes brasileiros de inglês ao de nativos. Para esse fim, uma das práticas do grupo era a de classificar os pacotes advindos da produção de aprendizes, segundo uma ou outra taxonomia, para que depois fosse possível realizar análises dos tipos de pacotes mais e menos utilizados. Porém, detectou-se uma grande dificuldade em classificar alguns desses pacotes lexicais. Ocasionalmente, não era possível extrair nenhuma função pragmático-funcional de um pacote lexical, ou encontrávamos outros muito parecidos entre si, com uma ou duas palavras distintas apenas. Isso nos fazia classificá-los de uma forma ou de outra, sem uma justificativa bem definida. Era comum a discordância de classificação entre os membros do grupo, sem todavia chegar a uma decisão em conjunto.

A inconsistência e dificuldade na classificação desses pacotes foram alguns dos atrativos para estudá-los com ênfase. Esta dissertação é resultante desse estudo e demonstra como pacotes lexicais cuja produção é influenciada pelo tema das redações, bem como pacotes lexicais que se sobrepõem um ao outro, afetam sua análise quantitativa e qualitativa.

1.2 Justificativa

O termo pacote lexical é proposto por linguistas que utilizam a Linguística de Corpus como ferramenta. Para que se entenda as razões por trás desse fato, bem como a justificativa deste estudo, é necessário que se atente para algumas das principais características dessa área metodológica.

A LC é uma área de pesquisa que cresce consistentemente ao longo das últimas décadas e aplica-se a estudos bastante diversos como os da Sociolinguística, Ensino e Aprendizagem de Línguas Estrangeiras, Análise do Discurso, Estilística, Linguística Forense, Pragmática, Tecnologia da Fala, Comunicação em Saúde, Lexicografia, entre outros (O'KEEFFE; MCCARTHY, 2010) A LC disponibiliza meios eficazes para a análise de dados oriundos da produção natural da língua, podendo responder perguntas de pesquisa que vão além da descrição do léxico e de aspectos gramaticais (O'KEEFFE; MCCARTHY, 2010).

Os estudos da LC baseiam-se em quatro características elementares: (1) sua natureza é empírica, uma vez que busca-se analisar padrões de uso em textos naturais; (2) utiliza-se uma

coletânea esquematizada de textos naturais, o *corpus*, como base para as análises; (3) utilizam-se software extensivamente; e (4) ambas as análises quantitativas e qualitativas são empregadas (BIBER *et al.*, 1998).

Além de sua utilização como ferramenta metodológica em diversas outras áreas de investigação, a LC vem contribuindo sistematicamente para as descobertas no âmbito da Linguística Aplicada, no qual este estudo se insere. Nesse contexto, inúmeros trabalhos puderam contribuir para a pesquisa concernente à área de Ensino e Aprendizagem de Inglês como Segunda Língua. Algumas das questões abordadas por essas pesquisas referem-se, por exemplo, ao uso de *corpora* dentro da sala de aula (REPPEN, 2010); à aplicação de métodos da LC em textos escritos acadêmicos (RÖMER; WULFF, 2010); às principais características da escrita avançada de alunos de diferentes disciplinas e níveis do *corpus* MICUSP[3] (ÄDEL; RÖMER, 2012); e às principais características discursivas em tarefas do teste TOEFL iBT[4] nas seções de escrita e fala (BIBER; GRAY, 2013).

Outra grande contribuição dos estudos baseados em *corpus* reflete-se nos estudos fraseológicos[5], fortemente relacionados à presente pesquisa. Tem havido, nas últimas décadas, uma mudança de foco da investigação de itens isolados para a investigação de itens fraseológicos. A LC favorece estudos com esse segundo foco, dando suporte a uma visão de linguagem que integra o léxico e a gramática (RÖMER, 2011), tornando possível oferecer aos aprendizes uma análise mais completa e integrada da linguagem. Esta mudança torna-se relevante uma vez que pôde ser verificado, também por meio de estudos empíricos, que aproximadamente 21% do discurso acadêmico do *Longman Spoken and Written English Corpus* (LSWE) é formado por expressões recorrentes, incluindo verbos frasais, preposicionais e pacotes lexicais de 3 e 4 palavras (BIBER; CONRAD; *et al.*, 1999). Examinando o *corpus* geral do discurso oral do inglês *London-Lund Corpus of Spoken English*[6], Altenberg (1998) estima que

[3] O *Michigan Corpus of Upper-level Student Papers* (MICUSP) é composto por trabalhos acadêmicos que receberam conceito A, de diferentes disciplinas ofertadas pelos cursos da Universidade de Michigan, e contém aproximadamente 2.6 milhões de palavras.

[4] O TOEFL iBT é um teste de proficiência amplamente reconhecido por instituições de ensino superior e universidades em mais de 130 países. Além das seções de leitura e escuta, o teste apresenta as seções de escrita e fala nas quais os candidatos devem redigir textos argumentativos e expressar sua opinião sobre um tópico familiar, respectivamente.

[5] Os estudos fraseológicos e sua conexão com a Linguística de Corpus serão explorados no próximo capítulo deste trabalho.

[6] http://www.helsinki.fi/varieng/CoRD/corpora/LLC/

em torno de 80% do *corpus* seja composto por combinações de palavras recorrentes, definidas como qualquer combinação de palavra que ocorra mais de uma vez e de forma idêntica. Não nos compreenderíamos se a língua não fosse formulaica. Mesmo que obedecêssemos à organização sintática do português, por exemplo, não seria possível comunicar o sentido desejado se organizássemos algumas das palavras de uma sentença, como esta que escrevo neste momento, randomicamente. A linguagem é, portanto, predominantemente construída por expressões pré-fabricadas, regulada pelo princípio idiomático e não pelo princípio da livre escolha[7] (SINCLAIR, 1991).

Estudos da área de Linguística Aplicada, realizados sob essa perspectiva, buscam contrastar a produção de falantes nativos à de aprendizes, como citado anteriormente, tanto no discurso oral como no escrito, a partir de uma investigação da frequência e das funções dessas expressões, os chamados pacotes lexicais (BIBER; CONRAD; CORTES, 2004; BIBER, 2009; CHEN; BAKER, 2010; DUTRA; BERBER-SARDINHA, 2013; SIMPSON-VLACH; ELLIS, 2010). Esses itens são tecnicamente entendidos como sequências de palavras que comumente co-ocorrem no discurso natural (BIBER; CONRAD; *et al.*, 1999) e que destacam-se como importantes fontes de significado no discurso (RÖMER, 2011). Os pacotes lexicais são comumente conhecidos na literatura por outros termos, salvo afiliações teóricas, como expressões pré-fabricadas ou pré-padronizadas, *chunks*, sequências formulaicas, colocações, *clusters*, n-gramas e *multi-word units* (MWUs). Pacotes lexicais são considerados de grande importância uma vez que "[...] são os tijolos da construção do discurso." (tradução minha)[8] (BIBER; CONRAD; CORTES, 2004, p.371); "[...] representam as mais importantes necessidades comunicativas de um registro." (tradução minha)[9] (BIBER, 2009, p. 285); "[São] um importante componente da fluência na produção linguística [...]" (tradução minha)[10] (HYLAND, 2008, p. 41); e, finalmente, "Servem como uma importante medida de desenvolvimento linguístico de aprendizes." (tradução minha)[11] (STAPLES *et al.*, 2013, p. 214).

O estudo dos tipos de pacotes lexicais utilizados por aprendizes mais ou menos proficientes e por nativos pode revelar informações importantes tanto para a descrição linguística

[7] Esses princípios serão detalhados no próximo capítulo, na seção 2.2.
[8] "*[...] they function as basic building blocks of discourse*."
[9] "*[... they] serve the most important communicative needs of a register.*"
[10] "*[Theyess are] an important component of fluent linguistic production […]*"
[11] "*[They have] shown to be an important measure of learner development*."

da produção desses indivíduos quanto para servir de insumo para o desenvolvimento de atividades a serem aplicadas na sala de aula. Essas atividades, por sua vez, podem otimizar o desenvolvimento da fluência e acuidade de aprendizes de inglês na escrita acadêmica, por exemplo. A familiarização de algumas construções formulaicas pode ser entendida como um dos diversos aspectos necessários para que se alcance o domínio desse tipo de escrita, em uma língua estrangeira. Mesmo pesquisadores brasileiros com muitos anos de experiência na academia enfrentam dificuldades para publicar seus trabalhos científicos em revistas internacionais, e uma das causas para isso pode estar associada à falta de domínio daquele gênero, incluindo o não conhecimento sobre quais pacotes lexicais são característicos do gênero textual em questão. Trabalhos desse tipo podem também contribuir para informar a criação de conteúdos programáticos, uma vez que atestam quais tipos de estruturas merecem maior atenção na sala de aula em detrimento de outras. Alguns dos estudos fraseológicos mais relevantes são apresentados no próximo capítulo.

Na literatura, porém, existem resultados divergentes quanto à correlação entre o número de pacotes lexicais utilizados por um indivíduo e o seu nível de proficiência. Em um de seus trabalhos, Hyland (2008) pôde verificar que estudantes de mestrado, alunos menos proficientes dentre os grupos pesquisados, fizeram uso de mais pacotes lexicais, enquanto que escritores especializados, habituados à redação de artigos acadêmicos e, portanto, o grupo mais proficiente, utilizaram menos pacotes lexicais em sua produção escrita, tanto em relação ao emprego de *types* quanto ao de *tokens*[12]. Por outro lado, Chen & Baker (2010) descobriram que o número de combinações de palavras recorrentes cresce de acordo com a proficiência, novamente, tanto em relação ao emprego de *types* quanto ao de *tokens*. É importante salientar que esses estudos utilizaram *corpora* e metodologias distintas. Outra possível razão para essa divergência, levantada por Chen & Baker no mesmo estudo supracitado, estaria relacionada à manutenção ou eliminação de *pacotes lexicais relacionados ao tópico* e/ou de *pacotes lexicais em contexto de*

[12] *Types* refere-se à quantidade de tipos diferentes de pacotes lexicais e *tokens* refere-se à quantidade total de ocorrências de pacotes lexicais.

sobreposição[13]. A seguir, as definições elaboradas por esses mesmos autores acerca desses tipos de pacotes lexicais são apresentadas.

Pacotes lexicais relacionados ao tópico são construídos por palavras de conteúdo presentes nas instruções da redação ou dependentes do contexto, e geralmente incorporam nomes próprios, *e.g., students using credit cards, the opium of the*. Esses pacotes são considerados sequências de itens que foram produzidas de maneira não natural, uma vez que só passaram a existir devido à presença dos tópicos utilizados para a produção do texto. Os exemplos de pacotes citados acima foram produzidos a partir das seguintes instruções: *Discuss the advantages and disadvantages of using credit cards* e *Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television*. Esse tipo de pacote não nos diz nada, ou muito pouco, a respeito da produção linguística dos autores dos textos.

Já os *pacotes lexicais em contexto de sobreposição* desdobram-se em duas situações. A primeira, chamada *sobreposição completa*, refere-se a, por exemplo, pacotes de 4 palavras que, na verdade, derivaram-se de uma combinação de 5 palavras. A título de ilustração, consideremos que ao gerar-se uma lista de pacotes lexicais de 4 palavras de um dos *subcorpora* do ICLEv2[14], os pacotes *there are a lot* e *are a lot of* foram produzidos, ambos com 28 ocorrências. Esses pacotes são muito parecidos entre si e parecem fazer parte de uma mesma sequência. Ao somarmos um pacote ao outro, obtemos o pacote maior de 5 palavras: *there are a lot of.* Através de uma checagem da ocorrência desse pacote maior no *corpus*, obtemos o mesmo número de ocorrências, ou seja, o pacote de 5 palavras ocorre também 28 vezes. É possível concluir, portanto, que esses dois pacotes de 4 palavras, na verdade, formam um único pacote de 5 palavras, *there are a lot of.* Eles não ocorrem nenhuma vez isoladamente: *there are a lot* não ocorre nenhuma vez sem a partícula *of* e *are a lot of* não ocorre nenhuma vez sem a partícula *there.* Além disso, não há nenhuma partícula que se repete antes ou depois de cada uma dessas combinações (*vide* Quadro 1).

---

[13] Os termos *pacotes lexicais relacionados ao tópico* e *pacotes lexicais em contexto de sobreposição*, bem como os desdobramentos do segundo termo, foram destacados em itálico nesta dissertação por serem esses os objetos de estudo desta pesquisa e tipos de pacotes lexicais específicos, definidos por Chen & Baker (2010).

[14] *International Corpus of Learner of English Version 2* (ICLEv2), compilado por Granger *et al.,* (2009). Esse *corpus* será descrito em detalhes no capítulo de metodologia desta dissertação.

Quadro 1 - Demonstração do contexto de *sobreposição completa* dos pacotes lexicais *are a lot of* e *there are a lot*

N° de palavras	**Frequência**	**Pacote Lexical**				
4	28		*are*	*a*	*lot*	*of*
4	28	*there*	*are*	*a*	*lot*	
5	28	*there*	*are*	*a*	*lot*	*of*

A segunda situação, chamada *subsunção completa*, refere-se a dois ou mais pacotes lexicais de 4 palavras, por exemplo, que se sobrepõem, e a ocorrência de um deles incorpora a ocorrência do outro pacote. Tomemos como exemplo a mesma lista de pacotes de 4 palavras na qual *at the same time* e *and at the same* foram produzidos. Desta vez, o primeiro pacote apresentou uma frequência de 29 ocorrências, enquanto que o segundo apresentou uma frequência de 6 ocorrências. A partir de uma checagem no *corpus*, é possível verificar que ao somarmos os dois pacotes, obtemos a sequência maior *and at the same time*, de 5 palavras, com também 6 ocorrências (*vide* Quadro 2).

Conclui-se, a partir daí que *at the same time* tem sua ocorrência inflacionada, pois incorpora as ocorrências de *and at the same time* em seu número total de frequência. Isso ocorre porque uma lista de pacotes de 4 palavras não permite que o pacote de 5 palavras seja visualizado. Porém, o pacote maior *and at the same time* existe com 6 ocorrências. Além disso, o fato de que o pacote de 4 palavras *and at the same* não ocorre nenhuma vez sem a partícula *time* demonstra que ele, na verdade, não é um pacote completo, ou seja, ele inexiste somente com 4 palavras. Como *and at the same* nunca existirá sem *time*, então, o pacote *at the same time* contém, dentro dele, a contagem do pacote maior de 5 palavras *and at the same time*. *At the same time* tem o seu número de ocorrências aumentado 6 vezes mais, em uma lista que só mostra pacotes de 4 palavras. Para que a frequência deste pacote seja real, é necessário retirar as ocorrências de *and at the same time* de sua contagem. Portanto, *at the same time* ocorre, na verdade, 23 vezes (29 – 6 = 23). Destacam-se em negrito no Quadro 2 os pacotes resultantes do processo de refinação, *and at the same*, de 4 palavras, com 6 ocorrências e *and at the same time*, de 5 palavras com 23 ocorrências. Essa incorporação indesejável será detalhada mais adiante.

Quadro 2 - Demonstração do contexto de *subsunção completa* dos pacotes lexicais *at the same time* e *and at the same*

N° de palavras	**Frequência**	**Pacote Lexical**				
4	29		*at*	*the*	*same*	*time*
4	6	*and*	*at*	*the*	*same*	
5	6	*and*	*at*	*the*	*same*	*time*
Processo de Refinamento da Frequência dos Pacotes						
4	29 -		*at*	*the*	*same*	*time*
5	**6 =**	***and***	***at***	***the***	***same***	***time***
4	**23**		***at***	***the***	***same***	***time***

Os software existentes na atualidade para a geração de listas de pacotes lexicais como WordSmith Tools (SCOTT, 1998), Collocate (BARLOW, 2004) e AntConc (ANTHONY, 2011) não lidam com essas questões. Neles, o pesquisador deve escolher o tamanho desejado para os pacotes lexicais – quantas palavras os pacotes devem ter – e o software gera a lista automaticamente. *Pacotes lexicais relacionados ao tópico* e/ou *pacotes lexicais em contexto de sobreposição* são, porém, extremamente frequentes nas listas geradas – como demonstrado no Quadro 3, a seguir. Pelo menos 19 dos 20 pacotes mais frequentes são de um desses dois tipos no *corpus* utilizado[15]. Esses pacotes distribuem-se ao longo da lista gerada, ou seja, não ocorrem somente nas primeiras posições de frequência. *Pacotes lexicais relacionados ao tópico* não apresentam funções pragmáticas e não são previstos pelas principais listas[16] de categorização de pacotes lexicais criadas até os dias de hoje (BIBER; CONRAD; CORTES, 2004; HYLAND, 2008; SIMPSON-VLACH; ELLIS, 2010) e tampouco devem ser, uma vez que não possuem uma função pragmática na escrita, a não ser talvez, a de atender às instruções dadas para a redação do texto.

[15] Utilizou-se para a demonstração da alta frequência de *pacotes lexicais relacionados ao tópico* e de *pacotes lexicais em contexto de sobreposição* o *subcorpus* dos aprendizes de inglês de língua materna chinesa do ICLEv2, detalhado no capítulo de metodologia desta dissertação.

[16] Essas listas serão apresentadas no capítulo de metodologia desta dissertação.

Quadro 3 - 20 pacotes lexicais de 4 palavras mais frequentes da lista gerada pelo programa Collocate (BARLOW, 2004) do *subcorpus* dos aprendizes de inglês de língua materna chinesa do ICLEv2

Ranque	**Frequência**	**Pacote lexical**
1	399	*banning smoking in restaurants*
2	399	*In this essay I*
3	241	*students using credit cards*
4	237	*advantages and disadvantages of*
5	211	*on the other hand*
6	204	*of banning smoking in*
7	194	*the advantages and disadvantages*
8	182	*of students using credit*
9	167	*professionals from Mainland China*
10	166	*pros and cons of*
11	158	*the pros and cons*
12	129	*method of waste management*
13	109	*cancer and heart disease*
14	106	*lung cancer and heart*
15	101	*this essay I will*
16	99	*ban on smoking in*
17	94	*importing professional from Mainland*
18	93	*smoking in restaurants is*
19	91	*According to R the*
20	90	*on smoking in restaurants*

Apesar disso, pesquisadores vêm realizando estudos a partir dessas listas, contabilizando esses tipos de pacotes, ou ignorando alguns, quando esses lhe parecem muito estranhos, sem nenhuma sistematicidade, ou sem explicitar de que maneira lidaram com esses tipos de sequências. A classificação pragmático-funcional dessas listas torna-se extremamente difícil, uma vez que muitos dos pacotes ali presentes apresentam-se divididos em partes, ou seja, são na verdade um só pacote. Além dessa dificuldade, a análise de listas de pacotes que não elimina *pacotes lexicais relacionados ao tópico* e/ou de *pacotes lexicais em contexto de sobreposição* apresentará números inflacionados já que um mesmo pacote será contabilizado mais de uma vez.

Além disso, a não eliminação desses pacotes favorecerá outros que não deveriam ser levados em consideração na análise e desfavorecerá o aparecimento de pacotes maiores, que estão escondidos nos contextos de sobreposição, afetando a análise quantitativa dos pacotes lexicais nos estudos.

Os exemplos de *pacotes lexicais relacionados ao tópico* e de *pacotes lexicais em contexto de sobreposição* discutidos acima só demonstraram pares para cada uma das situações. É comum, porém, que pacotes que encontram-se em uma posição de frequência alta estejam diretamente relacionados a vários outros, em posições de frequências mais baixas, em ambos as situações. No *corpus* do grupo de aprendizes de inglês de língua materna chinesa do ICLEv2, por exemplo, *in this essay I*, com 345 ocorrências, está relacionado aos pacotes *this essay I will*, com 100 ocorrências; *this essay I have*, com 68 ocorrências; *this essay I examine*, com 66 ocorrências; *this essay I would*, com 37 ocorrências; *this essay I am*, com 32 ocorrências; e *this essay I discuss*, com 16 ocorrências, ao mesmo tempo. *(vide* Figura 1). Ao gerarmos a lista de pacotes lexicais de quatro palavras no programa Collocate (BARLOW, 2004), os pacotes mencionados acima aparecem nas seguintes posições de acordo com o seu número de ocorrências: 1, 17, 53, 57, 223, 310 e 985, respectivamente.

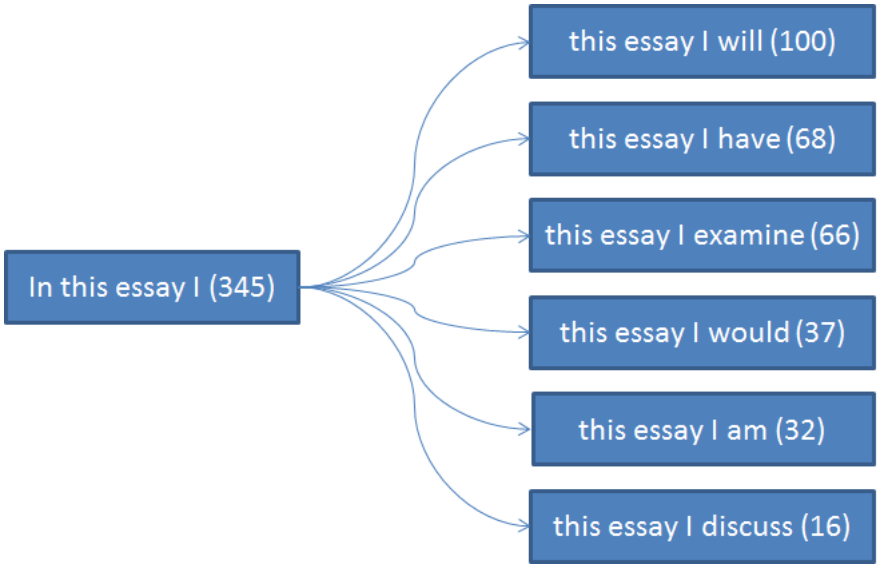

Figura 1 - Demonstração de algumas das relações do pacote lexical *in this essay I* no *corpus* de aprendizes de língua materna chinesa do ICLEv2

Além disso, as relações entre os pacotes podem ocorrer não somente para a direita, ou para a esquerda como nos exemplos apresentados até então, mas também para a direita e esquerda simultaneamente, como é o caso das sequências *advantages and disadvantages of, both advantages and disadvantages of* e *advantages and disadvantages of the*. Isso demonstra o grau de dificuldade apresentado aos pesquisadores que analisam listas de pacotes lexicais contendo cerca de 1900 itens, como ocorre no *corpus* utilizado para essa demonstração.

Além de Chen & Baker (2010), alguns estudos recentes começam a abordar essa questão para a investigação da relação entre proficiência e utilização de pacotes lexicais (BOHÓRQUEZ *et al.*, 2012; STAPLES *et al.*, 2013). Em linhas gerais, esses estudos corroboram a importância de eliminar um ou ambos os tipos de pacotes. O primeiro estudo corrobora a correlação entre maior nível de proficiência e maior uso de pacotes lexicais quando *pacotes relacionados ao tópico* e/ou *pacotes em contexto de sobreposição* são eliminados. A pesquisa propõe uma metodologia manual para a eliminação desses tipos de pacotes. Ressalta-se, porém, o longo período de tempo despendido no processo metodológico proposto, bem como a possibilidade da presença de erros, ambos referentes ao fato da eliminação ser realizada por humanos, além de seu caráter incipiente. Esta dissertação é resultante de um avanço da metodologia elaborada nesse trabalho. O segundo estudo argumenta que aprendizes menos proficientes utilizaram mais pacotes lexicais, mas também mais pacotes idênticos a partes das instruções. Após a exclusão desses pacotes, o estudo revelou que o nível intermediário de proficiência fez mais uso de pacotes lexicais, enquanto que o nível mais baixo e mais alto apresentaram resultados bastante similares (STAPLES *et al.*, 2013). Esse estudo, porém, não menciona *pacotes lexicais em contexto de sobreposição*.

Acredita-se, portanto, que outras pesquisas sejam necessárias para investigar a correlação entre nível de proficiência e uso de pacotes lexicais. Além disso, o desenvolvimento de uma metodologia eficaz e automatizada para a eliminação de *pacotes lexicais relacionados ao tópico* e de *pacotes lexicais em contexto de sobreposição* contribuiria diretamente para os estudos da produção de escrita acadêmica de aprendizes e nativos, e para as análises linguísticas concernentes aos pacotes lexicais. Dessa forma, as listas de pacotes lexicais geradas pelos software mais utilizados na atualidade poderão ser limpas, facilitando a classificação pragmático-funcional de pacotes lexicais e gerando resultados mais fidedignos à produção dos indivíduos

investigados. A otimização dessas listas poderá favorecer a descrição e análise dos traços de similaridade e discrepância entre a produção do *continuum* aprendiz-nativo. Professores da língua inglesa poderão basear-se nos resultados encontrados e desenvolver atividades e intervenções que favoreçam o desenvolvimento linguístico de seus alunos. A familiarização de estruturas como os pacotes lexicais por parte dos estudantes pode colaborar para a aquisição de fluência e acuidade na redação de textos acadêmicos em inglês (SIYANOVA; SCHMITT, 2008). As contribuições deste trabalho poderão ser estendidas a investigações da produção de aprendizes de outras línguas e aproximar as pesquisas da LC à sala de aula.

1.3 Objetivos

1.3.1 Objetivo geral

- Investigar a correlação entre o uso de pacotes lexicais e o nível de proficiência linguística, utilizando um *corpus* de textos argumentativos de aprendizes de inglês.

1.3.2 Objetivos específicos

- Desenvolver uma metodologia automatizada para a eliminação de *pacotes lexicais relacionados ao tópico* e de *pacotes lexicais em contexto de sobreposição*.
- Verificar se a eliminação de *pacotes lexicais relacionados ao tópico* e de *pacotes lexicais em contexto de sobreposição* pode ser um fator que influencie a percepção da correlação entre maior nível de proficiência e uso de mais pacotes lexicais na escrita acadêmica de inglês.

1.4 Perguntas de pesquisa

- A metodologia automatizada é capaz de eliminar *pacotes lexicais relacionados ao tópico* e *pacotes lexicais em contexto de sobreposição* de maneira eficaz?
- Como a eliminação dos tipos de pacotes lexicais supracitados afeta o resultado em relação às ocorrências de *p*acotes lexicais nos *corpora* investigados?
- Após as eliminações desses pacotes, é possível correlacionar maior nível de proficiência e maior uso de pacotes lexicais?

1.5 Organização da dissertação

Esta dissertação divide-se em 5 capítulos, sendo o primeiro “Introdução”, já apresentado ao leitor, dedicado à contextualização do leitor no escopo do estudo. Para isso, a motivação para a elaboração da pesquisa bem como as características primordiais à área da Linguística de Corpus e noções inerentes aos estudos fraseológicos são apresentadas. Este capítulo tratou também da relevância da pesquisa, dos objetivos a serem alcançados pelo estudo, e das perguntas de pesquisa a serem respondidas.

O segundo capítulo “Fundamentação Teórica” trata das principais contribuições dos estudos fraseológicos para a descrição linguística do gênero de escrita acadêmica e faz um apanhado dos principais e mais atuais estudos concernentes aos pacotes lexicais derivados de *corpora* de aprendizes. Além disso, são detalhados aqueles que abordam a problemática relacionada a *pacotes lexicais relacionados ao tópico* e *pacotes lexicais em contexto de sobreposição*.

O terceiro capítulo “Metodologia” discorre sobre os passos metodológicos adotados no estudo. Primeiramente, os *corpora* utilizados são definidos. Em seguida, os métodos utilizados para o desenvolvimento da metodologia automatizada para a eliminação dos pacotes em questão, bem como sua validação, são apresentados. Finalmente, os procedimentos de análise acerca da comparação dos resultados anteriormente e posteriormente à aplicação da metodologia proposta são expostos.

O quarto capítulo “Resultados” apresenta, primeiramente, uma análise preliminar à aplicação da metodologia desenvolvida. Em seguida, discorre-se sobre a análise dos resultados comparando-se a produção de *pacotes lexicais relacionados ao tópico* e de *pacotes lexicais em contexto de sobreposição* por parte dos aprendizes de inglês representados pelos *corpora* do estudo, antes e depois das eliminações e refinações realizadas.

O quinto e último capítulo “Conclusão”, apresenta as considerações finais do trabalho, suas principais contribuições, bem como suas principais limitações, e os possíveis desdobramentos para pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo busca contextualizar o leitor no escopo dos estudos concernentes aos pacotes lexicais principalmente no contexto da escrita acadêmica. Para isso, a primeira seção traz uma introdução sobre fraseologia e a segunda, uma síntese da linha de desenvolvimento dos estudos fraseológicos ligados à Linguística de Corpus. Em seguida, para que a natureza dos pacotes lexicais seja bem definida, a terminologia adotada pela LC acerca dos diferentes tipos de itens fraseológicos é detalhada. Na seção seguinte, são apresentados os trabalhos mais relevantes que correlacionam nível de proficiência e uso de pacotes lexicais na escrita acadêmica, buscando demonstrar a relevância desse tipo de pesquisa na contribuição para o melhor ensinar desse gênero textual. As taxonomias existentes para a classificação funcional de pacotes lexicais, utilizadas em cada estudo, são apresentadas. Discorre-se, ainda nessa parte, sobre as principais motivações para a divergência dos resultados encontrados em cada pesquisa. Por fim, as metodologias de eliminação de *pacotes lexicais relacionados ao tópico* e/ou de *pacotes lexicais em contexto de sobreposição* adotadas em trabalhos anteriores são descritas.

2.1 Panorama da fraseologia

Antes que se faça um breve panorama sobre a fraseologia e sua relação com a LC, parece-me interessante abrir um parêntese para um comentário acerca de uma percepção pessoal que me vem à mente, neste momento, e que não obstante, relaciona-se fortemente ao tema. Recentemente, precisei do número do CPF do meu pai. Perguntei a ele qual era o número. Enquanto eu anotava, percebi que quando meu pai pausava, ele não conseguia continuar de onde havia parado. Era necessário voltar ao primeiro número para falar a sequência novamente. Relacionei esse fato a outro. Percebo que tenho muito mais facilidade em informar o meu número de telefone quando o falo rapidamente, por inteiro, do que pausadamente, número a número. Seria possível relacionar esses dois fatos ao princípio básico da fraseologia? Assim como os itens de “unidades fraseológicas”, “combinações de palavras”, “lexemas frasais” (COWIE, 1998), ou “n-gramas” (SINCLAIR, 2004b *apud* SHEPHERD, 2009, p.157), “agrupamentos” e “feixes lexicais” (BIBER, 2004; BIBER *et al.*, 2004 *apud* SHEPHERD, 2009,

p.157)[17], não importa a terminologia, esses números são memorizados por nós como um todo. De fato, a literatura parece corroborar esse fenômeno. Estudos afirmam que sequências formulaicas são processadas de maneira mais eficaz pelo fato de seus componentes individuais formarem uma única unidade de significado. Essa unidade pode ser processada mais rapidamente e mais facilmente do que os seus itens separadamente, em contexto criativo (ELLIS; SIMPSON-VLACH; MAYNARD, 2008).

Outro estudo pôde demonstrar, além de apresentar um modelo de aquisição para linguagem formulaica e de um esquema de priorização de quais itens são mais válidos pedagogicamente, uma validação psicolinguística das fórmulas derivadas de *corpora* (ELLIS; SIMPSON-VLACH; MAYNARD, 2008). Para essa validação, experimentos foram realizados com o objetivo de investigar, por exemplo, a rapidez de leitura e aceitação no que diz respeito à gramaticalidade de diferentes fórmulas, algumas autênticas e outras não existentes na língua inglesa, por parte de aprendizes de inglês e de nativos. Esses pesquisadores também citam vários estudos que apontam para o fato do processamento linguístico ser sensível à formulaicidade. O presente trabalho não abordará essa característica mais a fundo – a de maior facilidade de estocagem de expressões pré-fabricadas no léxico mental – uma vez que essa questão foge do foco de investigação da presente pesquisa. Voltando a fraseologia *per se*, uma definição faz-se necessária nesta ocasião. Gries (2008, p. 5), considera que

> […] um fraseologismo é a co-ocorrência de uma forma ou lema de um item lexical e de qualquer outro tipo de elemento linguístico, o qual pode ser, por exemplo,
> – outro item lexical, ou outra forma de um item lexical, (*kith* e *kin* é um exemplo de uma co-ocorrência quase que determinística de dois itens lexicais, assim como *strong tea*);
> – um padrão gramatical (em oposição a, digamos, uma relação gramatical), *i.e.*, quando um item lexical específico tende a ocorrer em/co-ocorrer com uma construção gramatical específica (o fato do verbo *hem* ser mais frequentemente utilizado na voz passiva é um exemplo característico). (tradução minha)[18]

[17] Esses são alguns outros termos fraseológicos adotados na literatura, ainda não mencionados nesta dissertação.

[18] *[…] a phraseologism [… is] the co-occurrence of a form or lemma of a lexical item and any other kind of linguistic element, which can be, for example,*
– another (form of a) lexical item (kith and kin is a very frequently cited example of a nearly deterministic co-occurrence of two lexical items, as is strong tea);
– a grammatical pattern (as opposed to, say, a grammatical relation), i.e. when a particular lexical item tends to occur in/co-occur with a particular grammatical construction (the fact that the verb hem is mostly used in the passive is a frequently cited case in point).

O autor ainda argumenta que o interesse pela área tem crescido nos últimos 30 anos. É somente durante esse período que a fraseologia tem se consolidado como um campo de pesquisa da Linguística Teórica e Aplicada, tendo esse processo iniciado anteriormente na Europa oriental (COWIE, 1998). Os principais motivos para a negligência da fraseologia por parte das descrições linguísticas durante tanto tempo, relaciona-se a sua oposição à gramática tradicional, de natureza atomística e paradigmática, enraizada pela teoria gerativa (COWIE, 1998; SINCLAIR, 2008). Os dias atuais, porém, proporcionam oportunidades para uma crescente conscientização da prevalência de combinações pré-fabricadas no discurso escrito e oral e de seu papel central na aquisição de primeira e segunda língua (COWIE, 1998).

Os estudos fraseológicos são baseados em variadas abordagens teóricas, englobando interesses de linguistas teóricos, lexicógrafos, analistas do discurso, analistas computacionais, e pesquisadores das áreas de aquisição e de ensino de línguas e estilística (COWIE, 1994, 1998). Ainda de acordo com esse autor, existem três principais correntes dos estudos fraseológicos: aquela vinculada à análise colocacional da teoria lexical neo-Firthiana, liderada principalmente por Michael Halliday e John Sinclair; aquela mais essencialmente Firthiana e ligada à teoria fraseológica russa; e aquela com um viés antropológico, adicionando a dimensão cultural à tradição fraseológica russa, representada principalmente pelo trabalho de Veronika Teliya. Esta pesquisa insere-se na primeira perspectiva por afiliar-se aos estudos com ênfase na frequência observada de co-ocorrências advindas de grandes *corpora* digitais. A seção a seguir tratará das principais contribuições de Firth para os estudos fraseológicos da primeira perspectiva citada, bem como do trabalho desenvolvido por Halliday e Sinclair.

2.2 A Fraseologia e a Linguística de Corpus

O pai da linguística, como era conhecido John Rupert Firth em seu país, foi uma figura essencial no processo de consolidação de noções, na época visionárias, como a de contexto de situação, colocação, espectro de significado, entre outras, e de teorias para análise prosódica (GRANGER; MEUNIER, 2008; PALMER, 1968). Seu reconhecimento tardio se deu, muito provavelmente, por Firth ter sido um questionador das abordagens tradicionais, dominantes na época. Firth pôde desenvolver suas teorias e ao mesmo tempo contribuir para suas aplicações em

novas áreas da linguística como na tradução, na descrição do inglês e no ensino de línguas (PALMER, 1968).

No que tange o escopo da presente pesquisa, Firth (1968a, p. 12) ilustra bem a visão tradicionalista prevalente em sua época, a que ele se opõe, de que as palavras, isoladamente, contêm significado por si só. Ele argumenta por uma análise linguística em níveis mutualmente complementares, sejam eles situacionalmente contextuais, colocacionais ou estruturais, entre outros[19]:

> Todo pensamento sistemático deve começar a partir de pressuposições e, ao lidarem com o significado, alguns acadêmicos admitem a hipótese de que palavras isoladas listadas em um dicionário e orações isoladas, cada uma delas interpoladas por pontos finais, poderiam ser naturalmente examinadas quanto ao seu significado sob completa abstração de seu contexto circundante. [...] […] Lógicos continuam a tratar as palavras […] como se, de alguma maneira, elas tivessem o significado em si e por si só. Alguns linguistas seguem esse método secular de análise linguística simplesmente pelo peso da tradição filosófica e lógico-gramatical. Ambas essas pressuposições são ilusórias na linguística […]. (tradução minha)[20]

Em outro trabalho em que esquematiza os princípios de sua teoria de linguística geral, Firth (1968b) apresenta alguns termos essenciais, e dentre eles encontra-se o da colocação. Nesse trabalho, ele evidencia que o significado das palavras é dependente do seu contexto de situação, ou seja, de seu uso. É nesse trabalho também que Firth introduz sua famosa frase "Conhece-se uma palavra por suas companhias" (tradução minha p. 179)[21] e que de forma criativamente explicativa demonstra que as palavras possuem diferentes fisionomias, dependendo de seus colocados. Os exemplos reproduzidos a seguir sintetizam bem esse entendimento de Firth (1968b, p. 179):

> Segue-se que um texto em um tipo específico de uso pode conter sentenças tais quais *'Don't be such an ass!', 'You silly ass!', 'What an ass he is!'* Nesses exemplos, a palavra *ass* encontra-se em companhia familiar e habitual, comumente colocada com *you silly-*, *he is a silly-*, *don't be such an-*. […] Um dos significados de *ass* encontra-se

[19] A teoria de linguística geral elaborado por Firth assume que a língua deve ser descrita, ao mesmo tempo que integralmente, de acordo com diferentes níveis de análise, dentre eles os citados aqui. Para maior detalhamento sobre a teoria, recomenda-se a leitura de Firth (1968b).

[20] *All systematic thought must start from presuppositions and in dealing with meaning some scholars have supposed single words listed in a dictionary and single sentences each bounded by full stops could be safely examined as to their meaning in complete abstraction from specific environment.* [...] *[...] Logicians continue to treat words [...] as if they somehow could have meaning in and by themselves. Some linguists follow this centuries-old method of linguistic analysis merely because of the weight of philosophical and lógico-grammatical tradition. Both these pre-suppositions are misleading in linguistics […]*.

[21] *You shall know a word by the company it keeps.*

> em sua habitual colocação com outras palavras, como as citadas acima.[…]. (tradução minha)[22]

Em suma, Firth argumenta que cada palavra, quando empregada em um contexto diferente, é uma nova palavra, podendo esse contexto pertencer a diferentes níveis de significado.

Baseados nos pressupostos elaborados por Firth, os chamados neo-Firthianos, mais fortemente representados por Sinclair e Halliday, puderam contribuir para o desenvolvimento dessa visão distinta da língua, cujos pressupostos opunham-se aos elaborados pela teoria Chomskyana. Os princípios da teoria neo-Firthiana, sintetizados por Stubbs (1996, p. 23) e reproduzidos a seguir, formam os pilares da LC, principalmente como uma área movida a *corpus* (*corpus-driven*), mas também possíveis em um viés baseado em *corpus* (*corpus-based*)[23].

> - Sobre a natureza da linguística: é essencialmente uma ciência social e aplicada, com implicações práticas, principalmente na educação […].
> - Sobre a natureza dos dados para a linguística: a língua deve ser estudada como instâncias autênticas de uso (não como orações criadas, advindas da intuição); a língua deve ser estudada como textos íntegros (não como orações isoladas ou fragmentos de texto); textos devem ser estudados comparativamente a outros *corpora*.
> - Sobre os objetos de estudo da linguística: a linguística deve estudar o significado; forma e significado não são separáveis; léxico e gramática são interdependentes […].
> - Sobre a natureza do comportamento linguístico: a língua em uso envolve ambas rotina [ou fraseologismo] e criação; a língua em uso transmite cultura […].
> - Sobre a estrutura conceitual da disciplina: os dualismos estabelecidos por Saussure (principalmente *langue-parole* e sintagmático-paradigmático) necessitam de uma revisão radical. (tradução minha)[24]

---

[22] *It follows that a text in such stablished usage may contain sentences such as 'Don't be such an ass!', 'You silly ass!', 'What an ass he is!' In these examples, the word ass is in familiar and habitual company, commonly collocated with you silly-, he is a silly-, don't be such an-. […] One of the meanings of ass is its habitual collocation with such other words as the above quoted.*[…]

[23] A Linguística de Corpus pode ser entendida com base no reconhecimento ou não de uma dicotomia *corpus-driven* e *corpus-based*. Quando há esse reconhecimento, entende-se que os estudos *corpus-based* fazem uso de *corpora* com o objetivo de explorar uma teoria ou hipótese, validando, refutando, ou refinando-a, ao passo que os estudos *corpus-driven* tomam o próprio *corpus* como única fonte de hipóteses sobre a língua (MCENERY; HARDIE, 2012). O presente trabalho não entra no mérito da dicotomia apresentada.

[24] *- The nature of linguistics: that it is essentially a social science and an applied science, with practical implications, especially in education […].*
- The nature of data for linguistics: that language should be studied in attested, authentic instances of use (not as intuitive, invented sentences); that language should be studied as whole texts (not as isolated sentences or text fragments); and that texts must be studied comparatively across text corpora.
- The essential subjects of linguistics: that linguistics should study meaning; that form and meaning are inseparable; and that lexis and grammar are interdependent […].
- The nature of linguistic behavior: that language in use involves both routine and creation; and that language in use transmits culture […].
- The conceptual structure of the discipline: that Saussurian dualisms (especially langue-parole and syntagmatic-paradigmatic) require radical revision.

Sinclair (1991) ainda elabora dois diferentes princípios de interpretação do significado: o princípio da livre escolha e o princípio idiomático. O primeiro princípio interpreta o texto como o resultado de um grande número de escolhas complexas feitas pelo falante, restringidas pela gramática. Em suma, esse princípio prevê que existem espaços a serem preenchidos, em uma oração, por exemplo, e tais espaços podem ser ocupados por qualquer palavra, contanto que ela satisfaça as restrições gramaticais daquela posição. Sentenças do tipo *The farmer kills the ducklings; Pussy is beautiful; I have not seen your father's pen, but I have read the book of your uncle's gardner*[25] seriam produtivas nesse princípio, mas não tanto no segundo, uma vez que este último revela o caráter probabilístico da linguagem. Resumidamente, o segundo princípio prevê que as palavras não ocorrem randomicamente no texto. O falante, portanto, tem a sua disposição um grande número de expressões semi-pré-construídas que constituem escolhas e significados unificados. Em conclusão a esta seção, reproduz-se a observação de Sinclair (1991, p. 108) a respeito da fraseologia:

> A maior parte de um texto é de longe composta pela ocorrência de palavras frequentes, em padrões frequentes, ou em pequenas variações desses padrões. A maior parte das palavras mais comuns não possuem um significado independente. Essas são componentes de um rico repertório de padrões formados por blocos de palavras que formam o texto. Esse fato é totalmente ofuscado pelos procedimentos da gramática convencional. (tradução minha)[26]

2.3 Os itens fraseológicos e os pacotes lexicais

O avanço tecnológico favoreceu o desenvolvimento da perspectiva dos estudos fraseológicos explorada anteriormente. A tecnologia permite atualmente que processos automáticos sejam desenvolvidos para o resgate de padrões recorrentes advindos de *corpora*, geralmente de grandes proporções, e medidas estatísticas são priorizadas nesses processos (SINCLAIR, 2008).

---

[25] Essas sentenças foram retiradas de Stubbs (1996), em referência a exemplificação irônica de Firth quanto a sentenças gramaticalmente bem formadas, porém, dificilmente produtivas no uso, utilizadas com frequência por linguistas para explanações linguísticas.

[26] *By far the majority of text is made of the occurrence of common words in common patterns, or in slight variants of those common patterns. Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text. This is totally obscured by the procedures of conventional grammar.*

Diante das contribuições da LC para os estudos fraseológicos, serão apresentadas a seguir, as principais definições levantadas por Biber e seus colaboradores, principalmente na gramática baseada em *corpus* do inglês escrito e falado (BIBER *et al.*, 1999), dos diferentes tipos de expressões lexicais existentes. É importante ressaltar que, como em toda área cuja consolidação é relativamente recente, há divergências terminológicas. Há também pouco consenso em relação às características de cada uma das diferentes unidades pré-fabricadas e das metodologias utilizadas para identificá-las (BIBER; CONRAD; CORTES, 2004). A presente seção não tem o intuito de discutir o que diferencia as existentes perspectivas de pesquisas de agrupamentos lexicais, mas sim o de delimitar a terminologia *pacote lexical* adotada neste estudo contrastando-a com outros tipos de fraseologismos, assim como é realizado na gramática supracitada. Os tipos de expressões lexicais apresentados pelos autores e reproduzidos nesta seção são: verbos frasais, verbos preposicionados, expressões idiomáticas, colocações, associações léxico-gramaticais e pacotes lexicais (tradução minha)[27].

Os dois primeiros termos – verbos frasais e verbos preposicionados – referem-se a expressões bastante comuns na língua inglesa, e que constituem uma unidade semântica ou estrutural. Verbos frasais são compostos por um verbo e uma partícula adverbial *e.g.*, *look up*, *carry out*, enquanto que verbos preposicionados são compostos por um verbo e uma preposição *e.g.*, *look at*, *talk about*. Esses dois tipos de expressões lexicais podem aglutinar-se, formando verbos frasais preposicionados *e.g.*, *get away with*.

A maioria dos verbos frasais e uma parte considerável dos verbos frasais preposicionados podem também ser expressões idiomáticas se esses forem expressões relativamente invariáveis, das quais não é possível extrair seu significado dos significados de cada um dos itens que as compõem. O verbo frasal *carry out* e o verbo frasal preposicionado *get away with* encaixam-se nessa categoria, uma vez que o primeiro pode significar *undertake*, *perform*, e o segundo *escape*, e também pelo fato de que ambos os exemplos apresentam uma variabilidade restrita, em relação ao tempo, número e aspecto. Essa variabilidade é restrita, pois é necessário que se mantenham as palavras lexicais de cada expressão para que se possa extrair o significado idiomático de cada uma dessas expressões. Há também expressões idiomáticas mais longas, das quais muitas

[27] *phrasal verbs, prepositional verbs, idioms, collocations, léxico-grammatical associations, e lexical bundles*.

formam um predicado completo e podem ser substituídas por um único verbo lexical como *kick the bucket – die* e *bear in mind – remember*.

As chamadas colocações, por outro lado, são associações entre palavras lexicais que co-ocorrem com mais frequência do que devido ao acaso. Diferentemente das expressões idiomáticas, as colocações são associações estatísticas e não expressões relativamente fixas. O adjetivo *obvious*, por exemplo, no registro acadêmico, co-ocorre com palavras como *difference*, *difficulty*, *challenge*, e *example(s).* Os itens que compõem as colocações apresentam um significado transparente, mas esse significado é construído conjuntamente as suas partes. Algumas palavras como *little* e *small*, muitas vezes reconhecidos como sinônimos, têm o seu sentido construído juntamente aos seus colocados como *baby, devil, kitten(s), bag, dog, kid(s), thing, bit(s), girl(s), lad, while, boy(s), duck(s),* e *man* para *little* e *amount(s), piece, quantities, world, letters, print, sum, part, proportion* e *size* para *small*.

Outro tipo de padrão associativo engloba diferentes estruturas gramaticais e é chamado de associação léxico-gramatical. Verbos como *think* e *want*, por exemplo, co-ocorrem mais frequentemente com orações completivas introduzidas por *that* enquanto que verbos como *like*, *want* e *need* co-ocorrem mais frequentemente com orações completivas introduzidas por *to*.

Por fim, os pacotes lexicais podem ser formados por associações mais longas e podem ser considerados colocações estendidas. Pacotes lexicais comuns na conversação, por exemplo, incluem *do you want me to, I said to him, going to be a*, e *I don't know what*, enquanto que no discurso acadêmico incluem *in the case of the, there was no significant* e *it should be noted that*. Pacotes lexicais diferenciam-se de expressões idiomáticas por essas últimas serem relativamente invariáveis e pelo significado de seus itens ser opaco, além do fato de não serem expressões extremamente frequentes. Por esse motivo, expressões idiomáticas não podem ser automaticamente identificadas por um software da LC da mesma maneira que os pacotes lexicais são gerados (GREAVES; WARREN, 2010). Esses, por sua vez, são as sequências de palavras que mais ocorrem em um dado registro e geralmente não são expressões fixas e não é possível substituí-los por um único item lexical. Há, porém, expressões lexicais que podem ser categorizados tanto como expressões idiomáticas quanto como pacotes lexicais, *e.g., on the face of it*.

Biber *et al.*, (1999) ainda argumentam que a maior parte dos pacotes lexicais não é completa estruturalmente[28]. Os autores afirmam que, apesar disso, pacotes lexicais encaixam-se em tipos estruturais básicos, sendo possível constatar diferenças entre registros. Pacotes lexicais mais comumente encontrados na conversação, por exemplo, são construídos a partir de um sujeito pronominal seguido de um sintagma verbal e do início de uma oração completiva como *I don't know why* e *I thought it was*. Em textos acadêmicos, porém, pacotes lexicais são mais comumente formados por partes de sintagmas nominais e preposicionais como *the nature of the* e *as a result of*.

Além disso, pacotes lexicais estendem-se por unidades estruturais. Biber *et al.,* (1999) puderam constatar em seu estudo que muitos dos pacotes da conversação continham o início de uma oração principal seguida pelo início de uma oração completiva subordinada, cujo próximo espaço disponível seria utilizado para expressar o conteúdo específico para cada situação individual, como demonstram os exemplos do pacote *I don't know why*: *I don't know why he didn't play much at the end of the season* e *I don't know why Catherine finds that sort of thing funny*. Em outras palavras, esses pacotes lexicais funcionam como tijolos – metáfora introduzida por Biber *et a.l,* (1999) – para a construção de unidades estruturais dos tipos verbais e oracionais. A maior parte dos pacotes lexicais do discurso acadêmico[29], por outro lado, incorporam partes de sintagmas nominais e preposicionais, como demonstram os exemplos do pacote *the nature of the* – que se constrói a partir de um sintagma nominal incompleto que, por sua vez, contém um sintagma preposicional subordinado (incluindo o início de um sintagma nominal definido) – e do pacote *as a result of* – formado por um sintagma preposicional incompleto: uma preposição e um sintagma nominal seguidos pelo início de um sintagma preposicional subordinado. Itens complementares a esses pacotes lexicais, citados no estudo, são: *the nature of the physical world, the nature of the issues involved, as a result of his work, as a result of this change*.

Uma crítica quanto à limitação dos estudos de pacotes lexicais relaciona-se ao fato desse tipo de fraseologismo não permitir a identificação de estruturas descontinuadas, como *not only... but also...,* ou de palavras funcionais e seus colocados como *the... of* e *a/an... of* (NESI;

[28] Essa característica foi recentemente contestada em um estudo que encontrou pacotes compostos por estruturas gramaticais completas, orações e até frases (CORTES, 2013).

[29] Os gêneros textuais incluídos no *corpus* do discurso acadêmico da *Longman Grammar of Spoken and Written English* (BIBER *et al.*, 1999) abrangem artigos acadêmicos e excertos de livros, de diferentes áreas científicas como agricultura, biologia, química, etc., e somam mais de 5 milhões de palavras.

BASTURKMEN, 2006 *apud* GREAVES; WARREN, 2010), denominadas *concgrams* (CHENG *et al.*, 2006, 2009 *apud* GREAVES; WARREN, 2010). Acredita-se, porém, que expressões como *not only* e *but also*, por exemplo, separadamente gerariam frequências parecidas e o exame das linhas de concordância poderia comprovar a sua associabilidade. Em relação ao segundo caso, pacotes lexicais do tipo *a/the form of, in terms of the* e *nature of the*, por exemplo, foram identificados como pacotes lexicais do gênero acadêmico em estudos anteriores (SIMPSON-VLACH; ELLIS, 2010) e variações de *the... of* e *a/an... of* poderiam ser identificados em outros gêneros textuais. Por outro lado, há fraseologismos que exibem extrema variação de constituintes, resultando em distâncias maiores entre os colocados, como *think... because*, não tão facilmente passíveis de identificação pela abordagem utilizada para gerar pacotes lexicais, e.g., *but I think that a lot of people say no because of the media press* e *I think that this hasn't been looked at because its male dominated*.[30]

Os pacotes lexicais podem ser categorizados de acordo com padrões sistemáticos, identificados por meio de investigações em *corpora* de grandes proporções. Após a descrição apresentada pelo autor e reproduzida nos parágrafos acima, *Biber et al.*, (1999, p. 989) enfatizam o aspecto lexical da gramática que é revelado pelo estudo de pacotes lexicais, muitas vezes ignorado pelos estudos linguísticos mais tradicionais:

> [...] a gramática não envolve somente o estudo de classes e estruturas abstratas, mas também de palavras específicas e de suas funções específicas dentro dessas classes e [estruturas]. Esse tipo de informação é também importante para aprendizes de inglês como língua estrangeira: a produção natural e idiomática do inglês não é simplesmente uma questão de se construir sentenças bem formadas, mas também de usar expressões lexicais devidamente testadas em contextos apropriados. (tradução minha)[31]

Ainda de acordo com Biber *et al.*, (1999), pacotes lexicais são formados por sequências de três ou mais palavras, uma vez que pacotes menores são muitas vezes incorporados a outros pacotes maiores. Eles citam o pacote de três palavras *I don't think* como exemplo, usado em pacotes de quatro palavras como *but I don't think, well I don't think, I don't think so* e *I don't think I*. Esses pacotes encontram-se portanto em contexto de sobreposição. Como explicitado no capítulo anterior, mais especificamente na seção intitulada justificativa, a presente pesquisa julga

[30] Esses exemplos foram retirados de GREAVES; WARREN (2010).

[31] *[...] grammar is not just a study of abstract classes and structures, but of particular words and their particular functions within those classes and functions. Such information is also important for the learner of English as a foreign language: producing natural, idiomatic English is not just a matter of constructing well-formed sentences, but of using well-tried lexical expressions in appropriate places.*

que uma abordagem que não contabiliza a ocorrência desses tipos de pacote separadamente é parcial, e propõe uma metodologia integral para a análise de pacotes lexicais. Nessa metodologia, todos os pacotes que contém somente três palavras seriam levados em consideração e em seguida todos os pacotes com quatro, cinco, etc., evitando uma contagem inflacionada de cada tamanho de pacote.

Para a identificação de pacotes lexicais, Biber *et al.,* (1999) basearam-se em unidades de palavras ortográficas e itens com contrações foram considerados como uma só palavra. O mesmo critério foi adotado neste estudo, embora o número de contrações encontradas tenha sido muito pequeno devido ao caráter mais formal do discurso pesquisado, o acadêmico. Outra característica já mencionada refere-se ao fato de que uma sequência deve ser considerada um pacote lexical se ocorrer frequentemente. O índice de frequência é estabelecido pelo pesquisador. Além disso, pacotes lexicais devem ser recorrentes no discurso como um todo, e não somente em um registro específico, diferenciando-se, portanto, de idiossincrasias do discurso de um indivíduo. Pacotes lexicais são expressões recorrentes utilizadas por diferentes falantes em diferentes contextos. Portanto, pesquisadores estabelecem em quantos textos diferentes pacotes lexicais devem ocorrer. Essas questões serão retomadas no capítulo de Metodologia.

A partir do detalhamento acerca dos pacotes lexicais descrito acima, a próxima seção buscará apresentar os principais resultados dos estudos mais recentes cujo foco encontra-se na investigação da correlação entre nível de proficiência e uso de pacotes lexicais oriundos da produção de aprendizes de inglês e de nativos no discurso acadêmico. O principal objetivo dessa seção encontra-se em demonstrar como as investigações realizadas sob essa perspectiva poderão contribuir para o ensino e conscientização por parte dos aprendizes das características que constroem o gênero da escrita acadêmica.

2.4 Pacotes lexicais na escrita acadêmica do *continuum* aprendiz-nativo

Os estudos discutidos aqui buscam contrastar o uso de pacotes lexicais, tanto quantitativamente quanto qualitativamente, na produção acadêmica escrita de aprendizes de inglês e na de nativos. Como demonstrado pelas seguintes citações - (HYLAND, 2008, p. 42; PAQUOT; GRANGER, 2012, p. 5; CHEN; BAKER, 2010, p. 44;) respectivamente - essas pesquisas comprovam o papel primordial dos pacotes lexicais na construção do texto acadêmico:

> [...] combinações estatisticamente construídas são familiares aos autores e leitores que fazem uso de um gênero textual específico, e dessa maneira, sinalizam sua participação efetiva em uma dada comunidade de usuários. Por outro lado, a ausência de tais combinações pode revelar a falta de fluência por parte de um novato nessa comunidade. (tradução minha)[32]

> [Ferramentas e métodos da Linguística de Corpus] auxiliaram na identificação de uma ampla variedade de "associações recorrentes de forma e sentido" (cf. Pawley & Syder, 1983) que são comumente utilizadas para organizar o conteúdo de um texto e cumprir funções retóricas tão diversas quanto a de introduzir um tópico, comparar, expressar uma causa, sumarizar e concluir [...], ou envolver o leitor em um processo argumentativo [...]. (tradução minha)[33]

> [...] expressões formulaicas extraídas com base em frequência e encontradas na escrita de nativos expertos poder ser de grande valia para que escritores novatos possam alcançar um estilo mais apropriado de escrita acadêmica, e devem, portanto ser incluídas no currículo de ESL/EFL. (tradução minha)[34]

Para atingir o objetivo proposto pelos estudos citados, pesquisadores baseiam-se em taxonomias funcionais de pacotes lexicais elaboradas por diferentes acadêmicos. A subseção a seguir apresentará três principais delas. Em seguida, serão apontados os principais resultados gerados por estudos baseados em cada uma dessas taxonomias, elencando apenas aqueles que dizem respeito a correlação entre nível de proficiência e uso de pacotes lexicais. Por fim, as principais motivações para a divergência entre os resultados serão apresentadas. As descrições dos modelos para classificação funcional de pacotes lexicais discutidas aqui serão úteis

---

[32] *[...] statistically linked combinations are familiar to writers and readers who frequently use a particular genre, and so come to signal competence participation in a given community of users. In contrast, the absence of such clusters might reveal the lack of fluency of a novice or newcomer to that community*.

[33] *[Corpus linguistics tools and methods] helped identify a whole range of "regular form-meaning pairings" (cf. Pawley & Syder, 1983) that are commonly used to organize the content of a text, fulfil rhetorical functions as diverse as introducing a topic, comparing, expressing a cause, summarizing, and concluding [...], or engage the reader in the argumentation process [...]*.

[34] *[...] frequency-driven formulaic expressions found in native expert writing can be of great help to learner writers to achieve a more native-like style of academic writing, and should thus be integrated into ESL/EFL curricula.*

posteriormente, uma vez que os pacotes lexicais dos dois *corpora* investigados serão contrastados com base em uma das taxonomias apresentadas, após a aplicação da metodologia proposta.

2.4.1 Taxonomias elaboradas para a investigação de pacotes lexicais

Para a elaboração de cada taxonomia a ser apresentada, *corpora* de grandes proporções, representativos do discurso acadêmico escrito e/ou falado, foram utilizados. Desses *corpora*, listas de pacotes lexicais de 4 palavras foram extraídas para os primeiros estudos, e de 3, 4, e 5 palavras para o último. O limite de frequência estabelecido para a seleção de pacotes lexicais variou, para cada estudo, entre 10, 20 e 40 PMW[35], e a distribuição mínima exigida variou entre 10% dos textos, 5 textos diferentes, e aproximadamente 80% das áreas de conhecimento escolhidas no estudo para representatividade do discurso acadêmico.

A primeira taxonomia a ser apresentada foi adotada por Hyland (2008), baseado nas macrofunções linguísticas elaboradas por Halliday (Halliday, 1994 *apud* Hyland 2008). Ele separa os tipos de pacotes lexicais em três categorias, a saber, Direcionados à pesquisa, Direcionados ao texto, e Direcionados ao participante.[36] A primeira categoria abrange os pacotes lexicais que auxiliam os autores dos textos a estruturar suas atividades e experiências do mundo real. A segunda inclui pacotes que relacionam-se à organização do texto e ao significado de seus elementos como uma mensagem ou argumento. A última inclui pacotes focados no autor ou leitor do texto. Cada uma dessas categorias abarca subcategorias, esquematizadas no quadro abaixo. O autor argumenta que a possiblidade de um pacote lexical ser classificado em mais de uma categoria é pequena e que linhas de concordância[37] foram examinadas para que os contextos dos pacotes pudessem auxiliar na determinação de sua função (*vide* Quadro 4).

A segunda taxonomia (*vide* Quadro 5) foi elaborada por Biber *et al.*, (2004), através de uma metodologia indutiva, na qual os pacotes foram agrupados de acordo com suas funções,

---

[35] Do inglês *per million words*.

[36] As categorias e subcategories foram traduzidas a partir do original: *Research-oriented: location, procedure, quantification, description, topic-related; Text-oriented: transition signals, resultative signals, structuring signals, framing signals; Participant-oriented: stance features, engagement features* (HYLAND, 2008, p. 49).

[37] Linhas de concordância são derivadas de um *corpus* por meio de um software desenvolvido para esse fim - um concordanciador. "Concordanciadores tipicamente realçam e centralizam os exemplos encontrados, um por linha, juntamente com o seu contexto à direita e à esquerda." (MCENERY; WILSON, 2001).

usos e significados. Linhas de concordância também foram examinadas para que o uso de cada pacote fosse analisado em seu contexto discursivo. Em seguida, funções discursivas foram associadas a cada grupo. Como na taxonomia anterior, alguns pacotes apresentaram mais de uma função em um contexto discursivo único, ou quando ocorriam em contextos distintos. A maioria dos pacotes, porém, apresentou uma função fundamental.

Quadro 4 - Taxonomia funcional de pacotes lexicais do discurso acadêmico elaborada por Hyland (2008)

Direcionados à pesquisa	**Direcionados ao texto**	**Direcionados ao participante**
Contexto – indicam tempo e lugar (*at the beginning of, at the same time, in the present study*)	**Sinlizadores de transição** – estabelecimento de links aditivos ou contrastivos entre elementos *(on the other hand, in addition to the, in contrast to the)*	**Recursos atitudinais** – transmitem a opinião e avaliação do autor *(are likely to be, may be due to, it is possible that)*
Procedimento – *(the use of the, the role of the, the purpuse of the, the operation of the)*	**Sinalizadores resultativos** – marcam relações inferenciais ou causativas entre elementos *(as a result of, it was found that, these results suggest that)*	**Recursos de envolvimento** – referem-se diretamente ao leitor *(it should be noted that, as can be seen)*
Quantificação – *(the magnitude of the, a wide range of, one of the most)*	**Sinalizadores estruturais** – marcadores textuais que organizam porções do discurso ou direcionam o leitor para outra parte do texto *(in the present study, in the next section, as shown in fig.)*	
Descrição – *(the structure of the, the size of the)*	**Sinalzadores de construção** – situam argumentos a partir da especificação da limitação de condições *(in the case of, with respect to the, on the basis of, in the presence of, with the exception of)*	
Tópico – relacionados ao campo de pesquisa *(in the Hong Kong, the currency board system)*		

Quadro 5 - Taxonomia funcional de pacotes lexicais do discurso acadêmico elaborada por Biber *et al.*, (2004)

<table>
<tr><th>I. Expressões de opinião</th><th>II. Organizadores discursivos</th><th>III. Expressões referenciais</th></tr>
<tr><td>A. Postura epistêmica
Pessoais – (I don’t know if, I think it was)
Impessoais – (are more likely to, the fact that the)</td><td>A. Introdução ao tópico/foco
(what do you think, let’s have a look at)</td><td>A. Identificação/foco
(that’s one of the, one of the most)</td></tr>
<tr><td rowspan="2">B. Atitudinais/postura modal

B1. Desejo
Pessoais – (if you want to, do you want to)
B2. Obrigação/diretivos
Pessoais – (I want you to, and you have to)
Impessoais – (it is important to, it is necessary to)
B3. Intenção/predição
Pessoais – (I’m not going to, what we’re going to)
Impessoais – (It’s going to be, going to be the)
B4. Habilidade
Pessoais – (to be able to, to come up with)
Impessoais – (can be used to, it is possible to)</td><td>B. Elaboração do tópico/clarificação
(has to do with, on the other hand)</td><td>B. Imprecisão
(or something like that, and things like that)</td></tr>
<tr><td rowspan="2">C. Especificação de atributos

C1. Especificação de quantidade
(there’s a lot of, greater than or equal)
C2. Atributos de enquadramento tangíveis
(the size of the, in the form of)
C3. Atributos de enquadramento intangíveis
(the nature of the, in terms of the)</td></tr>
<tr><td>D. Referência temporal/de lugar/textual

D1. Referência de lugar
(the United States and, in the United States)
D2. Referência temporal
(at the same time, at the time of)
D3. Dêitico textual
(shown in figure N, as shown in figure)
D4. Referência multifuncional
(the end of the, the top of the)</td></tr>
</table>

Nessa taxonomia, três grandes categorias denominadas Expressões de opinião, Organizadores discursivos e Expressões referenciais[38], abarcam as principais funções dos pacotes lexicais. A primeira abrange os pacotes lexicais que expressam opiniões ou avaliações de certeza que constroem outra proposição.

A segunda categoria refere-se a pacotes que refletem relações entre discurso anterior e posterior. A última engloba pacotes lexicais que fazem referência direta a entidades físicas ou abstratas, ou ao contexto, tanto para identificar a entidade ou para sinalizar um atributo particular da entidade como algo importante. Essas três categorias abarcam também subcategorias, esquematizadas no Quadro 5.

A última taxonomia a ser apresentada nesta seção foi baseada na taxonomia supracitada, e elaborada por Simpson-Vlach & Ellis (2010). Além da taxonomia, o estudo apresenta uma lista de pacotes lexicais do gênero acadêmico, escrito e falado, denominada *Academic Formulas List* (AFL), gerada a partir de uma metodologia combinatória entre critérios quantitativos e qualitativos, baseados em medidas estatísticas da LC, em análises linguísticas, métricas de processamento psicolinguístico e avaliações de professores de inglês[39]. Os autores também argumentam que muitos dos pacotes apresentam funções múltiplas, mas que as linhas de concordância puderam auxiliar na elicitação daquelas mais salientes. É importante ressaltar ainda que a lista foi criada com fins pedagógicos, principalmente.

Assim como na taxonomia de Biber *et al.*, (2004), esta engloba três principais categorias, a saber, Expressões referenciais, Expressões de opinião e Organizadores discursivos[40]. Algumas das subcategorias diferem-se, uma vez que houve criação, modificação, e unificação de elementos. O Quadro 6 apresenta um esquema dessa taxonomia.

---

[38] As categorias e subcategorias foram traduzidas do original: *stance expressions: epistemic stance, attitudinal/modality stance (desire, obligation/directive, intention/prediction, ability); discourse organizers: topic introduction/focus, topic elaboration/clarification; referential expressions: identification/focus, imprecision, specification of attributes (quantity specification, tangible framing attributes, intangible framing attributes), time/place/text reference.*

[39] Parte da lista AFL foi utilizada na metodologia deste trabalho, e será explorada na seção 3.2 desta dissertação.

[40] As categorias e subcategorias foram traduzidas do original: *referential expressions: specification of attributes (intangible framing attributes, tangible framing attributes, quantity specification), identification and focus, contrast and comparison, deictics and locatives, vagueness markers; stance expressions: hedges, epistemic stance, obligation and directive, expressions of ability and possibility, evaluation, intention/volition, prediction; discourse organizing functions: metadiscourse and textual reference, topic introduction and focus, topic elaboration (non-causal, cause and effect), discourse markers.*

Quadro 6 - Taxonomia funcional de pacotes lexicais do discurso acadêmico elaborada por Simpson-Vlach & Ellis (2010)

A. Expressões referenciais	B. Expressões de opinião	C. Organizadores discursivos
I. Especificações de atributos **a. Atributos de enquadramento intangíveis** *(the concept of, the nature of the)* **b. Atributos de enquadramento tangíveis** *(the frequency of, the sum of)* **c. Especificação de quantidade** *(both of these, the first is)*	**I. Anguladores** *(more likely to be, may not be)*	**I. Referência textual e metadiscursiva** *(in the next section, I'll talk about)*
II. Identificação e foco *(it can be, as an example)*	**II. Postura epistêmica** *(according to the, assume that the)*	**II. Novo tópico e foco** *(what are the, when you look at)*
III. Contraste e comparação *(different from the, is much more)*	**III. Obrigação e diretivos** *(I want you to, you don't need to)*	**III. Elaboração de tópico** **a. Não-causal** *(but this is, any questions about)* **b. Causa e efeito** *(because it is, the reason for)*
IV. Dêiticos e Locativos *(a and b, the real world)*	**IV. Expressões de habilidade e possibilidade** *(can be used to, to use the)*	**IV. Marcadores discursivos** *(in other words, by the way)*
V. Marcadores de imprecisão *(and so on, and so forth)*	**V. Avaliação** *(the importance of, it doesn't matter)*	
	VI. Intenção/volição, predição *(I just wanted to, let me just)*	

2.4.2 Nível de proficiência e uso de pacotes lexicais

Na presente seção, decidiu-se por apresentar os principais resultados – relacionados à correlação entre nível de proficiência e uso de pacotes lexicais – de três trabalhos, cujas taxonomias para investigação de pacotes lexicais na escrita acadêmica apresentadas anteriormente foram adotadas. Um dos objetivos desta dissertação, como explicitado no capítulo anterior, é o de investigar essa correlação.

2.4.2.1 Estudo I: Correlação maior nível de proficiência e menor uso de pacotes lexicais - baseado na taxonomia desenvolvida por Hyland (2008)

O primeiro trabalho, desenvolvido por Hyland (2008), utilizou três *corpora* eletrônicos compostos por artigos acadêmicos, teses de doutorado e dissertações de mestrado cujos textos abarcaram, de maneira proporcional, quatro diferentes disciplinas: engenharia eletrônica, administração, linguística aplicada e microbiologia. As teses e dissertações foram produzidas por aprendizes de inglês de língua materna cantonesa, em sua maioria, e os artigos acadêmicos foram produzidos por acadêmicos experientes. Os *corpora* variaram entre 730.000, 1.900.000 e 825.000 palavras, respectivamente. Pacotes de 4 palavras foram gerados com a utilização do software WordSmith Tools (SCOTT, 1998) para cada um dos três *corpora*, selecionando-se apenas os pacotes com frequência maior ou igual à 20 PMW que ocorressem em, no mínimo, 10% dos textos. As listas geradas foram então classificadas de acordo com a taxonomia apresentada no próprio estudo e contrastadas para que se pudesse encontrar similaridades e discrepâncias entre os pacotes de cada *corpus*.

Em suma, o estudo demonstrou que o número de pacotes lexicais empregados diminui de acordo com o aumento do nível de expertise do autor do texto. A análise dos *corpora* de artigos científicos, teses de doutorado e dissertações de mestrado gerou 71, 95 e 149 tipos de pacotes lexicais diferentes, respectivamente. Além disso, 3,1%, 3,8% e 5,1% dos *corpora* eram compostos por pacotes lexicais. Portanto, tanto a análise de *types* quanto de *tokens* revelou uma correlação entre maior nível de proficiência e menor uso de pacotes lexicais. Da mesma maneira que muitos dos pacotes lexicais encontrados nas dissertações e teses não se mostraram produtivos nos artigos científicos, muitos dos pacotes gerados a partir dos artigos não ocorreram nas dissertações e teses. Por fim, o estudo revelou que muitos dos pacotes que ocorreram nos três *corpora*, foram mais frequentes nos textos dos aprendizes: *on the other hand*, por exemplo, ocorreu o dobro de vezes nas dissertações e o triplo de vezes nas teses quando comparado a sua ocorrência nos artigos.

2.4.2.2 Estudo II: Correlação maior nível de proficiência e menor uso de pacotes lexicais – baseado na taxonomia desenvolvida por Biber *et al.*, (2004) – Correlação encontrada no estudo de Staples *et al.*, (2013)

O segundo trabalho, desenvolvido por Staples *et al.*, (2013), investigou um *corpus* composto pelas respostas escritas a itens do teste *TOEFL Internet-based Test* (TOEFL iBT)[41], totalizando 960 textos e 249.417 palavras. A seção escrita do teste inclui duas diferentes tarefas. Na primeira, o candidato deve ler um excerto, escutar uma passagem sobre um tema, e sintetizar aquelas informações. Na segunda, o candidato deve expressar sua opinião acerca de um tema. O *corpus* do estudo incluiu as duas tarefas de um total de 480 candidatos. Em seguida, os textos receberam notas, baseadas nas pontuações estabelecidas pela *Educational Testing Service* (ETS)[42] que, basicamente classificam a tarefa nos níveis baixo, médio, ou alto. Em seguida, os textos foram divididos em três grupos de acordo com essa classificação. Posteriormente, pacotes de 4 palavras foram gerados para cada um dos três grupos. Somente aqueles que ocorreram em pelo menos dois textos diferentes, e no mínimo 25 vezes por 100 palavras, foram selecionados.

Nesse estudo, aborda-se a problemática em torno de *pacotes lexicais relacionados ao tópico*. Esse tipo de pacote não foi eliminado, uma vez que todos os candidatos receberam os mesmos tópicos. Porém, esses pacotes não foram analisados quanto a sua função, pois não encaixaram-se em nenhuma das três categorias da taxonomia adotada. Além disso, o estudo separa os resultados relacionados ao que denomina *prompt bundles* – pacotes cujas palavras, uma a uma, ocorreram nas instruções das tarefas, e portanto relacionam-se claramente ao tópico. Ressalta-se ainda que nesse estudo, a análise da frequência dos pacotes foi realizada individualmente, ou seja, a produção de cada candidato foi levada em consideração separadamente, permitindo o uso da estatística inferencial.

Os resultados da análise da frequência dos pacotes, incluindo *prompt bundles* e *pacotes lexicais relacionados ao tópico*, demonstraram que houve uma diminuição do uso de pacotes lexicais quando o nível de proficiência aumentava. Esse resultado pode indicar que a linguagem formulaica é um artifício necessário para aprendizes com um nível de proficiência menor. Os autores argumentam que esses aprendizes passam a produzir sequências próprias a medida que seu nível de proficiência aumenta e correlacionam esse fato a estudos de aquisição de segunda língua. Segundo Staples *et al.*, (2013) os resultados desses estudos apontam que sequências desenvolvimentais iniciam-se por processos de memorização e mapeamento de um-para-um de forma e função, e lentamente direcionam-se para uma produção mais aproximada à de nativos

[41] http://www.ets.org/toefl/ibt/about
[42] https://www.ets.org/

(ELLIS, 2006 *apud* STAPLES *et al.*, 2013). Além disso, o estudo demonstrou que *prompt bundles* foram mais utilizados pelos níveis menos proficientes. Quando esses pacotes foram eliminados, o nível intermediário utilizou mais pacotes dentre os três grupos. As ocorrências dos pacotes foram contabilizadas somente em relação aos *tokens*, uma vez que uma análise preliminar realizada revelou que houve muito pouca diferença entre *types* e *tokens*.

2.4.2.3 Estudo III: Correlação maior nível de proficiência e maior uso de pacotes lexicais – baseado na taxonomia desenvolvida por Biber *et al.*, (2004) – Correlação encontrada no estudo de Chen & Baker (2010)

O último trabalho, desenvolvido por Chen & Baker (2010), também foi baseado na taxonomia de Biber *et al.*, (2004), porém, apresentou resultados distintos dos encontrados por Staples *et al.*, (2013) que basearam-se na mesma taxonomia. Os resultados também foram distintos dos escontrados por Hyland (2008). O estudo comparou três *corpora*: o primeiro, composto por redações de aprendizes chineses de inglês, retirados do *British Academic Written English* (BAWE)[43] *corpus*, totalizando aproximadamente 150.000 palavras; o segundo, composto por redações de universitários nativos, também retirados do BAWE, totalizando aproximadamente 155.000 palavras; e o terceiro, composto por textos de acadêmicos, escritores expertos, retirados da seção acadêmica do *Freiburg-Lancaster-Oslo/Berger* (FLOB)[44] *corpus*, totalizando aproximadamente 165.000 palavras. Listas de pacotes de 4 palavras foram geradas, com o auxílio do software WordSmith Tools (SCOTT, 1998), para cada *corpus*, utilizando-se uma ocorrência mínima de, em média, 25 PMW, em, no mínimo, 3 textos diferentes.

Pacotes lexicais relacionados ao tópico foram manualmente excluídos das listas. *Pacotes lexicais em contexto de sobreposição* foram manualmente examinados nas linhas de concordância, e os pacotes de casos de *subsunção completa* e *sobreposição completa* – ambos explorados no capítulo 1, seção 1.2 – foram reduzidos ao pacote maior, evitando, dessa maneira, resultados inflacionados. Esse processo será explorado em detalhes na seção 2.5 deste capítulo. Os resultados do estudo mostraram uma correlação entre maior nível de proficiência e maior uso de pacotes lexicais, tanto para a análise de *types* quanto para a análise de *tokens*, tanto antes quanto após o refinamento da eliminação de *pacotes lexicais relacionados ao tópico* e de *pacotes*

[43] http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/
[44] http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/

lexicais em contexto de sobreposição. Antes da refinação, a ocorrência de pacotes no *corpus* dos expertos foi de 118 (*types*) e 749 (*tokens*); a ocorrência no *corpus* dos aprendizes nativos foi de 120 e 757; e a ocorrência no *corpus* dos aprendizes chineses foi de 90 e 554. Após o refinamento, os dados são 108 e 704; 104 e 667; e 80 e 507. É importante observar que, nesse estudo, o refinamento dos pacotes não provocou mudança na correlação encontrada, referente ao maior nível de proficiência e maior produção de pacotes lexicais.

2.4.3 Principais motivações para a divergência de resultados

O primeiro estudo analisado revelou uma correlação entre maior nível de proficiência e menor uso de pacotes lexicais. Hyland (2008) argumenta que uma possível razão para que muitos dos pacotes lexicais encontrados nas dissertações e teses não terem sido produtivos nos artigos científicos estaria relacionada aos tópicos trabalhados em cada um dos gêneros, como menções a Hong Kong nas dissertações de mestrado, por exemplo. O autor acredita que essa seja uma razão pouco provável, uma vez que essas referências foram raras no *corpus* como um todo. Outra explicação se deve à maior parte dos pacotes lexicais dos *corpora* menos proficientes conter mais sintagmas verbais do que os pacotes do *corpus* mais proficiente, revelando uma maior necessidade de utilizar-se expressões pré-fabricadas por parte de novatos e aprendizes na construção de seus argumentos.

Por fim, o pesquisador argumenta que pacotes lexicais produzidos nos três *corpora* podem ter ocorrido mais vezes nos grupos menos proficientes por ser este um gênero de natureza mais formulaica e por haver a necessidade de demonstrar-se uma atitude mais conciliadora, considerando-se vários pontos de vista na pesquisa, por parte dos estudantes. É importante enfatizar que o autor não relaciona essas diferenças necessariamente à deficiência linguística por parte dos autores cuja primeira língua não é o inglês, e tampouco à falta de habilidades para o domínio das convenções do texto acadêmico em outro idioma. Entretanto, as diferenças existem e podem contribuir para a prática pedagógica, como argumenta Hyland (2008, p. 60):

> Em primeiro lugar, evidências provenientes de *corpora* de aprendizes auxiliam descrições da língua alvo e fornecem modelos mais realistas para estudantes. Essas evidências nos alertam quanto à necessidade de entender-se os tipos de textos que nossos alunos precisam produzir ao invés de recorrermos à robusta literatura existente acerca do gênero artigo científico. Em segundo lugar, um entendimento aprofundado da

> produção de aprendizes elucida todos os aspectos da pedagogia, desde tarefas até o currículo. (tradução minha)[45]

O estudo descrito não elimina e tampouco separa os resultados quanto aos *pacotes lexicais relacionados ao tópico* e aos *pacotes lexicais em contexto de sobreposição*. Além das diferenças metodológicas de cada estudo, essa pode ser uma das razões pelas quais os resultados dos três trabalhos se diferenciam e será testada na presente pesquisa.

O segundo estudo, assim como o primeiro, revelou uma correlação entre menor nível de proficiência e maior uso de pacotes lexicais. Diferentemente do primeiro estudo, porém, essa pesquisa separou os resultados quanto aos *prompt bundles*. Eliminando-se os *prompt bundles*, o grupo com proficiência intermediária, e não mais o grupo com proficiência básica, passou a ser aquele que significativamente mais utilizou pacotes lexicais. Os candidatos menos proficientes utilizaram uma maior quantidade de *prompt bundles*, demonstrando assim uma dependência da utilização de linguagem provida pelo contexto do teste. Por outro lado, os candidatos mais proficientes fizeram uso de linguagem formulaica independente do contexto.

O último estudo, diferentemente do primeiro e segundo, revelou uma correlação entre maior proficiência e maior uso de pacotes lexicais. Umas das razões para essa divergência, levantada pelos pesquisadores do próprio estudo, estaria relacionada ao processo de refinamento dos pacotes lexicais. Os resultados, porém, mantiveram-se os mesmos antes e após as eliminações em relação à correlação pesquisada. Outra razão estaria relacionada ao tamanho dos *corpora* utilizados. Segundos os autores da pesquisa, *corpora* maiores geram menos combinações do que *corpora* menores, mesmo que o limite de frequência estabelecido seja o mesmo.

Em conclusão, os três estudos apresentados neste capítulo diferem-se amplamente quanto à metodologia adotada e uma comparação direta entre seus resultados pode não ser produtiva. A correlação entre nível de proficiência e uso de pacotes lexicais é, portanto, inconclusiva. A presente pesquisa, baseada nos estudos explorados, testará algumas das razões apresentadas para a divergência entre os resultados dos três trabalhos, principalmente a concernente à eliminação

[45] *First, evidence from learner corpora help improve descriptions of the target language and provide more realistic models for students. It alerts us to the need to understand the kinds of text our students need to write rather than rely on the massive literature which describes the research article. Second, an improved understanding of learner output illuminates all aspects of pedagogy from tasks to curriculum […].*

de *pacotes lexicais relacionados ao tópico* e *de pacotes lexicais em contexto de sobreposição*. Antes disso, porém, a próxima seção apresentará metodologias de eliminação desses tipos de pacotes, explicitadas em outros trabalhos.

É importante ressaltar ainda, como última observação desta seção, que alguns estudos revelaram uma tendência de itens fraseológicos – como colocações, verbos frasais e sequências formulaicas – ocorrerem com mais frequência em produções mais proficientes do que em produções menos proficientes (PAQUOT; GRANGER, 2012). Por outro lado, pacotes lexicais, tidos como colocações estendidas, todavia não encaixam-se conclusivamente nesse padrão de análise na literatura atual. Em outras palavras, os resultados de estudos que investigam pacotes lexicais, como explorado anteriormente, demonstram que seria possível correlacionar maior produção de pacotes a níveis menores de proficiência, ou o contrário. Qual seria a motivação para a diferenciação de resultados quando o objeto de investigação varia entre itens fraseológicos como colocações, verbos frasais e pacotes lexicais? A presente pesquisa busca investigar se a eliminação ou manutenção de *pacotes lexicais relacionados ou tópico* e *pacotes lexicais em contexto de sobreposição* podem ser fatores definitivos nessa correlação. A seção a seguir descreve metodologias de eliminação desses tipos de pacotes utilizadas em estudos anteriores.

2.5 Metodologias para eliminação de *pacotes lexicais relacionados ao tópico* e de *pacotes lexicais em contexto de sobreposição*

Esta seção apresenta metodologias de eliminação dos tipos de pacote lexicais supracitados em duas teses de doutorado. A primeira, intitulada *Lexical bundles in scientific English: a corpus-based study of native and non-native writing,* desenvolvida na Universidade de Barcelona por Salazar (2008), e a segunda, intitulada *Lexical bundles across learning development,* desenvolvida na Universidade de Lancaster por Chen (2009).

O primeiro estudo, cujo objetivo principal foi o de criar uma lista de pacotes lexicais de 3, 4, 5 e 6 palavras pedagogicamente relevantes, excluiu uma série de tipos de pacotes, dentre eles fragmentos de outros pacotes e pacotes específicos aos tópicos utilizados[46], além de utilizar

[46] Esses dois tipos de pacotes citados podem ser relacionados aos termos *pacotes relacionados ao tópico* e *pacotes em contexto de sobreposição* adotados nesta pesquisa.

o índice de informação mútua (*MI score – Mutual Information Score*)[47], descartando, dessa maneira, alguns pacotes com alta frequência que não tinham validade pedagógica[48]. Em relação aos fragmentos de outros pacotes, apesar de não utilizar a terminologia criada por Chen & Baker (2010), o estudo diferenciou pacotes lexicais em contexto de *sobreposição completa* e em contexto de *subsunção completa*.

No primeiro caso, os pacotes menores foram incorporados aos pacotes maiores, assim como na metodologia manual adotada como base neste estudo. Portanto, o pacote de 3 palavras *is likely that*, com 66 ocorrências e o pacote de 4 palavras *it is likely that*, também com 66 ocorrências foram unidos e considerou-se apenas o pacote maior. No segundo caso, pacotes do tipo *are consistent with*, com 93 ocorrências, *results are consistent with*, com 28 ocorrências, e *these results are consistent with*, com 21 ocorrências, tiveram suas linhas de concordância examinadas. O exame das linhas de concordância revelou que o pacote *are consistent with* possui outros possíveis colocados como *data*, *findings*, *observations* e *studies*, por exemplo. Diferentemente da metodologia adotada neste estudo, todos estes pacotes foram mantidos, e não houve um refinamento de suas frequências.

No segundo caso, pacotes do tipo *cells were transfected with* e *the x chromossome* foram eliminados por estarem especificamente relacionados aos tópicos dos artigos do *corpus* do estudo. Pacotes foram considerados específicos ao tópico quando encaixavam-se em uma das seguintes descrições: 1) ocorrem em um número limitado de artigos e/ou somente em uma revista específica; 2) sua palavra-chave pode ser encontrada como uma entrada na segunda edição do *Oxford Dictionary of Biochemistry and Molecular Biology*, uma vez que o *corpus* do estudo foi composto por uma amostra de um *corpus* maior de artigos científicos das áreas de biologia, bioquímica, biomedicina e medicina, o *Health Science Corpus*. Outros pacotes foram checados nas linhas de concordância para verificação de seu caráter terminológico na área, e consequente eliminação. Considera-se a estratégia adotada produtiva, uma vez que áreas afins

[47] "A medida estatística *MI score* compara a frequência de uma *multi-word unit* com as frequências gerais de cada um dos dois itens componentes dessa unidade, demonstrando assim que essas palavras co-ocorrem por uma razão, e não somente devido ao acaso" (Church & Hanks, 1990; Manning & Schütze, 1999; Oakes, 1998 *apud* Salazar, 2008, p.59), (minha tradução do original: "*The MI score compares the frequency of a multi-word unit to the overall frequencies of each of its component words, thereby reflecting the likelihood that the two words occur together for a reason and not just by random chance.").*

[48] Indica-se a leitura do estudo de Simpson-Vlach & Ellis (2010) para uma discussão acerca da relevância pedagógica de pacotes lexicais selecionados somente em termos de frequência ou baseados em *MI scores*.

provavelmente produzem pacotes parecidos. Porém, tal abordagem não poderia ter sido utilizada com os *corpora* do presente trabalho, uma vez que não seria possível estabelecer áreas científicas específicas nos mesmos. Todo o processo de exclusão de ambos os tipos de pacotes lexicais foi realizado manualmente.

A segunda tese de doutorado (CHEN, 2009) trabalhou com pacotes de 4 palavras, e trata os *pacotes lexicais em contexto de sobreposição* de maneira detalhada, dividindo-os em três categorias: a) *sobreposição completa*; b) *subsunção completa*; e c) *subsunção parcial*. A primeira categoria é equivalente à categoria *sobreposição completa* adotada na metodologia manual de eliminação adotada neste estudo, *i.e.,* o pacote maior é mantido quando a frequência de seus pacotes menores é exatamente a mesma. Portanto, os pacotes *this may be due* e *may be due to the*, cada um com 6 ocorrências, são eliminados, e mantêm-se o pacote maior *this may be due to the*, com 6 ocorrências.

Já a categoria b) difere-se da utilizada neste trabalho. Nela, *pacotes lexicais em contexto de sobreposição* com ocorrências desiguais são combinados, e mantêm-se a frequência do pacote com mais ocorrências. Logo, o pacote *in the context of*, com 9 ocorrências é combinado ao pacote *the context of the*, com 4 ocorrências, e a representação dessa combinação ocorre da seguinte maneira: *in the contexto of + (the)*, com 19 ocorrências. A metodologia manual utilizada neste trabalho utiliza uma estratégia divergente à essa, de separação e não de união dos pacotes. Dessa maneira, seria verificada a frequência do pacote maior *in the context of the* no *corpus*, e a partir dessa contagem, o número de ocorrências dos pacotes menores seria refinado. Portanto, se o pacote maior *in the context of the* tiver 4 ocorrências, o pacote menor *in the context of*, antes com 19 ocorrências, passaria a ter 15 (19 – 4 = 15) e o pacote *the context of the* não existiria (4 – 4 = 0). Obteríamos então dois pacotes diferentes *in the context of the*, com 4 ocorrências, e *in the context of*, com 15 ocorrências. Percebe-se, que nesse caso, os resultados são equivalentes, mas são obtidos por perspectivas diferentes. Como já explorado no capítulo da introdução, há casos em que mais de dois pacotes se sobrepõem, e seria interessante verificar qual é a vantagem de se obter a frequência dos pacotes separadamente, ou unificadas. O presente trabalho não tem o objetivo de discutir esse aspecto, mas sim de atestar a importância de realizar o processo de análise de *pacotes lexicais em contexto de sobreposição* antes da contagem da frequência dos pacotes em geral.

A última categoria não é utilizada na metodologia manual base para este estudo. Quando o pacote lexical maior não alcança a frequência mínima estabelecida, sua frequência é diminuída da frequência dos outros pacotes que se sobrepõem, para que depois os pacotes sobrepostos sejam unidos. Portanto, o pacote *the end of the*, com 10 ocorrências, e o pacote *at the end of*, com 6 ocorrências, somariam 16 ocorrências. O pacote maior *at the end of the*, porém, não alcançou a frequência mínima utilizada no estudo, de 4 ocorrências, uma vez que apresentou somente 3. Então, diminui-se sua frequência do conjunto da frequência dos pacotes sobrepostos (16 – 3 = 13) e sua representação ocorre da seguinte maneira: (*at)* + *the end of the*, com 13 ocorrências. Essa categoria não foi apresentada no artigo produzido pela pesquisadora no ano seguinte ao da publicação de sua tese de doutorado (CHEN; BAKER, 2010). Na metodologia do presente trabalho, os pacotes *the end of the* e *at the end of* manteriam suas ocorrências de 10 e 6, respectivamente, e não haveria processo de refinação, uma vez que a expressão maior *at the end of the* não seria considerada um pacote lexical, uma vez que essa definição é diretamente relacionada a um número mínimo de ocorrências, estabelecido pelo pesquisador, e baseado em pesquisas anteriores. Um apêndice ainda é adicionado à tese para discutir casos complexos de sobreposição.

Em relação aos *pacotes relacionados ao tópico*, foram eliminados manualmente sequências que continham palavras de conteúdo presentes nas instruções das redações, *e.g., financial and non-financial*, ou qualquer outro pacote relacionado ao tópico, geralmente incorporando nomes próprios, *e.g., in the UK, the Second World War*. Esses pacotes foram eliminados na metodologia base deste estudo da mesma maneira, e quando houve dúvidas, as linhas de concordância foram consultadas.

Através dessa descrição, juntamente com a explanação acerca da metodologia manual adotada neste estudo na seção 1.2 do capítulo 1, é possível perceber que a eliminação de *pacotes lexicais relacionados ao tópico* e de *pacotes lexicais em contexto de sobreposição* é uma tarefa extremamente complexa. O próximo capítulo discorrerá sobre os passos metodológicos adotados nesta pesquisa, e descreverá, em detalhes, uma proposta de metodologia automatizada para a eliminação desses tipos de pacotes. Tal proposta busca, além de estabelecer frequências mais realistas referentes aos pacotes lexicais em questão, permitir que a lista de pacotes a ser investigada pelo pesquisador esteja livre de lixo, e dessa maneira, facilitar a análise tanto

quantitativa quanto qualitativa de pacotes lexicais, permitindo ainda que o pesquisador tome decisões importantes, adequando-as aos seus objetivos do estudo. A aplicação dessa metodologia poderá contribuir para que os estudos que objetivam correlacionar nível de proficiência e uso de pacotes lexicais.

3 METODOLOGIA

O presente capítulo apresenta os passos metodológicos adotados para a elaboração desta pesquisa. Primeiramente, discorre-se sobre os *corpora* e instrumentos utilizados. Posteriormente, apresentam-se os procedimentos de análise e o passo a passo adotado para a elaboração e validação dos *scripts*[49] elaborados para a automatização da metodologia de eliminação de *pacotes lexicais relacionados ao tópico* e de refinação de *pacotes lexicais em contexto de sobreposição*. É importante ressaltar que a validação dos *scripts* permitirá que a metodologia desenvolvida seja implementada em qualquer linguagem de programação por parte de outros pesquisadores que tenham esse interesse.

3.1 *Corpora* do estudo

Foram examinados 2 *subcorpora* dentre os 16 que compõem o *International Corpus of Learner of English Version 2* (ICLEv2)[50], compilado por Granger *et al.*, (2009), composto em sua maioria, por redações argumentativas, somando um total de 3,7 milhões de palavras, de alunos universitários do curso de Letras de 16 línguas maternas diferentes: alemão, búlgaro, chinês, espanhol, finlandês, francês, holandês, italiano, japonês, norueguês, polonês, russo, sueco, tcheco, tsuana e turco. Um terceiro *subcorpus* foi ainda utilizado para que testes da metodologia proposta fossem realizados. Os detalhes dos testes serão apresentados mais adiante.

O primeiro *subcorpus* escolhido é representante dos aprendizes de inglês de língua materna chinesa – parte dele aqui chamada de Ch-ICLE – com um total de 490.787 palavras. O segundo é representante dos aprendizes de inglês de língua materna holandesa – aqui chamado de Dt-ICLE – com um total de 234.813 palavras. Essas escolhas foram baseadas na classificação amostral das redações de cada *subcorpus*, realizada pela equipe do ICLEv2 (*vide* Quadro 7), quanto ao nível de proficiência linguística de cada grupo. A classificação seguiu os parâmetros do *Common*

[49] *Scripts* são compostos por um conjunto de regras (ou algoritmos) desenvolvidas para que o computador realize uma ação X. Algoritmos são comumente comparados a uma receita de bolo. Os *scripts* desenvolvidos para esta pesquisa serão disponibilizados mediante pedidos. Vale ressaltar que esses *scripts* podem ser ainda melhorados, tanto em relação a sua sintetização quanto à otimização de tempo de processamento.

[50] http://www.uclouvain.be/en-277586.html

European Framework of Reference for Languages (CEFR)[51], que define a proficiência em uma língua estrangeira em seis níveis: A1, A2, B1, B2, C1 e C2, sendo C2 o nível mais alto de proficiência. Para essa classificação, a equipe coletou 20 redações de cada *subcorpus* randomicamente[52], e as categorizou nos níveis B2 (e menor), C1 e C2. Como pode ser observado no quadro, o Ch-ICLE teve a maior parte de suas redações classificadas no nível B2 (e menor) enquanto que a maior parte das redações do Dt-ICLE foram classificadas nos níveis C1 e C2.

Quadro 7 - Classificação de 20 redações de cada *subcorpus* do ICLEv2 de acordo com o *Common European Framework of Reference for Languages* (GRANGER et al., 2009)

Língua materna	**B2 (e menor)**	**C1**	**C2**
Alemão	1	12	7
Búlgaro	2	16	2
Chinês	19	1	0
Espanhol	12	8	0
Finlandês	3	8	9
Francês	3	11	6
Holandês	1	11	8
Italiano	10	9	1
Japonês	18	2	0
Norueguês	8	7	5
Polonês	1	12	7
Russo	3	15	2
Sueco	0	14	6
Tcheco	11	9	0
Tsuana	18	0	2
Turco	16	4	0

Com o objetivo de verificar se a escolha era realmente apropriada, ou seja, se poderíamos entender que os grupos eram distintos em termos estatísticos, decidimos fazer uma análise de agrupamentos baseada na classificação apresentada acima. O dendograma evidencia a distância entre os *subcorpora* do ICLEv2, demonstrando a diferenciação estatística entre os dois grupos escolhidos para o estudo (*vide* Figura 2). Como pode ser verificado na figura, os *subcorpora* cuja língua materna é o espanhol e o tcheco, por exemplo, assemelham-se em termos de proficiência,

[51] http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

[52] A equipe do ICLEv2 enfatiza que o número de redações escolhido para a classificação em níveis de proficiência é experimental, e que além disso, apenas um profissional se encarregou dessa tarefa (GRANGER *et al.*, 2009).

assim como os falantes de Russo e Búlgaro. Além disso, o dendograma evidencia três principais grupos semelhantes entre se. O primeiro grupo é constituído pelos falantes de espanhol, tcheco, italiano, russo e búlgaro. O segundo grupo é constituído pelos falantes de norueguês, finlandês, polonês, alemão, holandês, francês e sueco. O terceiro grupo é constituído pelos falantes de japonês, chinês, turco e tsuana. Em contrapartida, os *subcorpra* investigados neste trabalho cuja língua materna é o holandês e o chinês, distanciam-se no dendograma, e demonstram-se distintos estatisticamente. Os *corpora* Ch-ICLE e Dt-ICLE são, portanto, representantes dos grupos de proficiência menor e proficiência maior, respectivamente, nesta pesquisa.

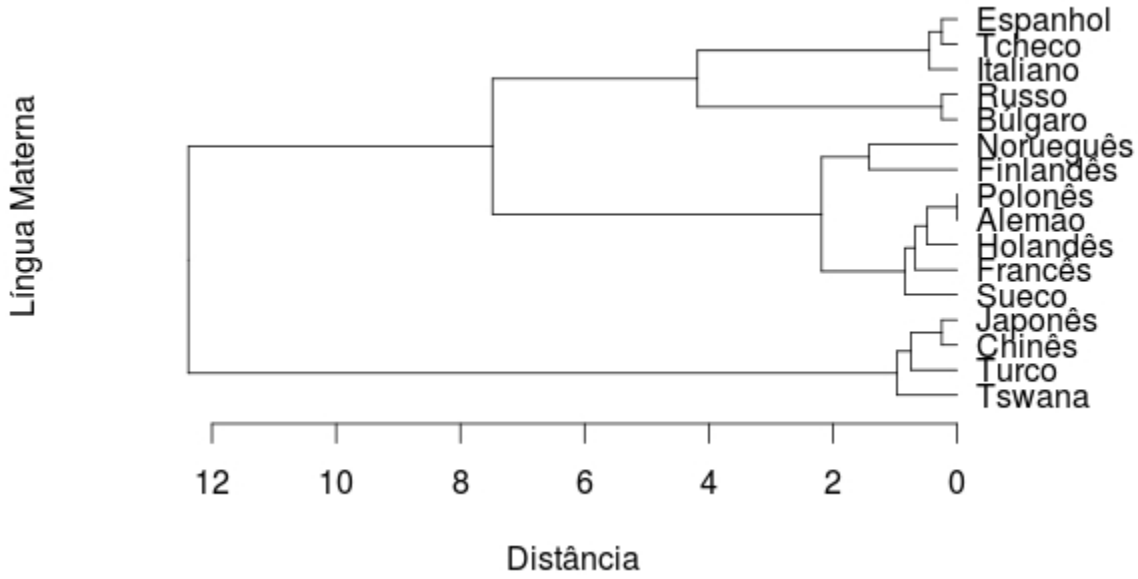

Figura 2 - Dendograma (distância euclidiana) baseado na classificação de 20 textos de cada *subcorpus* do ICLEv2 de acordo com o *Common European Framework of Reference for Languages*

Para que se pudesse controlar as variáveis do estudo, decidiu-se por diminuir o tamanho do *corpus* maior, aproximando-o ao do *corpus* menor. Como mencionado no capítulo anterior, pesquisadores puderam concluir que *corpora* maiores geram menos combinações do que *corpora* menores, mesmo que o limite de frequência tenha se mantido o mesmo (CHEN; BAKER, 2010). Portanto, o tamanho do Ch-ICLE foi reduzido para 234.826 palavras e o Dt-ICLE manteve-se com 234.813 palavras.

Um terceiro *subcorpus*, a saber, o dos aprendizes de inglês de língua materna alemã, com 230.690 palavras, foi escolhido para teste da metodologia desenvolvida, por apresentar um

número similar de número de palavras. Análises manuais foram realizadas utilizando-se esse *subcorpus* para validação dos resultados obtidos pela automatização da metodologia.

3.2 Instrumentos

Caracterizadas pela investigação de dados linguísticos de grandes proporções, as pesquisas da LC beneficiam-se da utilização de software e cálculos estatísticos para uma melhor manipulação dos mesmos[53]. Isso não é diferente em nosso estudo que tem como foco pacotes lexicais que são unidades cuja noção é concebida na sua frequência de ocorrência, como já discutido anteriormente. É possível a criação de *scripts* próprios por meio da linguagem R (R CORE TEAM, 2014), por exemplo, e também utilizar metodologias já implementadas em software fechados, como o Collocate (BARLOW, 2004). Como utilizamos ambos neste estudo, eles serão apresentados a seguir.

O processo de eliminação de *pacotes lexicais relacionados ao tópico* e de eliminação e refinação de *pacotes lexicais em contexto de sobreposição* pôde ser automatizado, com base na metodologia manual desenvolvida por Bohórquez *et al.,* (2012), no software R, com a utilização do editor RStudio[54]. O R pode funcionar tanto como uma calculadora, quanto como um programa estatístico, um programa gráfico, e uma linguagem de programação ao mesmo tempo (GRIES, 2009). Por meio dessa ferramenta é possível personalizar *scripts* com base nas necessidades de cada estudo.

Os *scripts* para a automatização da metodologia de eliminação foram criados em conjunto com um especialista na linguagem R, co-orientador da presente pesquisa. Além do R, o concordanciador Collocate foi utilizado para servir de parâmetro de comparação entre os resultados gerados por ele e pelos *scripts* desenvolvidos. O software Collocate permite que listas de pacotes lexicais sejam geradas, além de possibilitar buscas de itens lexicais e seu respectivo número de ocorrências e outras medidas estatística, como o MI, por exemplo. Outros dois software - AntConc (ANTHONY, 2011) e WordSmith Tools (SCOTT, 1998), similares ao Collocate, foram eventualmente utilizados nas análises para comparação de resultados.

[53] Existe atualmente, portanto, a crescente necessidade de uma formação estatística e de programação por parte dos pesquisadores dessa área. Os programas já existentes nem sempre são capazes de nos ajudar a responder perguntas de pesquisa das mais diversas. Esse foi um dos motivos que levou à criação de uma metodologia nesta dissertação.
[54] www.rstudio.com

A metodologia automatizada, diferentemente da metodologia manual, requer que pacotes lexicais sejam definidos a priori. Para isso, 17,7% dos pacotes lexicais (*vide* Quadro 8) de uma versão da lista *Academic Formulas List* (AFL) (SIMPSON-VLACH; ELLIS, 2010), apresentada no mesmo artigo cuja taxonomia pragmático-funcional foi explorada no capítulo anterior desta dissertação, foi selecionado. Foram feitas buscas a fim de verificarmos as ocorrências desses pacotes lexicais nos *corpora* Ch-ICLE e Dt-ICLE. Em seguida, aplicou-se a metodologia desenvolvida neste estudo para refinar esses pacotes. A lista AFL inclui sequências formulaicas com as seguintes características: a) são padrões frequentes produtivos tanto em *corpora* gerais do discurso oral quanto do discurso escrito em inglês; b) ocorrem significativamente mais no discurso acadêmico do que no discurso não-acadêmico; c) são produtivas em diferentes tipos de gêneros acadêmicos. Como já mencionado anteriormente, a AFL foi elaborada com base em uma metodologia combinatória de critérios quantitativos e qualitativos, abrangendo medidas estatísticas da LC, análises linguísticas, métricas de processamento psicolinguístico e avaliações de professores de inglês.

Neste momento, é importante ressaltar que os pacotes lexicais da lista AFL refletem o uso de um gênero específico que se difere do gênero dos *corpora* investigados nesta pesquisa em relação aos níveis de expertise e tessitura textual. Os pacotes da AFL advêm de *corpora* distintos como o *Michigan Corpus of Academic Spoken English* (MICASE) e o *British National Corpus* (BNC), abarcando palestras, seminários, apresentações de trabalhos de alunos, artigos científicos, livros didáticos, entre outros. Os pacotes dos *corpora* investigados pelo presente estudo, por outro lado, representam, em sua maior parte, redações argumentativas de aprendizes de inglês de diferentes línguas maternas. O presente estudo reconhece que as diferenças entre os *corpora* citados devem ser levadas em consideração na análise dos resultados desta pesquisa.

A versão da AFL utilizada apresenta os 207 pacotes lexicais da lista AFL núcleo, cujos pacotes lexicais são produtivos tanto no discurso acadêmico oral quanto no escrito; os 200 pacotes lexicais prioritariamente orais mais comuns; e os 200 pacotes lexicais prioritariamente escritos mais comuns, categorizados pela taxonomia supracitada. A seleção dos 17,7% da lista respeitou a proporção de distribuição dos pacotes lexicais de cada categoria.

Quadro 8 - Pacotes lexicais da lista AFL selecionados para refinação de *pacotes lexicais em contexto de sobreposição* nos *corpora* Ch-ICLE e Dt-ICLE

A. Expressões referenciais	**B. Expressões de opinião**	**C. Organizadores discursivos**
I. Especificações de atributos **a. Atributos de enquadramento intangíveis** *[a/the] form of* *focus on the* *in relation to* *in the context (of)* *(in) terms of (the)* *the nature of (the)* *the ability to* *the definition of* *the existence of* *the idea that* *the presence of (a)* *the question on* *the study of* *the work of* *(as) a function (of)* *form of the* *(from) (the) point of view (of)* *(in the case (of)* *the kind of* *by virtue of* **b. Atributos de enquadramento tangíveis** *(as) part of [a/the]* *an increase in the* **c. Especificação de quantidade** *a list of* *all sorts of* *a high degree*	**I. Anguladores** *(more) likely to (be)* *a kind of* *appear(s) to be*	**I. Referência textual e metadiscursiva** *come back to* *I was gonna say* *as shown in* *at the outset*
II. Identificação e foco *a variety of* *different types of* *is for the* *is the case* *it does not* *referred to as* *that is the* *that we are* *this type of* *[an/the] example of (a)* *[has/have] to do with* *(as) can be seen (in)*	**II. Postura epistêmica** *according to the* *out that the* *[and/as] you can (see)* *assumed to be*	**II. Novo tópico e foco** *for example [if/in/the]* *a look at*

III. Contraste e comparação *and the same* *the same thing* *be related to the*	**III. Obrigação e diretivos** *do* *you want (me) (to)* *(it should) be noted*	**III. Elaboração de tópico** **a. Não-causal** *but this is* *any questions about* *are as follows* **b. Causa e efeito** *[a/the] result of* *end up with* *as a consequence*
IV. Dêiticos e Locativos *a and b* *(at) the end (of) (the)* *at the time of*	**IV. Expressões de habilidade e possibilidade** *can be used (to)* *(gonna) be able (to)* *allows us to*	**IV. Marcadores discursivos** *and in the* *and if you* *even though the*
V. Marcadores de imprecisão *and so on* *and so forth*	**V. Avaliação** *the importance of* *it doesn't matter* *important role in*	
	VI. Intenção/volição, predição *I just wanted to* *to do so*	

Dessa maneira, dos 433 pacotes da AFL, 77 foram selecionados. Esses 77 pacotes lexicais foram checados quanto à sua ocorrência de *types* e *tokens* nos *corpora* pesquisados e em seguida, foram eliminados e/ou refinados quanto aos pacotes lexicais que se sobrepunham a eles. A ordem de escolha dos pacotes de cada categoria foi randômica, e pelo menos 1 pacote dos tipos prioritariamente oral e prioritariamente escrito foi escolhido. O Quadro 8 mostra os pacotes selecionados de cada categoria. A cor vermelha evidencia os pacotes prioritariamente característicos do discurso oral e a cor verde evidencia os pacotes prioritariamente característicos do discurso escrito.

3.3 Procedimentos de análise

3.3.1 Pré-processamento automatizado dos *corpora* e geração de pacotes lexicais

Primeiramente, foi necessário tratar os três *corpora* utilizados no estudo para que, posteriormente, os pacotes lexicais de cada um deles pudessem ser gerados. Esse tratamento foi

realizado no R. Utilizou-se o pacote tm[55] (HORNIK; FEINERER; MEYER, 2008) para que as alterações pudessem ser realizadas no *corpus* em questão, como um todo. A seguir explicitamos os procedimentos realizados. Primeiramente, eliminamos os códigos de identificação das redações. Cada uma das redações componentes dos três *corpora* apresenta um código de identificação *e.g.,* <ICLE-GE-AUG-0001.1>. As informações presentes nesse código, por exemplo, significam que essa redação faz parte do *subcorpus* alemão – *German subcorpus* – do ICLE (ICLE-GE), que essa redação foi coletada na Universidade de Augsburgo – *University of Augsburg* – (AUG), e que ela é a primeira redação da primeira leva (0001.1). Esses códigos foram eliminados para evitar que seus itens pudessem formar pacotes lexicais. A eliminação dos códigos foi realizada por meio de expressões regulares (JARGAS, 2012) e funções disponibilizadas na biblioteca tm. Em seguida, eliminaram-se todos os espaços em branco extras, presentes nos textos digitados involuntariamente pelos próprios autores dos textos, otimizando, dessa maneira, a rapidez de processamento do software. Além dessas transformações, foi necessário padronizar todas as letras dos textos para minúsculo, evitando que palavras como *Is* e *is*, por exemplo, fossem tratadas de maneira diferente. Uma vez que pacotes lexicais não ultrapassam fronteiras demarcadas por pontuação, foi necessário separar os textos dos *corpora* em trechos delimitados por qualquer caractere de pontuação (. , ! ? etc...) e em seguida a pontuação foi removida. Como exemplo do procedimento, tomemos parte da redação <ICLE-GE-AUG-0001.1>: *The person, whom I admire most is my girl-friend. One reason why I admire her is her beauty.* Essa parte foi dividida em três trechos diferentes, delimitados pelos caracteres de pontuação. Trecho 1: *the person* - *Trecho 2: whom i admire most is my girl-friend* - Trecho 3: *one reason why i admire her is her beauty*. Dessa maneira, os pacotes lexicais foram gerados respeitando o limite de cada trecho.

Para que a metodologia automatizada pudesse ser o mais abrangente possível, pacotes lexicais de 2 a 10 palavras[56] foram gerados para cada um dos *corpora* no R. Para isso, foi estabelecido que qualquer sequência de palavras com frequência igual ou maior que 5 formaria um pacote lexical. Do *corpus* Al-ICLE, foi gerado um total de 8.262 pacotes lexicais, do *corpus* Ch-ICLE, 18.176 e do Dt-ICLE, 9.014. Em seguida, a lista de pacotes lexicais de cada *corpus* foi

[55] A sigla *tm* advém dos termos *text mining*, e pode ser resumida como o processo de extração automática de informações de textos, abrangendo as etapas de organização, transformação e análise textual.

[56] Pacotes lexicais com mais de 5 palavras são extremamente raros. Essa decisão permitiu, porém, que *pacotes lexicais relacionados ao tópico* fossem facilmente identificados. Essa questão será discutida no próximo capítulo.

salva em três arquivos separados, contendo o pacote lexical e o número de sua frequência, como demonstrado na figura 3, a seguir. As listas foram salvas em arquivos separados, pois o tempo de processamento é alto. Espera-se, porém, reduzir esse tempo com a otimização futura dos *scripts*. Os arquivos separados permitem que eles sejam utilizados em outras análises, evitando que o processo de geração de pacotes lexicais tivesse que ser realizado repetidamente. Para a validação desse *script*, checou-se o número de ocorrências de 10 pacotes no programa Collocate (BARLOW, 2004) e os resultados foram comparados aos obtidos no R. Houve 100% de correspondência do número de ocorrências gerado para esses 10 pacotes nos dois meios utilizados[57].

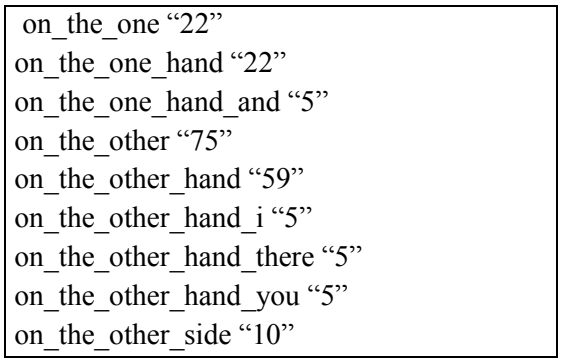

Figura 3 - Visualização de parte da lista de pacotes lexicais gerados no Al-ICLE pelo *script* desenvolvido no R

3.3.2 Eliminação automatizada de *pacotes lexicais relacionados ao tópico*

Para que os *pacotes lexicais relacionados ao tópico* pudessem ser eliminados, os 14 títulos de redações sugeridos pela equipe do ICLEv2 foram coletados. Além desses títulos, instituições de outros países elaboraram temas próprios – alguns deles descritos por Granger *et al.*, (2009). O Quadro 9 reproduz os títulos utilizados para a eliminação de *pacotes lexicais relacionados ao tópico*. Esses pacotes foram eliminados uma vez que não foram naturalmente produzidos por parte dos aprendizes de inglês deste estudo.

Dessa maneira, na metodologia, foi desenvolvido um procedimento para eliminar pacotes lexicais formados por sequências presentes nas instruções dadas para a escrita dos textos. Tomando o primeiro título como exemplo, *i.e.*, *Crime does not pay*, pacotes como *crime does*,

[57] Similaridades e discrepâncias em relação ao número de ocorrências de pacotes lexicais gerados por diferentes software serão discutidas no capítulo de resultados desta dissertação.

crime does not, *crime does not pay*, *does not pay*, *not pay*, e *does not*, foram eliminados. Um problema inevitável nesta fase do desenvolvimento da metodologia foi a eliminação de sequências como *does not*, que não são produzidas somente por estarem relacionadas ao tema *Crime does not pay*. Uma otimização da metodologia desenvolvida deve, futuramente, levar essa questão em consideração. Sequências do tipo *x does not pay*, *x does not*, *not pay x*, onde *x* é um item diferente daquele presente no título, não foram eliminadas.

Quadro 9 - Títulos utilizados para a elaboração de redações nos *subcorpora* do ICLEv2

1. Crime does not pay.
2. The prison system is outdated. No civilised society should punish its criminals: it should rehabilitate them.
3. Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.
4. A man/woman's financial reward should be commensurate with their contribution to the society they live in.
5. The role of censorship in Western society.
6. Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.
7. All armies should consist entirely of professional soldiers: there is no value in a system of military service.
8. The Gulf War has shown us that it is still a great thing to fight for one's country.
9. Feminists have done more harm to the cause of women than good.
10. In his novel Animal Farm, George Orwell wrote "All men are equal: but some are more equal than others". How true is this today?
11. In the words of the old song "Money is the root of all evil".
12. Europe: loss of sovereignty or birth of a nation?
13. In the 19th century, Victor Hugo said: "How sad it is to think that nature is calling out but humanity refuses to pay heed." Do you think it is still true nowadays?
14. Some people say that in our modern world, dominated by science technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?
15. Poverty is the cause of the HIV/AIDS epidemic in Africa.
16. Discuss the advantages and disadvantages of using credit cards.
17. Discuss the advantages and disadvantages of banning smoking in restaurants.

Para a eliminação automatizada de *pacotes lexicais relacionados ao tópico*, os temas descritos acima foram transformados em pacotes lexicais de 2 a 20 palavras, uma vez que algumas sequências relacionadas ao tópico foram formadas por mais de 10 palavras. Em seguida, eliminaram-se os pacotes produzidos dentro do arquivo de pacotes lexicais gerados no Ch-ICLE e no Dt-ICLE. O método apresentado eliminou os chamados *prompt bundles*, conceito explorado por Staples *et al.*, (2013) e descrito no capítulo anterior e também parte dos *pacotes*

relacionados ao tópico. O termo *prompt bundle* refere-se aos pacotes lexicais cujos itens, um a um, estão presentes nas instruções.

Após a eliminação dos *prompt bundles*, foi necessário eliminar os outros *pacotes lexicais relacionados ao tópico:* aqueles que são claramente relacionados ao tópico, mas não necessariamente estão presentes nas instruções das redações, como previsto por Chen & Baker (2010). A estratégia utilizada para esse fim foi de gerar pacotes lexicais de no mínimo 6 e no máximo 10 palavras a partir dos *copora* Ch-ICLE e Dt-ICLE. Uma vez que pacotes lexicais constituídos por mais de 6 palavras são considerados extremamente raros, sequências de até 10 palavras refletiriam o uso de expressões baseadas em fontes externas[58]. A partir dos pacotes criados, um *script* foi desenvolvido para eliminar esses pacotes da lista de pacotes originais do Ch-ICLE e Dt-ICLE.

3.3.3.1 Comparação dos resultados antes e depois da aplicação da metodologia automatizada para a eliminação de pacotes *lexicais relacionados ao tópico*

A metodologia automatizada para a eliminação de pacotes lexicais relacionados ao tópico foi aplicada em dois momentos distintos. Primeiramente, em relação aos *prompt bundles* nas listas de pacotes lexicais geradas para os *corpora* Ch-ICLE e Dt-ICLE. Posteriormente, eliminou-se o restante dos *pacotes lexicais relacionados ao tópico* nessas mesmas listas. O número de pacotes antes e depois das aplicações foi comparado nos dois momentos.

3.3.4 Refinação automatizada de *pacotes lexicais em contexto de sobreposição*: *subsunção completa*

A metodologia manual supracitada inicia o processo de refinamento do número de ocorrências de pacotes lexicais a partir de uma lista de pacotes de 4 palavras, gerada pelo software Collocate (BARLOW, 2004) com um limite de frequência de 20PMW. A presente explicitação apresenta o processo de eliminação manual gradativamente, partindo dos exemplos mais simples até os mais complexos, buscando facilitar sua compreensão por parte do leitor da pesquisa, até que se atinja o nível do processo de eliminação automatizado.

[58] Essa questão será explorada em detalhes no próximo capítulo desta dissertação.

O pacote lexical de 4 palavras mais frequente do Al-ICLE foi *on the other hand*. Esse pacote foi escolhido para exame e consequente elaboração de regras a serem seguidas pelos *scripts* da metodologia automatizada. O processo aqui apresentado foi repetido com 10 outros pacotes lexicais de tamanhos diferentes. O software gerou um total de 59 pacotes *on the other hand*. Gerou-se então, as 59 linhas de concordância referentes a cada uma das ocorrências. O quadro 10 traz um recorte das 10 primeiras linhas de concordância contendo *on the other hand* no Al-ICLE.

Quadro 10 - 10 primeiras de linhas de concordância de *on the other hand* no Al-ICLE

1	*that it would be something terrible.* ***On the other hand*** *it has also been one of the grey little*
2	*alone in the house to go to the phone box.* ***On the other hand*** *you couldn't take a little boy is*
3	*the main destructor of nature, but* ***on the other hand****, we depend on our ecologic system.*
4	*done his best to save the world. But* ***on the other hand*** *pessimists say that it is worthless to*
5	*world will be one of consent. But* ***on the other hand*** *there are some good arguments for the*
6	*are lucky onto the pavement.* ***On the other hand*** *you have to dodge the approaching cyclist,*
7	*you neither wanted nor needed.* ***On the other hand*** *T.V. commercials might be seen as an*
8	*despise them, but never worship them.* ***On the other hand*** *I think that it is certainly*
9	*enormous, horrible noise of traffic but* ***on the other hand*** *I want to consider the following*
10	*your rare, relaxing, spare hours. But* ***on the other hand****, there's that lonely, sharp-tongued,*

Como pode ser observado no Quadro 10, *on the other hand* tem alguns itens que se repetem à direita *e.g.*, *you*, nas linhas 2 e 6, e também à esquerda *e.g.*, *but*, nas linhas 3, 4, 5, 9, e 10. Acredita-se que uma sequência deve ser considerada um pacote lexical quando esta ocorrer em, no mínimo, 5 diferentes textos, sendo o ponto de corte uma decisão deliberada por parte dos pesquisadores (BIBER; CONRAD; CORTES, 2004). Nesta pesquisa, portanto, o ponto de corte para uma sequência ser considerada um pacote lexical foi de 5 ocorrências. Não foi checado, porém, se essas ocorrências eram produtivas em 5 textos diferentes. Como argumentam Biber *et al.*, (2004), essa restrição tem pouco efeito na prática, uma vez que a maioria dos pacotes distribui-se amplamente pelos textos de um *corpus*. Dessa forma, caso *you* se repita, juntamente com *on the other hand*, 5 vezes ou mais, *on the other hand you* seria um pacote lexical de 5 palavras, e as 59 ocorrências do pacote menor *on the other hand* teria que ser refinada. O mesmo ocorreria com o pacote *but on the other hand*. Ambos os pacotes estariam em contexto de *subsunção completa*, explicitado na seção 1.2.

De fato, examinando-se as 59 linhas de concordância de *on the other hand*, os itens à direita *you*, *there*, e *I*, repetem-se exatamente 5 vezes cada um. À esquerda, somente o item *but* se repete 5 vezes ou mais, como demonstra o Quadro 11.

Quadro 11 - Pacotes lexicais relacionados a *on the other hand* em contexto de *subsunção completa*

Pacote Lexical	**Nº de palavras**	**Frequência**
on the other hand	4	59
on the other hand I	5	5
on the other hand you	5	5
on the oher hand there	5	5
but on the other hand	5	14

Além das relações apresentadas, os pacotes de 5 palavras também se sobrepõem: *on the other hand I*, *on the other hand you* e *on the other hand there*, apresentam à sua esquerda o item *but*, uma vez cada. Essa observação levou-nos a tomar a decisão de que os pacotes lexicais devem ser formados à direita, ou à esquerda, independentemente.

A refinação de *on the other hand*, levando-se em conta os pacotes sobrepostos à direita, ocorreria subtraindo-se os pacotes com 5 ocorrências ou mais de seu total (59 – 5 – 5 – 5 = 44). A lista final da análise à direita seria composta por 4 pacotes diferentes*: on the other hand*, com 44 ocorrências, *on the other hand I*, com 5 ocorrências, *on the other hand you*, com 5 ocorrências, e *on the other hand there*, com 5 ocorrências.

Já a refinação de *on the other hand*, levando-se em conta os pacotes sobrepostos à esquerda, ocorreria da seguinte maneira: 59 – 14 = 45. A lista final dessa análise seria composta por 2 pacotes diferentes: *on the other hand*, com 45 ocorrências e *but on the other hand*, com 14 ocorrências.

Uma perspectiva abrangente deve levar em conta os pacotes menores, precedentes de *on the other hand*, como *on the other*, ou *on the*, cuja frequência total também encontra-se inflacionada antes de descontar as ocorrências do pacote de 4 palavras *on the other hand*, por exemplo. Essa perspectiva deve também selecionar os pacotes que ocorrem à direita ou à esquerda de *on the other* e *on the*. Para que isso ocorra, é necessário iniciar o processo do menor pacote possível. Pacotes lexicais são compostos por duas ou mais palavras. No caso de *on the other hand*, suas unidades mínimas de análise são *on the* e *other hand*. Examinando-se cada

unidade para a direita e para a esquerda pacotes do tipo *on the one hand*, *on the contrary*, *on the whole*, e *but on the other hand* poderão ser formados. Define-se então, neste momento, a primeira regra a ser adotada para a elaboração da metodologia automatizada para a eliminação de *pacotes lexicais em contexto de sobreposição* nos casos de *subsunção completa*:

1ª regra: A partir do pacote lexical escolhido para refinação, suas unidades mínimas devem ser examinadas, à direita e à esquerda. Quando o número de repetições de itens à direita ou à esquerda da unidade mínima for igual a ou maior do que 5, forma-se um pacote. Os pacotes menores terão a sua frequência refinada a partir da frequência dos pacotes maiores quando o número de ocorrências dos pacotes relacionados for diferente, o que constitui o contexto de *subsunção completa*.

Para ilustração da 1ª regra, todos os pacotes relacionados às unidades mínimas de *on the other hand* à esquerda e à direita foram examinados manualmente e suas ocorrências foram refinadas, como demonstrados nos Quadros 12 e 13. Essa análise manual é importante também para validar a metodologia proposta, uma vez que o processamento dos pacotes pelo *script* desenvolvido reflete na íntegra o que é descrito na metodologia apresentada.

A metodologia, além de refinar as contagens dos pacotes relacionados a *on the other hand*, sublinhados em ambas as tabelas, encontra todos os pacotes relacionados a *on the* e a *other hand* como *on the contrary*, *on the other side*, e *on the way* e refina suas ocorrências. É possível observar, nessas mesmas tabelas, pacotes com ocorrências iguais antes da refinação, e um deles com frequência 0, após a refinação, como os pacotes nº 6 e 7 no Quadro 12 e 1, 2 e 3 no Quadro 13. Esses pacotes encontram-se em contexto de *sobreposição completa* e serão discutidos na próxima seção. Antes, faz-se necessário acrescentar um detalhe importante à 1ª regra.

Quadro 12 - Pacotes lexicais relacionados à direita da unidade mínima de *on the other hand*

Nº	**Pacote lexical**	**Frequência sem refinação**	**Frequência com refinação**
1	*on the*	502	305
2	*on the contrary*	15	15
3	*on the floor*	11	11
4	*on the ground*	7	7
5	*on the loose*	9	9
6	*on the one*	22	0
7	*on the one hand*	22	17
8	*on the other hand and*	5	5
9	*on the other*	75	6
10	*on the other hand*	59	44
11	*on the other hand I*	5	5
12	*on the other hand there*	5	5
13	*on the other hand yoy*	5	5
14	*on the other side*	10	10
15	*on the pavement*	5	5
16	*on the phone*	5	5
17	*on the screen*	5	5
18	*on the street*	6	6
19	*on the streets*	16	16
20	*on the table*	6	6
21	*on the way*	6	6
22	*on the whole*	9	9

Quadro 13 - Pacotes lexicais relacionados à esquerda da unidade mínima de *on the other hand*

Nº	**Pacote lexical**	**Frequência sem refinação**	**Frequência com refinação**
1	*other hand*	59	0
2	*the other hand*	59	0
3	*on the other hand*	59	45
4	*but on the other hand*	14	14

A partir da observação do Quadro 14, é possível descrever o motivo dessa necessidade. O quadro mostra os pacotes lexicais relacionados a *it is importante to*, à direita. Após o processo de refinação, alguns dos pacotes obtiveram frequência menor do que 5, como os pacotes nº 9 e 11, por exemplo. A sequência nº 9 *it is a wonderful* ocorre somente 3 vezes sem o item *world*. Consequentemente, *it is a wonderful* não é um pacote lexical, pois não respeita a frequência mínima exigida de 5 vezes para ser considerado um pacote. A frequência desses pacotes não pode ser descontada do pacote menor, *i.e.*, a frequência de *it is a wonderful* não pode ser descontada da frequência de *it is*. Por essa razão, a frequência desses tipos de pacote deve ser considerada igual a 0. A 1ª regra redefine-se, portanto, da seguinte maneira:

Quadro 14 - Pacotes lexicais relacionados à direita da unidade mínima de *it is importante to*

Nº	**Pacote lexical**	**Frequência sem refinação**	**Frequência com refinação**
1	*it is*	754	315
2	*it is a*	84	66
3	*it is a brave*	6	6
4	*it is a fact*	7	0
5	*it is a fact that*	7	7
6	*it is also*	20	14
7	*it is also true*	6	6
8	*it is an*	6	6
9	*it is a wonderful*	8	3=0
10	*it is a wonderful world*	5	5
11	*it is better*	7	2=0
12	*it is better to*	5	5
13	*it is certainly*	6	6
14	*it is for*	5	5
15	*it is important*	9	3=0
16	*it is important to*	6	6
17	*it is in*	9	9
18	*it is just*	11	11
19	*it is like*	5	5
20	*it is more*	9	9
21	*it is much*	9	9
22	*it is necessary*	6	6
23	*it is no*	14	14
24	*it is not*	81	64
25	*it is not only*	8	8
26	*it is not the*	9	9
27	*it is now*	6	6
28	*it is of*	5	5
29	*it is only*	11	11
30	*it is possible*	16	10
31	*it is possible to*	6	6
32	*it is probably*	5	5
33	*it is quite*	5	5
34	*it is really*	5	5
35	*it is so*	6	6
36	*it is still*	10	10
37	*it is the*	29	29
38	*it is their*	7	7
39	*it is therefore*	6	6
40	*it is to*	9	9
41	*it is too*	5	5
42	*it is true*	13	3=0
43	*it is true that*	10	10
44	*it is very*	28	15
45	*it is very difficult*	8	1=0
46	*it is very difficult to*	7	7
47	*it is very important*	6	6

1ª regra: A partir do pacote lexical escolhido para refinação, suas unidades mínimas devem ser examinadas, à direita e à esquerda. Quando o número de repetições de itens à direita ou à esquerda da unidade mínima for igual ou maior do que 5, forma-se um pacote. Os pacotes menores terão a sua frequência refinada a partir da frequência dos pacotes maiores quando o número de ocorrências dos pacotes relacionados for diferente, o que constitui o contexto de *subsunção completa*. Quando, após a refinação, a frequência de uma sequência for menor do que 5, ela deve ser considerada igual a 0, para não ser descontada indevidamente do pacote menor.

3.3.5 Refinação automatizada de *pacotes lexicais em contexto de sobreposição*: *sobreposição completa*

Examinando-se os pacotes relacionados às unidades mínimas de *on the other hand*, verificam-se os pacotes *on the one* e *on the one hand*, cada um com 22 ocorrências na análise feita à direita (pacotes nº 6 e 7 no Quadro 12) e os pacotes *other hand*, *the other hand* e *on the other hand*, cada um com 59 ocorrências na análise feita à esquerda (pacotes nº 1, 2 e 3 no Quadro 13). Esses pacotes representam casos de *sobreposição completa*. Tomemos como exemplo os dois primeiros pacotes citados, do Quadro 12. A sequência menor *on the one*, nesse *corpus*, não ocorre nenhuma vez sem o item *hand*, e portanto, não é um pacote completo. Na metodologia, *on the one* é eliminado e por isso sua ocorrência refinada torna-se igual a 0. Da mesma forma, em relação aos pacotes do Quadro 13, as sequências *other hand* e *the other hand* não existem isoladamente, mas somente como *on the other hand*. Desse modo, define-se a 2ª regra, que é dependente da 1ª regra.

2ª regra: Quando os pacotes relacionados às unidades mínimas do pacote escolhido tiverem o mesmo número de ocorrências, o pacote menor deve ser eliminado, o que constitui o contexto de *sobreposição completa*.

A validação dos *scripts* para refinação automatizada de *pacotes lexicais em contexto de sobreposição* tanto nos casos de *sobreposição completa* quanto nos de *subsunção completa* foi realizada checando-se manualmente o contexto de 10 pacotes lexicais diferentes. Houve 100% de correspondência entre os resultados encontrados.

3.3.4.1 Comparação dos resultados antes e depois da aplicação da metodologia automatizada para a refinação de *pacotes lexicais em contexto de sobreposição*

Para a comparação, gerou-se uma lista de cada um dos pacotes lexicais selecionados da AFL utilizando-se o software Collocate (BARLOW, 2004) para os *corpora* Ch-ICLE e Dt-ICLE. Em seguida, aplicou-se a metodologia automatizada para esses mesmos pacotes nos dois *corpora* e os resultados foram contrastados.

4 RESULTADOS

O presente capítulo foi dividido da seguinte maneira: em primeiro lugar, apresenta-se uma análise preliminar do número de pacotes gerados para os *corpora* Ch-ICLE e Dt-ICLE pelo *script* desenvolvido a partir da metodologia proposta neste estudo e por outros software disponíveis. Em seguida, a análise dos resultados oriundos da aplicação da metodologia automatizada para a eliminação de *pacotes lexicais relacionados ao tópico* é discutida. Posteriormente, discorre-se sobre a análise dos resultados advindos da aplicação da metodologia automatizada para refinação e eliminação de *pacotes lexicais em contexto de sobreposição*, tanto em relação aos *types* quanto em relação aos *tokens*.

4.1 Análises preliminares

O número de pacotes lexicais gerados[59] foi sempre maior no *corpus* dos aprendizes menos proficientes nas contagens realizadas antes das refinações e eliminações de pacotes lexicais propostas pela metodologia deste estudo. Esse fato ocorreu tanto para a geração de pacotes feita pelo programa Collocate (BARLOW, 2004) quanto pelo *script* desenvolvido na linguagem R, como demonstrado pelos Quadros 15 e 16. As listas de pacotes lexicais geradas pelo software utilizado são feitas separadamente, para cada tamanho de pacote lexical desejado. O Quadro 15, portanto, mostra as ocorrências de pacotes lexicais com frequência mínima de 5 ocorrências para os *corpora* Ch-ICLE e Dt-ICLE de 2, 3, 4, 5, 6, 7, 8, 9 e 10 palavras. Em contrapartida, o *script* desenvolvido gerou uma única lista com pacotes lexicais de 2 a 10 palavras, com frequência mínima de 5 ocorrências para ambos os *corpora*. Outros software como o AntConc (ANTHONY, 2011) e o WordSmith Tools (SCOTT, 1998), assim como o *script* desenvolvido, possibilitam que uma única lista seja gerada. Os resultados gerados por esses software serão também analisados mais adiante nesta seção, com o objetivo de validar os resultados encontrados pelo *script*.

[59] A análise descrita a seguir refere-se ao número total de *types* de pacotes lexicais produzidos. Na presente etapa do estudo, *types* e *tokens* são equivalentes no sentido de que quando um *corpus* específico produziu mais *types*, o mesmo *corpus* também produziu mais *tokens*. Portanto, acredita-se que a distinção entre *types* e *tokens* não foi necessária nas seções 1.1, 1.2, e 1.3.

Quadro 15 - Contagem de pacotes lexicais de 2 a 10 palavras (*types*) gerados pelo programa Collocate nos *corpora* Ch-ICLE e Dt-ICLE com frequência mínima de 5 ocorrências

Nº de palavras	**Nº de pacotes lexicais no Ch-ICLE**	**Nº de pacotes lexicais no Dt-ICLE**
2	7.302	6.596
3	5.327	2.377
4	2.885	490
5	1.715	92
6	1.197	23
7	909	14
8	727	14
9	578	12
10	468	10
Total	21.108	9.631

Quadro 16 - Contagem de pacotes lexicais de 2 a 10 palavras (*types*) gerados pelo *script* desenvolvido no R nos *corpora* Ch-ICLE e Dt-ICLE com frequência mínima de 5 ocorrências

Nº de palavras	**Nº de pacotes lexicais no Ch-ICLE**	**Nº de pacotes lexicais no Dt-ICLE**
2 a 10	18.125	8.964

Como pode ser observado comparando-se os quadros acima, a soma total das ocorrências dos pacotes gerados pelo programa Collocate (BARLOW, 2004) não coincide com o número de pacotes lexicais gerados pela metodologia proposta. Porém, o padrão de maior produção de pacotes lexicais por parte dos aprendizes de inglês de língua materna chinesa permanece o mesmo em ambas as abordagens. Desse modo, fica demonstrada a equivalência dos resultados obtidos através do software Collocate e da metodologia proposta para geração de pacotes lexicais quanto ao maior número de pacotes produzidos pelos aprendizes menos proficientes. Pode-se concluir, portanto, que o *corpus* Ch-ICLE produziu uma quantidade superior ao *corpus* Dt-ICLE de pacotes lexicais. Esses resultados corroboram os estudos que atestam maior produção de pacotes lexicais aos aprendizes menos proficientes (HYLAND, 2008; STAPLES, 2013). É importante salientar, porém, que esses pacotes lexicais ainda não foram refinados quanto aos contextos de sobreposição ou em relação aos tópicos utilizados nas instruções das redações.

Voltando à problemática da diferença entre a soma das ocorrências dos pacotes lexicais gerados pelo programa Collocate e pela metodologia proposta, ressaltam-se algumas questões. A primeira delas refere-se à diferença de ocorrências de pacotes lexicais encontrada pelos software existentes. O Quadro 17 mostra a frequência de ocorrências de alguns pacotes lexicais escolhidos

aleatoriamente do *corpus* Dt-ICLE gerada pelos software AntConc (ANTHONY, 2011), Collocate (BARLOW, 2004), WordSmith Tools (SCOTT, 1998), e pelo *script* desenvolvido baseado na metodologia proposta.

Quadro 17 - Frequências de pacotes lexicais do *corpus* Dt-ICLE geradas por três diferentes software e pelo *script* desenvolvido

Pacote lexical	**AntConc**	**Collocate**	**WordSmith Tools**	***script* desenvolvido**
advantages and	10	11	10	10
on the other hand	70	70	70	70
it is a	68	69	68	68
that is the	18	20	18	16
the importance of	15	15	15	15

Apesar de não haver diferenças entre os resultados gerados pelos software AntConc e WordSmith Tools para os pacotes escolhidos, os resultados encontrados pelo software Collocate e pelo *script* desenvolvido ora equivalem-se aos resultados gerados pelos outros dois software, ora diferenciam-se. O mesmo ocorre com os resultados obtidos pelo *script* e pelo Collocate quando comparados entre si. Contagens diferentes em relação às ocorrências de itens lexicais com a utilização de diferentes programas parecem ser comuns, como citado por Gries (2009, p. 2):

> [...] quando a ocorrência do item "perl" é checada [...] nos programas AntConc 3.2.1w, WordSmith Tools 4.0, e MonoConc Pro 2.2, utilizando-se a configuração padrão, o AntConc encontra 253 ocorrências enquanto que o WordSmith Tools e MonoConc Pro encontram 248 ocorrências. Os usuários, além de se depararem com o dilema do que fazer com esses resultados divergentes, também precisam compreender a razão pela qual eles se diferenciam, ou melhor, os usuários necessitam compreender de que maneira os programas definem *palavra* e como seria possível alterar suas configurações, etc. (minha tradução)[60]

No caso das diferenças encontradas e exemplificadas no Quadro 17, é possível afirmar que existem evidências estatísticas de que os resultados dos software e do *script* desenvolvido são homogêneos, como pôde ser verificado pelo teste qui-quadrado (χ^2=0,4226, p-value=1,0000). Apesar disso, se todas as ocorrências de todos os pacotes lexicais dos *corpora* trabalhados forem somadas, os resultados gerados por cada software podem diferenciar-se mais

[60] *[...] when you a concordance of the string "perl" [...] with the default setting in the programs AntConc 3.2.1w, WordSmith Tools 4.0, and MonoConc Pro 2.2, then AntConc finds 253 matches whereas WordSmith Tools and MonoConc Pro 2.2 find 248 matches. Users then not only face the problem of what to do with these conflicting results, but are then basically required to figure out why the counts differ or, put differently, how the programs have defined what a word is and how you can change their settings, etc.*

claramente, o que configura a segunda questão. O Quadro 18 mostra o número das ocorrências somadas de pacotes lexicais de 2 a 10 palavras, com frequência mínima igual a 5, nos *corpora* Ch-ICLE e Dt-ICLE, gerados pelo software Collocate (BARLOW, 2004) e o número total de pacotes lexicais com as mesmas configurações citadas acima gerados pelos software AntConc (ANTHONY, 2011), WordSmith Tools (SCOTT, 1998), e pelo *script* desenvolvido.

Quadro 18 - Ocorrências somadas de pacotes lexicais de 2 a 10 palavras com frequência igual a 5 gerados por três diferentes software e pelo *script* desenvolvido

Meio	Ch-ICLE	Dt-ICLE
AntConc	20.834	9.534
Collocate	21.108	9.631
WordSmith Tools	19.294	9.605
***Script* desenvolvido**	18.125	8.964

Como pode ser observado no Quadro 18, nenhum dos quatro meios utilizados para a geração de pacotes lexicais produziu números idênticos. Houve diferença significativa entre o número total de pacotes lexicais gerados por cada um dos quatro meios utilizados, como pode ser verificado pelo teste qui-quadrado (χ^2= 45,026, p-value<0,001). Apesar disso, a relação de maior produção de pacotes lexicais por parte dos aprendizes de inglês menos proficientes do *corpus* Ch-ICLE mantém-se nos quatro meios utilizados.

Diante da variação apresentada, é possível que seja mais vantajoso utilizar software livres como o R, uma vez que não se conhece a fundo os critérios escolhidos para a realização de uma certa tarefa em outros software disponíveis atualmente, como por exemplo, a tarefa de gerar listas de pacotes lexicais de diferentes tamanhos. Ressalta-se, além dessa justificativa, algumas outras exploradas por Gries (2009), e como elas se aplicaram no presente trabalho.

A primeira delas refere-se ao fato de que o esforço despendido na elaboração de um *script* no R é feito uma única vez, e o mesmo *script* pode ser utilizado inúmeras vezes, em diferentes estudos. O *script* desenvolvido para gerar pacotes lexicais, por exemplo, foi utilizado neste trabalho em diferentes *corpora*, e em diferentes momentos, para responder diferentes perguntas. Pequenos ajustes foram realizados para cumprir os diferentes objetivos. Além disso, o presente trabalho pôde ser otimizado com a utilização de pacotes desenvolvidos em outros estudos, disponibilizados pela equipe do R, como foi o caso do pacote tm (HORNIK;

FEINERER; MEYER, 2008). Esse pacote nos permitiu, por exemplo, tratar todas as redações do *corpus* de uma só vez, além de já disponibilizar comandos prontos para a sua preparação. Outra razão a favor do R está ligada ao maior controle por parte de quem desenvolve os *scripts*. Neste trabalho, o conceito de *palavra*, por exemplo, foi por nós definido e uma série de decisões foi tomada para que as eliminações e refinações pudessem ser realizadas.

Além disso, nenhum software disponível atualmente possibilita as eliminações e refinações automatizadas que realizamos com o *script* desenvolvido. Nesse sentido, o R permite que o pesquisador possa desenvolver uma ferramenta customizada que atenda às necessidades de seu estudo. Como discutido no capítulo de metodologia, os *scripts* criados foram baseados em uma metodologia manual para eliminação de pacotes lexicais desenvolvida por Bohórquez *et al.*, (2012) e posteriormente otimizada, também manualmente, para esta pesquisa. O R possibilitou que grande parte da base lógica para essa eliminação pudesse ser automatizada, demonstrando a sua versatilidade para realizar ações tão específicas quanto as que foram realizadas neste trabalho. A reprodução dos *scripts* desenvolvidos é possibilitada pela exemplificação da metodologia manual desenvolvida. Como será explicitado mais adiante, o R nos permitiu vislumbrar resultados antes dificilmente atingíveis através de uma metodologia manual, ou ainda resultados cuja confiabilidade poderia ser considerada menor devido à grande proporção de dados a serem analisados manualmente.

Voltando a análise dos resultados preliminares quanto ao número de pacotes lexicais gerados a partir dos *corpora* Ch-ICLE e Dt-ICLE, além de sempre produzirem mais pacotes lexicais, o grupo menos proficiente produziu uma grande quantidade de pacotes lexicais longos, de até 10 palavras, como pode ser observado no quadro 15. É interessante ressaltar que esse fato pôde ser observado justamente pela diferenciação da maneira de funcionamento do software Collocate (BARLOW, 2004) em relação aos outros programas utilizados neste estudo. O software não permite que uma única lista de pacotes lexicais de 2 a 10 palavras, por exemplo, seja gerada. É necessário que se gere 9 listas diferentes, nesse caso. As listas separadas puderam revelar resultados que nos chamaram a atenção quanto aos pacotes mais longos. A partir desse resultado, as linhas de concordância com pacotes lexicais longos, de 6 a 10 palavras, foram analisados em cada um dos dois *corpora* pesquisados.

O exame dos pacotes lexicais mais longos possibilitou a identificação de dois tipos diferentes no *corpus* Ch-ICLE. Ambos os tipos identificados categorizam-se como *pacotes lexicais relacionados ao tópico*. O Quadro 19 exemplifica os dois tipos de pacotes lexicais longos identificados. O primeiro tipo faz referências a estudos realizados e seus resultados, reproduzindo dados quantitativos que poderiam fortalecer a argumentação dos textos dos aprendizes de inglês de língua materna chinesa. Acredita-se que essas informações foram retiradas de pesquisas que foram utilizadas como fontes para consulta por parte dos aprendizes do Ch-ICLE. O segundo tipo apresentado evidencia pacotes lexicais idênticos à parte das instruções recebidas pelos participantes para a redação do texto, o que configura esses pacotes lexicais como *prompt bundles*, ou similares.

Quadro 19 - Tipos de pacotes lexicais longos, de 6 a 10 palavras, encontrados no Ch-ICLE

Tipo 1: Referência a fontes externas	- *30% of the catering industry's customers are smokers* (22 ocorrências) - *53 bartenders before and after california's prohibition on smoking in* (12 ocorrências) - *a recent survey by scientists at the boston university school* (8 ocorrências) - *a survey conducted by kpmg consulting asia* (27 ocorrências) - *a method which involves reuse of waste materials* (19 ocorências)
Tipo 2: *Prompt bundles* e similares às instruções das redações	- *advantages and disadvantages of banning smoking in restaurants* (43 ocorrências) - *advantage of students using credit card is* (5 ocorrências) *advantages and disadvantages of recycling as a method of waste* (14 ocorrências) - *discuss the pros and cons of importing professionals* (6 ocorrências) - *constructing a second railway link to the mainland* (14 ocorrências)

Ainda em relação ao exame desses pacotes longos, foi possível perceber que alguns deles poderiam ser ainda mais longos se o limite de palavras não tivesse sido estabelecido para o máximo de 10 no *script* desenvolvido. Sequências extremamente longas foram descobertas no exame das linhas de concordância, como por exemplo *breathing secondhand smoke increases the risk of lung cancer and heart disease by about 25%* (de 15 palavras, com 14 ocorrências) e *tobacco-specific carcinogens have been found in the blood and urine of nonsmokers exposed to environmental tobacco smoke* (de 17 palavras, com 10 ocorrências). Essas sequências parecem apresentar uma função de indexação de autoridade a fontes externas. O fato de pacotes extremamente longos terem sido encontrados acima do corte de frequência mínima estabelecido

demonstra que alguns dos aprendizes do Ch-ICLE copiaram esse tipo de informação, exatamente da maneira em que essa informação foi escrita nos textos utilizados para pesquisa.

Por outro lado, o exame dos poucos pacotes lexicais longos do Dt-ICLE permitiu verificar que somente o tipo 2, os pacotes lexicais considerados *prompt bundles*, teve número de ocorrência considerável. O Quadro 20 mostra alguns exemplos desse tipo de pacote encontrado no Dt-ICL. Uma vez que os aprendizes de inglês de língua materna holandesa demonstraram-se totalmente independentes de insumo linguístico do tipo 1, ao menos no que diz respeito aos pacotes longos examinados, parece haver uma maior independência na escrita por parte desses indivíduos.

Quadro 20 - Tipos de pacotes lexicais longos, de 6 a 10 palavras, encontrados no Dt-ICLE

Tipo 2: *Prompt bundles*	- *there is no longer a place for dreaming and imagination* (6 ocorrências) - *television is the opium of the masses* (5 ocorrências)

Os resultados encontrados pela análise de pacotes lexicais longos nos dois *corpora* evidenciam um sobreuso dessas sequências, possivelmente demonstrando uma menor capacidade de argumentação própria, por parte dos aprendizes do Ch-ICLE. As diferenças encontradas entre os resultados do Ch-ICLE e Dt-ICLE mostraram que, em média, o grupo dos aprendizes menos proficientes produz 99,49% mais pacotes longos do que os aprendizes mais proficientes. O fato de que o grupo menos proficiente produziu uma quantidade muito maior de sequências dos tipos de pacotes longos apresentados indica a importância deles serem eliminados, se uma correlação entre proficiência e uso de pacotes lexicais for desejável. A seção seguinte tratará dos resultados da aplicação da metodologia automatizada para a eliminação de *pacotes lexicais relacionados ao tópico*. Primeiramente, discorre-se sobre a eliminação do tipo 2 de pacotes lexicais relacionados ao tópico, *i.e.*, dos chamados *prompt bundles*, e em seguida do tipo 1, *i.e.*, do restante dos *pacotes relacionados ao tópico.*

4.2 Eliminação automatizada de *pacotes lexicais relacionados ao tópico – prompt bundles* – na lista de pacotes lexicais do Ch-ICLE e Dt-ICLE

Supreendentemente, a aplicação da metodologia para a eliminação automatizada de *pacotes lexicais relacionados ao tópico* do tipo *prompt bundles* – pacotes formados por

sequências exatamente iguais a porções das instruções recebidas pelos participantes para a redação dos textos – excluiu mais pacotes no *corpus* Dt-ICLE, representante do grupo mais proficiente, do que no Ch-ICLE, representante do grupo menos proficiente, como pode ser observado no Quadro 21.

Quadro 21 - Resultados da aplicação da metodologia de eliminação automatizada de *pacotes lexicais relacionados ao tópico: prompt bundles*

Corpus	**Nº de pacotes antes da eliminação**	**Nº de pacotes eliminados**	**Nº de pacotes após a eliminação**	**Porcentagem de diminuição de nº de pacotes**
Ch-ICLE	18.125	83	18.042	0,45%
Dt-ICLE	8.946	195	8.751	2,17%

Dos 8.946 pacotes produzidos pelos aprendizes de inglês de língua materna holandesa, 195 *prompt bundles* foram eliminados, o que configura uma diminuição de 2,17% do total de pacotes lexicais, enquanto que dos 18.125 pacotes produzidos pelos aprendizes de inglês de língua materna chinesa, 83 *prompt bundles* foram eliminados, o que configura 0,45% do total de pacotes lexicais. Houve, portanto, uma redução maior de pacotes lexicais no Dt-ICLE do que no Ch-ICLE, estatisticamente verificada pelo teste qui-quadrado (χ^2= 173,02, p-value<0,001). A diminuição desses pacotes foi de 1,31%, em média, nos dois *corpora*. Esses resultados são diferentes e ao mesmo tempo semelhantes aos resultados encontrados pelo estudo explorado nesta dissertação (STAPLES *et al.*, 2013), que também exclui *prompt bundles* para testar a relação entre nível de proficiência e uso de pacotes lexicais.

Em primeiro lugar, os resultados diferenciam-se uma vez que neste estudo, a metodologia utilizada pôde identificar mais *prompt bundles* no *corpus* dos aprendizes mais proficientes, enquanto que no estudo supracitado, foram os aprendizes menos proficientes que produziram mais *prompt bundles*. Isso indica que os aprendizes de inglês de língua materna holandesa se mostraram mais dependentes do insumo linguístico disponível nas instruções das redações do que os aprendizes de língua materna chinesa, neste trabalho. Por outro lado, os resultados assemelham-se uma vez que no presente estudo e na pesquisa citada não foi o grupo dos aprendizes mais proficientes que produziu mais pacotes lexicais em geral, após as eliminações dos *prompt bundles*: aqui, o grupo que apresentou maior produção de pacotes lexicais foi o grupo

menos proficiente, após as eliminações, com um total de 18.042 pacotes contra 8.751 dos aprendizes mais proficientes, e no outro trabalho, o grupo que apresentou maior produção de pacotes lexicais foi grupo intermediário. Os resultados não podem ser comparados mais criteriosamente, pois um estudo comparou somente dois níveis - menos e mais proficiente - enquanto que o outro estudo comparou três diferentes níveis - básico intermediário e avançado. Além dos resultados descritos, é possível concluir que a metodologia de eliminação automatizada de *prompt bundles* não alterou a correlação encontrada até então quanto ao número de pacotes lexicais produzidos: mesmo após as eliminações, o grupo dos aprendizes menos proficiente continuou a produzir mais pacotes lexicais do que o grupo mais proficiente.

Especula-se ainda outra motivação para a diferença encontrada entre os resultados quanto à maior produção de *prompt bundles*. É possível que nem todos os títulos utilizados nas redações dos *corpora* do estudo tenham sido contemplados na eliminação realizada. A equipe do ICLEv2 (GRANGER *et al.*, 2009) sugere uma lista de tópicos a serem utilizados pelas universidades participantes do projeto para a compilação dos *subcorpora*, reproduzida no capítulo de metodologia desta dissertação. Muitos dos coordenadores responsáveis pela compilação em cada país fizeram uso dessa lista, e outros criaram seus próprios temas. Alguns desses temas foram citados pela publicação supracitada, mas não todos. O nosso grupo de pesquisa entrou em contato com as equipes responsáveis pela compilação dos *subcorpora* do ICLEv2 nos diferentes países para que pudéssemos ter acesso às diferentes instruções utilizadas para redação dos textos. Porém, não houve resposta por parte de muitos deles, assim como não houve resposta por parte da equipe responsável pela compilação dos *corpora* Ch-ICLE e Dt-ICLE. Portanto, não se sabe se os grupos do estudo receberam outras instruções, e os *prompt bundles* dessas outras instruções não foram eliminados nesta etapa da pesquisa.

4.3 Eliminação automatizada do restante dos *pacotes lexicais relacionados ao tópico* na lista de pacotes lexicais do Ch-ICLE e Dt-ICLE

Como explicitado no capítulo de metodologia, a estratégia utilizada neste estudo para recuperar e em seguida eliminar os *pacotes lexicais relacionados ao tópico*, além daqueles exatamente iguais a porções das instruções utilizadas para a redação dos textos – os chamados *prompt bundles* – e disponibilizadas pelo ICLEv2 (GRANGER *et al.*, 2009), foi a de gerar uma

lista de pacotes lexicais longos, de 6 a 10 palavras, dos *corpora* Ch-ICLE e Dt-ICLE, utilizando a metodologia proposta para geração de pacotes lexicais. A frequência mínima estabelecida foi também de 5 ocorrências. No Quadro 22, estão relacionados alguns desses pacotes longos, provenientes do *corpus* de aprendizes de inglês de língua materna chinesa.

Quadro 22 - Exemplificação da lista de pacotes lexicais de 6 a 10 palavras geradas do Ch-ICLE

Pacote lexical	**Número de palavras**
10 per cent of a cross-section	6
10 per cent of a cross-section of	7
10 per cent of a cross-section of 88	8
10 per cent of a cross-section of 88 companies	9
10 per cent of a cross-section of 88 companies were	10

Em princípio, essa metodologia não precisaria ter sido aplicada no *corpus* Dt-ICLE, uma vez que ele não apresentou nenhum pacote lexical longo que não fosse um *prompt bundle*. A seção anterior já apresentou a análise das eliminações de *prompt bundles* nos dois *corpora* do estudo. Uma vez que o *corpus* Dt-ICLE não apresentou *pacotes lexicais relacionados ao tópico* do tipo 1, que são aqueles que reproduzem dados quantitativos de pesquisas, os resultados da aplicação da metodologia de eliminação de pacotes lexicais relacionados ao tópico além dos *prompt bundles*, seriam em tese, equivalentes aos resultados da metodologia de eliminação de *prompt bundles*. Decidiu-se, porém, aplicar a metodologia mencionada para verificar se os resultados seriam realmente iguais. Essa questão será explorada mais adiante.

A partir das listas geradas de pacotes longos para cada um dos *corpora*, foi criada uma nova lista de pacotes lexicais de 2 a 10 palavras desses mesmos pacotes longos. Desse modo, os pacotes do quadro acima, por exemplo, foram considerados como tópicos. O "tópico" *10 per cent of cross-section of 88 companies were*, por exemplo, gerou os pacotes exemplificados no Quadro 23.

Com base nessa nova lista de *pacotes lexicais relacionados ao tópico* baseada em pacotes longos, aplicou-se novamente a metodologia proposta para remover os pacotes criados dos *corpora*. Desse modo, foi possível eliminar os pacotes lexicais exemplificados no quadro acima da lista original de pacotes lexicais de 2 a 10 palavras do Ch-ICLE e Dt-ICLE, automaticamente.

Quadro 23 - Exemplificação da lista de pacotes lexicais de 2 a 10 palavras gerados a partir da lista de pacotes lexicais longos, considerados como tópicos, no corpus Ch-ICLE

Pacote lexical	**Número de palavras**
10 per	2
10 per cent	3
10 per cent of	4
10 per cent of a	5
10 per cent of a cross-section	6
10 per cent of a cross-section of	7
10 per cent of a cross-section of 88	8
10 per cent of a cross-section of 88 companies	9
10 per cent of a cross-section of 88 companies were	10
per cent	2
per cent of	3
per cent of a	4
per cent of a cross-section	5
per cent of a cross-section of	6
per cent of a cross-sections of 88	7
per cent of a cross-section of 88 companies	8
per cent of a cross-section of 88 companies were	9
cent of	2
cent of a	3
cent of a cross-section	4
cent of a cross-section of	5
cent of a cross-section of 88	6
cent of a cross-section of 88 companies	7
cent of a cross-section of 88 companies were	8
of a	2
of a cross-section	3
of a cross-section of	4
of a cross section of 88	5
of a cross-section of 88 companies	6
of a cross-section of 88 companies were	7
a cross-section	2
a cross-section of	3
a cross section of 88	4
a cross-section of 88 companies	5
a cross-section of 88 companies were	6
cross-section of	2
...	...

Em outras palavras, a lista de pacotes lexicais longos gerada, que funcionou como tópicos, pôde ser manipulada da mesma maneira que os temas disponibilizados pelo ICLEv2 foram para a eliminação automatizada dos *prompt bundles*. Essa estratégia permitiu que todos os

pacotes longos fossem eliminados: aqueles copiados dos textos fonte utilizados pelos aprendizes e também aqueles adaptados dos textos fonte, uma vez que, em uma sequência como *10 per cent of a total of 88 companies* os pacotes *10 per cent of* a e o pacote *of 88 companies* seriam eliminados.

A aplicação da metodologia diminuiu o tamanho da lista de pacotes lexicais do Ch-ICLE consideravelmente, como pode ser visualizado no Quadro 24. Os aprendizes do Ch-ICLE, que produziram 18.125 pacotes lexicais antes das eliminações, passaram a produzir 10.095 pacotes, evidenciando um decréscimo de 44,30%. Por outro lado, os aprendizes do Dt-ICLE, que produziram 8.964 pacotes lexicais antes das eliminações, passaram a produzir 8.900, evidenciando um decréscimo de 0,71%. Isso demonstra que quase a metade dos pacotes lexicais do *corpus* dos aprendizes menos proficientes é composto por *pacotes lexicais relacionados ao tópico* do tipo 1. A aplicação da metodologia eliminou 8.030 desses pacotes lexicais no Ch-ICLE e apenas 64 no Dt-ICLE. A redução no Ch-ICLE foi significativamente maior do que no Dt-ICLE, verificada pelo teste do qui-quadrado (χ^2= 5437,10, p-value<0,001). Houve uma diminuição de 22,50%, em média, desse tipo de pacote nos dois *corpora*.

Apesar disso, a eliminação automatizada desse tipo de pacotes não inverteu o padrão que relaciona menor proficiência ao maior uso de pacotes lexicais encontrados até então. Os aprendizes de inglês de língua materna chinesa, após as eliminações, apresentaram um total de 10.095 pacotes lexicais e os aprendizes de inglês de língua materna holandesa apresentaram um total de 8.900 pacotes lexicais.

Quadro 24 - Resultados da aplicação da metodologia de eliminação automatizada do restante de *pacotes lexicais relacionados ao tópico*

Corpus	**Nº de pacotes antes da eliminação**	**Nº de pacotes eliminados**	**Nº de pacotes após a eliminação**	**Porcentagem de diminuição de nº de pacotes**
Ch-ICLE	18.125	8.030	10.095	44,30%
Dt-ICLE	8.964	64	8.900	0,71%

Um exame mais cuidadoso da lista de *pacotes lexicais relacionados ao tópico* restantes, após essas eliminações, pôde evidenciar que nem todos eles puderam ser eliminados com a aplicação da metodologia proposta, por dois motivos principais.

Para que se compreenda o primeiro deles, tomemos como exemplo o pacote longo, também considerado como um tópico, *advantages and disadvantages of banning somoking in restaurants*. A partir dele, seria possível eliminar da lista geral de pacotes as sequências *advantages and, advantages and disadvantages, advantages and disadvantages of,* etc. Entretanto, os seguintes pacotes foram encontrados na lista dos pacotes do *corpus* Ch-ICLE, mesmo após a aplicação da metodologia proposta: *advantages and disadvantage* e *advantage of using*. O primeiro exemplo não foi eliminado pela metodologia automatizada, pois o terceiro item do pacote não tem a letra *s* contida no pacote originado da lista de pacotes longos. A metodologia automatizada não pôde eliminá-lo por esse detalhe. Apesar disso, *advantages and disadvantage* não deixa de ser um *pacote lexical relacionado ao tópico*. Algo parecido ocorre com o segundo exemplo. O pacote longo *advantages and disadvantages of banning somoking in restaurants* não poderia produzir pacotes que em um segundo momento eliminaria *advantage of using* simplesmente porque o pacote maior não apresenta essa sequência em sua estrutura. De todo modo, acredita-se que ele também deveria ser eliminado da lista final por ser considerado um *pacote lexical relacionado ao tópico*.

O segundo motivo pode ser contextualizado a partir do pacote *abortion is the,* encontrado na lista após a aplicação da metodologia para eliminação do restante de *pacotes lexicais relacionados ao tópico*. O fato de que nenhum pacote lexical com mais de seis palavras tenha sido produzido sobre o tema *abortion*, impediu que esse pacote de apenas três palavras tenha sido eliminado, assim como os pacotes *abortion should, abortion was, abortion made, abortion is, abortion is a.* Novamente, essas sequências puderam ser claramente caracterizadas como *pacotes lexicais relacionados ao tópico*.

Muitos pacotes dos dois tipos citados acima foram encontrados no Ch-ICLE após a aplicação da metodologia automatizada para a eliminação de *pacotes lexicais relacionados ao tópico*. Isso indica que a lista final de pacotes lexicais do Ch-ICLE diminuiria ainda mais se eles fossem eliminados e a lista de pacotes no Dt-ICLE também. Porém, a metodologia automatizada desenvolvida não é capaz de recuperar esses casos e eliminá-los. Acredita-se que uma limpeza manual seria necessária nessa fase para testar a hipótese de que a eliminação de *pacotes lexicais relacionados ao tópico* pode ser um passo definitivo para correlacionar nível de proficiência e uso de pacotes lexicais.

Outro resultado interessante diz respeito à comparação entre a aplicação da metodologia de eliminação de *prompt bundles* e a aplicação da metodologia de eliminação do restante de *pacotes lexicais relacionados ao tópico* no *corpus* Dt-ICLE. Como explorado na seção anterior, houve uma diminuição de 2,17% de pacotes lexicais após a eliminação de *prompt bundles*. Porém, a diminuição após a eliminação do restante dos *pacotes lexicais relacionados ao tópico* foi de 0,71%. Esperava-se que a porcentagem de diminuição fosse equivalente, uma vez que o Dt-ICLE não apresentou nenhum pacote longo que não fosse um *prompt bundle*. A diferença encontrada entre os dois índices de diminuição parece indicar que alguns dos *promt bundles* do Dt-ICLE não atingiram o tamanho do pacote lexical escolhido para a fase atual de eliminação e continham menos de 6 palavras ou ainda que, como previsto na seção de metodologia, pacotes lexicais menores, de 2 a 3 palavras, não necessariamente *pacotes lexicais relacionados ao tópico*, seriam excluídos pela eliminação automatizada.

Através do exame da lista de pacotes do Dt-ICLE, antes e depois das eliminações dos *prompt bundles*, é possível comprovar as hipóteses mencionadas acima. Para a comprovação da primeira hipótese, tomemos o seguinte tópico como exemplo, previsto pela lista de tópicos sugeridos pela equipe do ICLEv2: *In the 19th century, Victor Hugo said: "How sad it is to think that nature is calling out but humanity refuses to pay heed." Do you think it is still true nowadays?* Ao checarmos a lista final de pacotes lexicais após as eliminações do restante dos *pacotes lexicais relacionados ao tópico*, encontramos o pacote *19th century*. Esse mesmo pacote, porém, não se encontra na lista final após as eliminações dos *prompt bundles*. A estratégia adotada para a presente fase de eliminações não permitiu que o pacote *19th century* fosse eliminado, pois a sequência maior até a fronteira de pontuação *In the 19th century* não possui 6 ou mais palavras. Como outros casos como esse ocorreram, as eliminações de *prompt bundles* foram maiores do que as eliminações do restante de *pacotes lexicais relacionados ao tópico* no Dt-ICLE.

Para a comprovação da segunda hipótese, tomemos o tópico *Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value*. Dele, vários *pacotes lexicais relacionados ao tópico* do tipo *prompt bundle*, de diferentes tamanhos, foram gerados e eliminados da lista geral de pacotes do Dt-ICLE. Inevitavelmente, porém, pacotes como *and do not, do not, and do, for the, they are, are therefore of*, etc., também

foram eliminados por estarem contidos no tópico citado. Verificando os dados, é possível observar que anteriormente às eliminações, o pacote *and do* ocorreu 9 vezes no *corpus* Dt-ICLE. Após a aplicação da metodologia, ele foi eliminado. As 9 ocorrências poderiam, ou não, ter sido originadas da instrução utilizada como exemplo. A checagem dos pacotes do *corpus* dos aprendizes de inglês de língua materna holandesa antes das eliminações, mostrou que esse tema foi produtivo, uma vez que há pacotes lexicais como *university degrees, university education, for the real world*, etc. Examinando-se as linhas de concordância de *and do* no *corpus* Dt-ICLE, reproduzidas no Quadro 25, é possível observar que nenhuma delas relacionam-se ao tema citado.

Quadro 25 - Linhas de concordância do pacote lexical *and do* no *corpus* Dt-ICLE

1	*keep studying only for the exams **and do** not see the importance of a general knowledge. To*
2	*subject of our imagination. We can dream **and do** dream about a holiday we had,*
3	*So, when children watch the right programmes **and do** not watch TV for too long, then it is*
4	*lost his dreams and imagination. so much of what you say **and do**, is part of a dream you*
5	*get caught in this mill of things **and do** not see any way out. Imagination and dreaming are*
6	*because Belgians think they are strange **and do** not belong here. I am totally against this to*
7	*their office to work. No, they can stay at home **and do** everything there on their own*
8	*argue about the usefulness of circumcision **and do** you really need to wear the traditional*
9	*which gives him plenty of time to go **and do** things that add to his knowledge of the country*

Conclui-se, a partir desses resultados, que a metodologia automatizada para a eliminação de *prompt bundles* eliminou mais pacotes do que metodologia automatizada para eliminar o restante dos *pacotes lexicais relacionados ao tópico*. Isso ocorreu pois a metodologia para a eliminação de *prompt bundles* baseia-se em pacotes gerados a partir de tópicos longos, maiores do que 6 palavras, enquanto que a metodologia para a eliminação do restante dos pacotes lexicais relacionados ao tópico baseia-se em pacotes de 6 a 10 palavras. Portanto, a primeira metodologia elimina mais pacotes do que a segunda, uma vez que mais combinações são possíveis a partir de sequências mais longas.

A próxima seção discorrerá sobre a análise dos dados referentes ao total de pacotes lexicais produzidos pelo Ch-ICLE e Dt-ICLE em relação aos *types*, antes e depois da aplicação da metodologia de eliminação de p*acotes lexicais em contexto de sobreposição completa*. A seção seguinte a essa apresentará a análise dos dados referentes ao total de pacotes lexicais produzidos pelos *corpora* em relação aos *tokens*, comparando-se também os resultados obtidos

antes e depois da aplicação da metodologia de refinação para *pacotes lexicais em contexto de subsunção completa*.

4.4 Eliminação automatizada de *pacotes lexicais em contexto de sobreposição completa* - análise dos *types* produzidos

As contagens apresentadas nesta e na próxima seção foram realizadas com os resultados obtidos pelo *script* desenvolvido no R, baseado na metodologia proposta, para a geração de pacotes lexicais. Os pacotes escolhidos para as análises, como explicitado na seção de metodologia, foram retirados da lista AFL e somam 77 pacotes, produtivos em todas as categorias da lista, o que configura 17,7% do total de pacotes da lista AFL. A análise da produção dos *types* antes e depois da aplicação da metodologia nos dois *corpora* do estudo corresponde a eliminações de *pacotes lexicais em contexto de sobreposição* do tipo *completa*, uma vez que quando um pacote representante de um *type*, ou de uma categoria, é eliminado, ele passa a ter frequência 0, pois descobre-se que existe um pacote maior, como foi o caso do pacote *(in) terms of (the)* e *terms of time and energy*, que será explicitado mais adiante.

A análise da produção dos *types* revelou que, mesmo antes da aplicação da metodologia, o grupo mais proficiente produziu mais pacotes lexicais, enquanto que o grupo menos proficiente produziu menos pacotes lexicais. De um total de 77 tipos de pacotes, o Ch-ICLE produziu 26, e o Dt-ICLE produziu 42. Esse resultado pode estar relacionado ao fato de que os pacotes escolhidos para análise advêm de *corpora* mais expertos, atestando dessa maneira, maior produção desses tipos de pacotes por parte do grupo mais proficiente deste estudo. Além disso, esses resultados se opõem aos dos estudos que relacionam menor nível de proficiência e maior uso de pacotes lexicais (HYLAND, 2008; STAPLES *et al.*, 2013). Não é possível, porém, estabelecer neste momento do estudo que a metodologia de eliminação e refinação é um fator definitivo para correlacionar nível de proficiência e uso de pacotes lexicais.

A comparação entre as contagens realizadas antes e depois da aplicação da metodologia automatizada para refinação e eliminação de *pacotes lexicais em contexto de sobreposição completa* pode ser verificada nas análises a seguir. Em primeiro lugar, o *corpus* Ch-ICLE, que anteriormente apresentou 26 *types*, o que corresponde à 33,76% dos 77 pacotes escolhidos da AFL, passou a apresentar 24,6% dos pacotes escolhidos, em média. Houve, portanto, uma diminuição de 26,91%, em média, após a aplicação da metodologia. Pelo teste qui-quadrado (χ^2=

0,0977, p-value<0,7546), conclui-se que as porcentagens de redução equivalem-se para as análises realizadas à direita e à esquerda. O Quadro 26 resume os resultados apresentados. Para as médias descritas, levou-se em consideração as análises feitas à direita e à esquerda dos pacotes lexicais pesquisados.

Quadro 26 - Contagem de *types* após a aplicação da metodologia de eliminação automatizada para eliminação e refinação de *pacotes lexicais em contexto de sobreposição* no *corpus* Ch-ICLE, anteriormente com 26 *types*

***Corpus* Ch-ICLE**	**Nº de *types* eliminados**	**Nº de *types* posteriormente à eliminação**	**Porcentagem de diminuição de nº de *types***
direita	6	20	23,07%
esquerda	8	18	30,76%
média	7	19	26,91%

Já o Dt-ICLE, que apresentou 42 *types* antes da aplicação da metodologia, o que corresponde à 54,54% dos 77 pacotes escolhidos da AFL, passou a apresentar 38, o que corresponde à 49,35% dos pacotes escolhidos, tanto para a análise à esquerda quanto para a análise à direita. Houve, portanto, uma diminuição de 9,52% de *types* após a aplicação da metodologia, como pode ser observado no Quadro 27.

Quadro 27 - Contagem de *types* após a aplicação da metodologia de eliminação automatizada para eliminação e refinação de *pacotes lexicais em contexto de sobreposição* no *corpus* Dt-ICLE, anteriormente com 42 *types*

***Corpus* Dt-ICLE**	**Nº de *types* eliminados**	**Nº de *types* posteriormente à eliminação**	**Porcentagem de diminuição de nº de *types***
direita	4	38	9,52%
esquerda	4	38	9,52%
média	4	38	9,52%

É importante salientar que a metodologia diminuiu os *types* dos pacotes utilizados pelos dois grupos de aprendizes de inglês, Ch-ICLE e Dt-ICLE, em 18,21%, em média. Portanto, um pacote de uma categoria x que era produtivo em um *corpus*, após as eliminações passou a não existir. Como exemplo, tomemos o pacote *(in) terms of (the)*, do Ch-ICLE, que apresentou 12 ocorrências antes da aplicação da metodologia à direita, e passou a ter 0 ocorrências após a aplicação. Isso aconteceu pois a aplicação da metodologia revelou que dessas 12 ocorrências, 8

são do pacote lexical *terms of time and energy* e as 4 restantes não atingiram a frequência mínima exigida para serem consideradas um pacote lexical. O pacote *(in) terms of (the)*, portanto, não existe no *corpus* pesquisado. O pacote lexical *(in) terms of (the)* faz parte da categoria A1a da AFL (Expressões Referenciais - Especificações de atributos - atributos de enquadramento intangíveis), uma das categorias mais produtivas em *corpora* de nativos (DUTRA; BERBER-SARDINHA, 2013; SIMPSON-VLACH; ELLIS, 2010). O pacote maior descoberto *terms of time and energy* poderia também ser classificado nessa mesma categoria? Se a resposta for negativa, a mudança trazida pela metodologia é importante para os estudos que contrastam a produção de pacotes lexicais por parte de aprendizes e nativos baseada em categorias desses itens.

Tentando responder à pergunta feita acima, uma observação se faz pertinente neste momento. O pacote *terms of time and energy* encontrado no *corpus*, após a aplicação da metodologia, com 8 ocorrências, parece relacionar-se ao tópico de reciclagem, evidenciado pela ocorrência do pacote longo relacionado ao tópico *advantages and disadvantages of recycling as a method of waste*, exemplificado na primeira seção do presente capítulo. Uma vez que a perspectiva deste trabalho é a de que *pacotes relacionados ao tópico* devem ser eliminados da contagem total de uma análise que busca correlacionar nível de proficiência e uso de pacotes lexicais, o pacote *terms of time and energy* não deve ser entendido como um pacote pertencente à categoria A1a da AFL. Esse tipo de pacote não é considerado advindo de uma produção natural. É possível ainda que esse pacote tenha sido reproduzido de uma fonte de consulta utilizada pelos aprendizes para a escrita do texto, evidenciando o uso do item *terms of* por parte, possivelmente de um nativo. Portanto, *terms of time and energy* só apareceu com frequência no *corpus* do Ch-ICLE pois ele pode ter sido copiado de um texto fonte de um nativo. A metodologia proposta neste estudo é capaz de revelar essa informação, uma vez que ela descobre o pacote maior *terms of time and energy* que encontra-se em contexto de sobreposição com o pacote da AFL *(in) terms of (the)*. Se os aprendizes de inglês de língua materna chinesa tivessem produzido o pacote *(in) terms of (the)* seguido de itens variados, como ocorre na produção natural de nativos, além de descobrirmos o pacote *terms of time and energy* após a aplicação da metodologia, outros pacotes *(in) terms of (the)* alcançariam a frequência mínima de 5 ocorrências e a categoria A1a seria produtiva no *corpus*. Porém, isso não ocorreu.

Há ainda outra circunstância que explica a diminuição do número de *types* produzidos e revelada pela aplicação da metodologia. O pacote da AFL *(in) the case (of)*, também da categoria A1a, antes da aplicação da metodologia, ocorreu 9 vezes. Após a aplicação da metodologia à direita, esse número diminuiu para 0, pois o pacote produtivo, na verdade, foi *the case that*, com 5 ocorrências. Da mesma maneira do caso do pacote descrito anteriormente, o restante das sequências não alcançou a frequência mínima estabelecida para formar um pacote. A partir dos resultados alcançados, pode-se chegar a algumas conclusões. Como *(in) the case (of)* é diferente de *the case that*, podemos concluir que os aprendizes de inglês de língua materna chinesa produzem um pacote parecido ao dos nativos, mas não idêntico ao previsto pela AFL, indicando sua condição de não expertos no que tange a redação de textos acadêmicos em inglês. Da mesma maneira, é possível questionar se *(in) the case (of)* e *the case that* poderiam ser classificados da mesma maneira. Em princípio, a resposta poderia ser negativa, uma vez que esse pacote não é previsto na AFL e pode revelar um uso não experto, configurando *the case that* como um pacote de aprendiz. Porém, essa questão deve ser investigada com mais cuidado.

Em suma, os resultados descritos nesta seção mostraram que houve uma diminuição maior de *types* no *corpus* dos aprendizes menos proficientes de inglês após a aplicação da metodologia - diminuição de 26,91%, em média - do que no *corpus* dos aprendizes mais proficientes - diminuição de 9,52%, em média. Neste estudo, isso parece indicar que os aprendizes menos proficientes utilizam mais formas parecidas aos pacotes lexicais da AFL, ou mais repetições de formas desses pacotes, mas na verdade, quando os pacotes sobrepostos são levados em consideração, é possível verificar que eles são também relacionados ao tópico, ou apresentam uma composição de aprendiz, diferente daquela prevista pela AFL.

4.5 Eliminação e refinação automatizada de *pacotes lexicais em contexto de sobreposição completa* e de *pacotes lexicais em contexto de subsunção completa* - análise dos *tokens* produzidos

As eliminações dos *tokens* englobam tanto os *pacotes lexicais em contexto de sobreposição completa* quanto os *pacotes lexicais em contexto de subsunção completa* uma vez que realiza-se a contagem da soma dos *tokens* dos pacotes lexicais que antes apresentavam uma

frequência x e passam a apresentar uma frequência 0, como os pacotes que apresentavam uma frequência x e passam a apresentar uma frequência x - y (x menos y).

Assim como os resultados da análise de *types* revelaram, os resultados da análise de *tokens* demonstraram que antes da aplicação da metodologia de refinação de *pacotes lexicais em contexto de subsunção completa*, o grupo mais proficiente produziu mais pacotes lexicais. Os aprendizes do Ch-ICLE produziram 745 *tokens* dos 77 tipos de pacotes lexicais escolhidos da AFL e os aprendizes do Dt-ICLE produziram 1.167 *tokens*.

Quadro 28 - Contagem de *tokens* antes da aplicação da metodologia automatizada para eliminação e refinação de *pacotes lexicais em contexto de sobreposição completa* e de *pacotes lexicais em contexto de subsunção completa* nos *corpora* Ch-ICLE e DT-ICLE

Ch-ICLE	**Dt-ICLE**
745	1.167

Novamente, acredita-se que esses resultados indicam a maior proficiência por parte dos aprendizes do Dt-ICLE, uma vez que os aprendizes de inglês de língua materna holandesa usaram mais pacotes lexicais da AFL, uma lista de pacotes produtivos na escrita acadêmica de nativos. Esses resultados diferenciam-se dos estudos que correlacionam maior produção de pacotes lexicais a níveis de proficiência mais baixos (HYLAND, 2008; STAPLES *et al.*, 2013).

Uma comparação dos resultados antes e depois da aplicação da metodologia para o Ch-ICLE é apresentada no Quadro 29. Os aprendizes de inglês de língua materna chinesa, que anteriormente produziram 745 *tokens*, passaram a produzir 402 *tokens*, em média, evidenciando uma diminuição de 46,03%, em média. Dessa vez, as análises à esquerda e à direita demonstraram-se diferentes pela aplicação do teste qui-quadrado (χ^2= 8,777, p-value<0,003). Essas diferenças serão investigadas em estudos futuros.

Já os aprendizes de inglês de língua materna holandesa, que apresentaram 1.167 *tokens* antes da aplicação da metodologia, passaram a apresentar 819 *tokens*, em média. Houve uma diminuição de 29,81%, em média, de pacotes lexicais em relação aos *tokens* no *corpus* do Dt-ICLE, como pode ser observado no Quadro 30. Novamente, não existem evidências estatística de diferenças entre as reduções causadas pelos métodos à direita e à esquerda, como pôde ser verificado pelo teste qui-quadrado (χ^2= 0,73, p-value<0,3900).

Quadro 29 - Contagem de *tokens* após a aplicação da metodologia automatizada para eliminação e refinação de *pacotes lexicais em contexto de sobreposição completa* e de *pacotes lexicais em contexto de subsunção completa* no *corpus* Ch-ICLE, anteriormente com 745 *tokens*

Corpus **Ch-ICLE**	**Nº de *tokens* eliminados**	**Nº de *tokens* posteriormente à eliminação**	**Porcentagem de diminuição de nº de *tokens***
direita	372	373	49,93%
esquerda	314	431	42,14%
média	343	402	46,03%

Quadro 30 - Contagem de *tokens* após a aplicação da metodologia automatizada para eliminação e refinação de *pacotes lexicais em contexto de sobreposição completa* e de *pacotes lexicais em contexto de subsunção completa* no *corpus* Dt-ICLE, anteriormente com 1.167 *tokens*

Corpus **Dt-ICLE**	**Nº de *types* eliminados**	**Nº de *types* posteriormente à eliminação**	**Porcentagem de diminuição de nº de *types***
direita	338	829	28,96%
esquerda	358	809	30,67%
média	348	819	29,81%

É possível verificar que a aplicação da metodologia diminuiu as ocorrências dos pacotes em 41,88%, em média, nos dois *corpora*, eliminando alguns pacotes que não existiam isoladamente e refinando a contagem de outros.

Assim como para a análise de *types* após a aplicação da metodologia, houve uma diminuição maior de *tokens* no *corpus* dos aprendizes de inglês de língua materna chinesa – 53,95% – do que no *corpus* dos aprendizes de inglês de língua materna holandesa – 29,81% –, em média. Esses resultados ecoam os resultados encontrados a partir das eliminações realizadas na seção anterior. Como já mencionado, houve uma diminuição maior de pacotes que eram produtivos em uma categoria e passaram a não ser mais produtivos na mesma no *corpus* Ch-ICLE do que no *corpus* Dt-ICLE, evidenciando, ao mesmo tempo, cópias de porções de textos utilizados para pesquisa por parte dos aprendizes de inglês de língua materna chinesa e pacotes lexicais não previstos pela AFL. Há evidências, portanto, que a análise da produção de *tokens* corroborou os resultados encontrados pela análise de *types* uma vez que houve uma porcentagem maior de diminuição de pacotes no *corpus* Ch-ICLE do que no *corpus* Dt-ICLE, após a aplicação da metodologia.

5 CONCLUSÃO

O capítulo final desta dissertação busca, em primeiro lugar, apresentar as principais contribuições que a presente pesquisa pôde trazer para os estudos acerca dos pacotes lexicais no contexto da escrita acadêmica no que tange a correlação entre a produção desses itens e o nível de proficiência de quem escreve os textos. Em um segundo momento, as principais limitações do trabalho serão apontadas. Por fim, serão apresentados os possíveis desdobramentos desta pesquisa.

Para alcançar os propósitos do presente capítulo, retomemos os objetivos e perguntas de pesquisa apresentados no capítulo de introdução desta dissertação. Quanto ao objetivo principal de investigar a correlação entre a produção de pacotes lexicais e o nível de proficiência linguística, foi possível chegar a algumas conclusões importantes. Em primeiro lugar, sem nenhum tipo de eliminação e em relação aos pacotes lexicais em geral, foi possível verificar que os aprendizes menos proficientes do *corpus* Ch-ICLE produziram mais pacotes lexicais do que os aprendizes mais proficientes do *corpus* Dt-ICLE, tanto em relação aos *types* quanto aos *tokens*, corroborando estudos anteriores que correlacionam menor nível de proficiência à maior produção de pacotes lexicais (HYLAND, 2008; STAPLES *et al.*, 2013).

O estudo comprovou, porém, que grande parte dos pacotes lexicais encontrados no *corpus* dos aprendizes menos proficientes eram pacotes longos, de 6 palavras ou mais. O exame desses tipos de pacote revelou que os aprendizes menos proficientes tiveram acesso a textos para consulta e reproduziram informações retiradas desses textos em suas redações, ou ainda utilizaram o insumo linguístico presente nas instruções para elaborar seus argumentos. Desse modo, acredita-se que esses tipos de pacotes, que caracterizam-se como *pacotes lexicais relacionados ao tópico* de dois tipos distintos, explicitados no estudo, devem ser eliminados para que se possa traçar uma correlação mais fidedigna entre a produção de pacotes lexicais e o nível de proficiência linguística, como argumentam Chen & Baker (2010). O estudo também revelou que o *corpus* dos aprendizes mais proficientes também apresentou pacotes longos, porém, somente do tipo *prompt bundle*, uma vez que esse grupo não produziu pacotes lexicais longos retirados de textos utilizados para consulta.

Em relação ao objetivo da pesquisa de se desenvolver uma metodologia automatizada para a eliminação de *pacotes lexicais relacionados ao tópico* e de eliminação e refinação de p*acotes lexicais em contexto de sobreposição*, é possível concluir que o software R permitiu que *scripts* fossem elaborados para eliminar grande parte desses itens de maneira automatizada, refinando a frequência dos itens pesquisados e eliminando alguns deles. Além disso, os padrões encontrados pelos *scripts* puderam ser corroborados pelos resultados encontrados com a utilização de outros software.

Entretanto, as análises mostraram que alguns pacotes lexicais não foram eliminados, o que nos remete a primeira pergunta de pesquisa do trabalho – A metodologia automatizada é capaz de eliminar *pacotes lexicais relacionados ao tópico* e *pacotes lexicais em contexto de sobreposição* de maneira eficaz? Sobre os *pacotes relacionados ao tópico*, aqueles que apresentaram uso criativo e não exatamente iguais a porções das instruções e *pacotes relacionados ao tópico* menores que não foram contemplados pela estratégia utilizada de se gerar uma lista de pacotes a partir da lista de pacotes lexicais longos, de 6 palavras ou mais, não foram eliminados pela metodologia desenvolvida. Além disso, em relação aos *pacotes lexicais em contexto de sobreposição*, só foi possível realizar as eliminações e refinações a partir de uma lista de pacotes escolhida previamente, uma vez que a metodologia desenvolvida parte da unidade mínima de um pacote qualquer.

Apesar dessas limitações, a metodologia automatizada permitiu responder à segunda pergunta de pesquisa – Como a eliminação de *pacotes lexicais relacionados ao tópico* e de *pacotes lexicais em contexto de sobreposiçã*o afeta o resultado em relação às ocorrências de pacotes lexicais nos *corpora* investigados? Houve uma diminuição de 1,31%, em média, de pacotes lexicais do tipo *prompt bundle*, de 22,50% do restante de *pacotes lexicais relacionados ao tópico*, de 18,21% dos *types* examinados da amostra da lista AFL (SIMPSON-VLACH; ELLIS, 2010) por estarem inseridos em contextos de sobreposição, e de 37,92% dos *tokens* examinados da lista AFL, também por estarem inseridos em contextos de sobreposição, nos dois *corpora*. Foi possível verificar que as eliminações ocorreram em maior proporção em relação ao restante dos *pacotes lexicais relacionados ao tópico*, aos *pacotes lexicais em contexto de sobreposição completa* e aos *pacotes lexicais em contexto de subsunção completa* no *corpus* dos aprendizes menos proficientes do que no *corpus* dos aprendizes mais proficientes. Dessa

maneira, foi possível concluir que a maior produção de pacotes lexicais por parte dos aprendizes menos proficientes parece estar relacionada ao fato de que eles produzem mais pacotes lexicais dos tipos que deveriam ser eliminados. Já em relação aos *prompt bundles*, ao contrário do esperado, a metodologia automatizada revelou que os aprendizes mais proficientes produziram mais pacotes desse tipo. Foi possível, porém, elaborar uma hipótese para explicar esse fato, e se confirmada, os resultados podem ser interpretados de outra maneira.

Caso a correlação entre maior nível de proficiência e maior uso de pacotes lexicais seja corroborada em estudos futuros com o uso de *corpora* diferentes, de níveis de proficiência variados, entre outras decisões metodológicas, acredita-se que esse resultado atestaria a natureza desses itens como itens fraseológicos, produzidos de maneira natural e consequentemente com maior frequência por parte dos nativos de uma língua, ou por falantes mais proficientes. Esse resultado também contribuiria para justificar o ensino desses itens na sala de aula e contestaria uma das justificativas mencionadas por estudos que correlacionam menor nível de proficiência e maior uso de pacotes lexicais. Como discutido anteriormente, autores argumentam que aprendizes menos proficientes dependem mais da linguagem formulaica em estágios iniciais e passam a produzir sequências próprias a medida que seu nível de proficiência aumenta e correlacionam esse fato a estudos de aquisição de segunda língua (STAPLES *et al.*, 2013), que por sua vez, apontam que sequências desenvolvimentais iniciam-se por processos de memorização e mapeamento de um-para-um de forma e função, e lentamente direcionam-se para uma produção mais aproximada à produção de nativos (ELLIS, 2006 *apud* STAPLES *et al.*, 2013). Uma vez que a correlação encontrada entre menor nível de proficiência engloba a produção de pacotes lexicais em geral, e não somente a de *prompt bundles*, por exemplo, acredita-se que a argumentação apresentada pode ser considerada paradoxal, já que pacotes lexicais são unidades que refletem a produção de nativos, justamente por serem unidades geradas estatisticamente, com base em frequência.

É importante ressaltar também que o presente trabalho tratou *pacotes relacionados ao tópico* e *prompt bundles* separadamente, diferentemente de estudos anteriores. A metodologia adotada aponta questões relevantes que podem influenciar os resultados concernentes à produção de pacotes lexicais e, consequentemente, a interpretação dos dados.

Quanto ao terceiro objetivo da pesquisa de verificar se a eliminação de *pacotes lexicais relacionados ao tópico* e a eliminação e refinação de *pacotes lexicais em contexto de sobreposição* pode ser um fator que influencie a percepção da correlação entre maior nível de proficiência e uso de mais pacotes lexicais na escrita acadêmica de inglês, não foi possível alcançá-lo por meio da metodologia automatizada desenvolvida neste trabalho, pois ela não pôde eliminar todos os pacotes lexicais dos tipos supracitados. Portanto, a terceira pergunta de pesquisa – Após as eliminações desses pacotes, é possível correlacionar maior nível de proficiência a maior uso de pacotes lexicais? – não foi completamente respondida. Pode-se concluir, porém, que se há uma indicação de que se todos esses pacotes pudessem ter sido eliminados, os aprendizes de inglês do Ch-ICLE teriam produzido menos pacotes lexicais do que o Dt-ICLE, uma vez que foi demonstrado que o *corpus* dos aprendizes menos proficientes produziram sempre mais pacotes lexicais dos tipos pacotes lexicais *relacionados ao tópico*, *pacotes lexicais em contexto de sobreposição completa* e *pacotes lexicais em contexto de subsunção completa*, com exceção dos *prompt bundles*. Neste momento questiona-se se esses resultados configuram uma coincidência, ou se aprendizes menos proficientes realmente produziriam mais pacotes lexicais desses tipos por alguma razão. Acredita-se que há uma justificativa clara para a maior produção de *pacotes lexicais relacionados ao tópico* por parte de aprendizes menos proficientes. Aprendizes menos proficientes dependem mais do insumo linguístico disponível em textos utilizados para pesquisa, e como demonstrado por outros estudos, do insumo linguístico disponível nas instruções dos textos (STAPLES *et al.*, 2013). Como demonstrado no presente trabalho, a justificativa para a maior produção de *pacotes lexicais em contexto de sobreposição completa* e de *pacotes lexicais em contexto de subsunção completa* por parte dos aprendizes menos proficientes parece estar relacionada a duas questões principais. Uma vez que houve uma diminuição maior desses tipos de pacotes no Ch-ICLE, em média, foi possível concluir que aprendizes menos proficientes produziram mais versões diferentes, ou repetições, dos pacotes da AFL, pois, em primeiro lugar, muitos deles eram, na verdade *pacotes lexicais relacionados ao tópico* do tipo 1. As refinações evidenciaram os pacotes maiores que por sua vez estavam sobrepostos aos pacotes maiores e esses últimos eram contabilizados equivocadamente como pacotes pertencentes a categorias da AFL. Já que *pacotes lexicais relacionados ao tópico* do tipo 1 evidenciam a reprodução de excertos de textos especializados utilizados para consulta por parte dos aprendizes de língua materna chinesa, é

natural que partes deles pudessem ser categorizadas em uma taxonomia que identifica a produção da escrita acadêmica em inglês. Em segundo lugar, muitos dos pacotes produzidos eram parecidos, mas não idênticos aos da AFL, o que pode configurá-los como pacotes lexicais de aprendizes. Como antes das refinações e eliminações esses pacotes encontravam-se quebrados, muitos deles encaixavam-se nas categorias da taxonomia e após a aplicação da metodologia, muitos deixaram de fazer parte delas. Esses pacotes, portanto, inflam os resultados e parecem refletir o menor nível de expertise por parte de quem escreve o texto. A proporção maior de eliminação de *pacotes lexicais em contexto de sobreposição completa* e de *pacotes lexicais em contexto de subsunção completa* no *corpus* dos aprendizes menos proficientes – média de 26,91% no Ch-ICLE e média de 9,52% no Dt-ICLE para *types* e média de 46,03% no Ch-ICLE e média de 29,81% no Dt-ICLE para *tokens* – pôde evidenciar as conclusões citadas acima, além dos exemplos dos tipos de pacotes eliminados.

Como contribuição para a área destaca-se a possibilidade de disponibilizar-se os *scripts* desenvolvidos para utilização por parte da comunidade científica que investiga pacotes lexicais e, dessa maneira, otimizar as análises desses itens facilitando sua classificação pragmático-funcional bem como refinando sua frequência. Os *scripts* serão disponibilizados mediante pedidos. Essa decisão foi tomada, pois ainda é necessário otimizá-los tanto em relação a seu tempo de processamento quanto em relação à economia de linhas para objetivos didáticos, bem como desenvolver um manual de como utilizar os *scripts*. Ressalta-se ainda que o grupo de pesquisa do qual faço parte e outros pesquisadores podem utilizar esses *scripts* para analisar e comparar outros *subcorpora* do ICLEv2, por exemplo, ou utilizar a lista AFL (SIMPSON-VLACH; ELLIS, 2010) integralmente para as eliminações. Esses passos podem contribuir para a discussão da relação entre nível de proficiência e uso de pacotes lexicais.

Além das limitações mencionadas a respeito da metodologia desenvolvida, a presente pesquisa possui algumas outras que merecem ser citadas para que pesquisas futuras possam tentar superá-las. A primeira delas refere-se ao fato de que apenas dois *corpora* foram pesquisados, representantes dos níveis menos e mais proficientes. Para corroborar os resultados encontrados neste trabalho, é necessário que mais *corpora* representantes de cada nível sejam investigados. Além disso, uma variedade de *corpora* representantes do discurso acadêmico também seria interessante, uma vez que nesta ocasião, investigou-se apenas *corpora* de textos

argumentativos de estudantes de Letras, aprendizes de inglês de língua materna chinesa e holandesa. Ressalta-se ainda que o presente trabalho, que investiga especificamente o gênero redação argumentativa, faz comparações com resultados de estudos que utilizam gêneros textuais variados da escrita acadêmica. Além disso, utilizam-se neste trabalho listas de pacotes lexicais gerados de também outros tipos de gêneros textuais da escrita acadêmica, por dois motivos principais. O primeiro deles refere-se ao fato de não haver listas dos pacotes lexicais mais utilizados em redações de alunos universitários na literatura atual. O segundo deles diz respeito ao fato de que listas de pacotes lexicais baseadas em textos acadêmicos trazem a linguagem que espera-se que alunos universitários venham a utilizar.

Por fim, conclui-se que os resultados encontrados neste trabalho apontam para a importância de considerar *pacotes lexicais relacionados ao tópico* e *pacotes lexicais em contexto de sobreposição* nas análises realizadas, principalmente se uma correlação entre nível de proficiência e uso de pacotes lexicais for desejável. Ressalta-se a importância de elaborar um processo para eliminar esses pacotes automaticamente, e a metodologia apresentada nesta pesquisa pode ser considerada um primeiro passo para que esse objetivo seja alcançado.

A partir dos resultados apresentados, espera-se determinar um índice que indique quantas vezes a contagem de um pacote está aumentada. Dessa forma, será possível calcular, a partir de sua contagem bruta, a frequência verdadeira desse pacote, além de comparar esse índice em diferentes *corpora*. Como o *script* demanda otimizações e um equipamento que permita um processamento mais intenso, existe a possibilidade de disponibilizar um site para que os usuários possam submeter os seus *corpora* e realizarem as análises apresentadas neste trabalho. Esperamos ainda abordar os temas que ficaram sem respostas, para ampliar a aplicação da metodologia proposta. Pretende-se ainda utilizar cálculos estatísticos para descrever características peculiares aos *corpora* investigados, como por exemplo tamanho das redações e categorias de listas como a AFL mais e menos utilizadas por cada um dos grupos.

REFERÊNCIAS

ÄDEL, A.; RÖMER, U. Research on advanced student writing across disciplines and levels: Introducing the Michigan Corpus of Upper-level Student Papers. *International Journal of Corpus Linguistics*, v. 17, n. 1, p. 3–34, 2012.

ALTENBERG, B. On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. In: COWIE, A.P. (Ed.). *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, 1998. cap.5, p. 101-122.

ANTHONY, L. AntConc. Tokyo: Waseda University. Disponível em: <http://www.antlab.sci.waseda.ac.jp/>. , 2011

BARLOW, M. Collocate. 1.0: Locating collocations and terminology. Houston: Athelstan, 2004

BIBER, D.; *et al., Corpus Linguistics: Investigating Language Structure and Use*. New York: Cambridge University Press, 1998.

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

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. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, v. 14, n. 3, p. 275–311, 2009.

BIBER, D.; GRAY, B. *Discourse Characteristics of Writing and Speaking Task Types on the TOEFL iBT® Test : A Lexico-Grammatical Analysis.* TOEFL iBT® Research Report, 2013. 127 p. Relatório.

BOHÓRQUEZ, C. *et al*., O Impacto da Eliminação de Pacotes Lexicais Relacionados ao Tópico e em Contexto de Sobreposição. In: XI Encontro de Linguística de Corpus, 2012, São Carlos: Anais, 2012. Disponível em <http://nilc.icmc.usp.br/elc-ebralc2012/anais/andamento/104021.pdf>. Acesso em: 10 nov. 2013.

CHEN, Y.-H. *Lexical Bundles across Learner Writing Development*. 2009. 339 f. - Universidade de Lancaster, Lancaster, 2009.

CHEN, Y.-H.; BAKER, P. Lexical Bundles in L1 And L2 Academic Writing. *Language Learning & Technology*, v. 14, n. 2, p. 30–49, 2010.

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, p. 33–43, 2013.

COWIE, A. P. Phraseology. In: ASHER, R. E. (Ed.). *The Encyclopedia of Language and Linguistics.* Oxford: Oxford University Press, 1994.

COWIE, A. P. *Phraseology Theory, Analysis, and Applications.* Oxford: Oxford University Press, 1998.

DUTRA, D.; BERBER-SARDINHA, T. Referential expressions in English learner argumentative writing. In: GRANEGR, S.; GILQUIN, G.; MEUNIER, F. (Ed.). *Twenty Years of Learner Corpus Research: Looking back, Moving ahead.* Louvain-la-Neuve: Presses Universitaires de Louvain, 2013. p. 117–127.

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.

FIRTH, J. R. Linguistic analysis as a study of meaning. In: PALMER, F.R. (Ed.). *Selected Papers of J. R. Firth 1952-59*. London/Harlow: Longmans, Green and Company, 1968a. cap.1, p. 12–26.

FIRTH, J. R. A synopsis of linguistic theory, 1930-55. In: PALMER, F.R. (Ed.). *Selected Papers of J. R. Firth 1952-59*. London/Harlow: Longmans, Green and Company, 1968b. cap. 11, p. 168–205.

GRANGER, S. *et al*., (Ed.). *International Corpus of Learner English v.2.* Louvain-la-Neuve: Presses Universitaires de Louvain, 2009.

GRANGER, S.; MEUNIER, F. (Ed.). *Phraseology:* An interdisciplinary perspective. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2008.

GREAVES, C.; WARREN, M. What can a corpus tell us about multi-word units? In: O'KEEFFE, A.; MCCARTHY, M. (Ed.). *The Routledge Handbook of Corpus Linguistics.* London and New York: Routledge Taylor & Francis Group, 2010. cap.16, p. 212–226.

GRIES, S. Phraseology and linguistic theory: A brief survey. In: GRANGER, S.; MEUNIER, F. (Ed.). *Phraseology:* An interdisciplinary perspective. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2008. cap. 1, p. 3–26.

GRIES, S. T. *Quantitative Corpus Linguistics with R*: a practical introduction. New York: Routledge, 2009.

HORNIK, K.; FEINERER, I.; MEYER, D. Text Mining Infrastructure in R. *Journal of Statistical Software*, v. 25, n. 5, p. 1–54, 2008.

HYLAND, K. Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, v. 18, n. 1, p. 41–62, 2008.

JARGAS, A. *Expressões Regulares:* uma abordagem divertida. 4. ed. São Paulo: Novatec Editora Ltda., 2012.

MCENERY, T.; HARDIE, A. *Corpus Linguistics Method, Theory and Practice*. New York: Cambridge University Press, 2012.

MCENERY, T.; WILSON, A. Early corpus linguistics and the Chomskyan revolution. In: MCENERY, T.; WILSON, A. (Ed.). *Corpus Linguistics*: An Introduction. 2. ed. Manchester: Edinburgh University Press, 2001.

O'KEEFFE, A.; MCCARTHY, M. (Ed.). *The Routledge Handbook of Corpus Linguistics*. London and New York: Routledge Taylor & Francis Group, 2010.

PALMER, F. R. (Ed.). *Selected Papers of J. R. Firth 1952-59*. London and Harlow: Longmans, Green and Co Ltda., 1968.

PAQUOT, M.; GRANGER, S. Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, v. 32, p. 1–19, 2012.

R CORE TEAM. . Vienna: R Foundation for Statistical Computing, Vienna. Disponível em: <http://www.r-project.org/>. , 2013

REPPEN, R. *Using Corpora in The Language Classroom*. New York: Cambridge University Press, 2010.

RÖMER, U. Corpus Research Applications in Second Language Teaching. *Annual Review of Applied Linguistics*, v. 31, p. 205–225, 2 set. 2011.

RÖMER, U.; WULFF, S. Applying corpus methods to written academic texts : Explorations of MICUSP. *Journal of Writing Research*, v. 2, n. 2, p. 99–127, 2010.

SALAZAR, D. *Lexical bundles in scientific English:* A corpus-based study of native and nonnative writing. 2008. 304 f. - Universidade de Barcelona, Barcelona, 2008.

SCOTT, M. WordSmith Tools. Oxford: Oxford University Press, 1998

SHEPHERD, T. M. G. O Estatuto da Linguística de Corpus : Metodologia ou Área da Linguística? *Matraga*, v. 16, n. 24, p. 150–172, 2009.

SIMPSON-VLACH, R.; ELLIS, N. C. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, v. 31, n. 4, p. 487–512, 2010.

SINCLAIR, J. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

SINCLAIR, J. Prefácio. In: GRANGER, S.; MEUNIER, F. (Ed.). *Phraseology:* An interdisciplinary perspective. Amsterdam/Philadelphia: Jonh Benjamins Publishing Company, 2008.

SIYANOVA, A.; SCHMITT, N. L2 Learner Production and Processing of Collocation: A Multi-study Perspective. *Canadian Modern Language Review/ La Revue canadienne des langues vivantes,* v. 64, n. 3, p. 429–458, 2008.

STAPLES, S. *et al*,. Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, v. 12, n. 3, p. 214–225, 2013.

STUBBS, M. *Text and Corpus Analysis:* Computer Assisted Studies of Language and Culture. Oxford: Blackwell Publishers Ltda., 1996.