PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE LETRAS

LÍVIA PRETTO MOTTIN

ANÁLISE DA PRODUÇÃO METAFÓRICA NO *BRAZILIAN ENGLISH LEARNER CORPUS*

PORTO ALEGRE
2012

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS
ÁREA: LINGUÍSTICA
LINHA DE PESQUISA: TEORIAS E USOS DA LINGUAGEM

ANÁLISE DA PRODUÇÃO METAFÓRICA NO *BRAZILIAN ENGLISH LEARNER CORPUS*

LÍVIA PRETTO MOTTIN

ORIENTADOR: PROF. DR. AUGUSTO BUCHWEITZ

Dissertação de mestrado em Letras, apresentada como requisito parcial para a obtenção do título de mestre pelo Programa de Pós-Graduação em Letras da Pontifícia Universidade Católica do Rio Grande do Sul.

PORTO ALEGRE
2012

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS
ÁREA: LINGUÍSTICA
LINHA DE PESQUISA: TEORIAS E USOS DA LINGUAGEM

ANÁLISE DA PRODUÇÃO METAFÓRICA NO *BRAZILIAN ENGLISH LEARNER CORPUS*

LÍVIA PRETTO MOTTIN

ORIENTADOR: PROF. DR. AUGUSTO BUCHWEITZ

BANCA EXAMINADORA

Profª. Drª. Cristina Becker Lopes Perna
Faculdade de Letras – PUCRS

Profª. Drª. Simone Sarmento
Instituto de Letras – UFRGS

Ao Bruno.

AGRADECIMENTOS

Esses dois anos de mestrado foram um período de mudanças, assimilações, perdas e ganhos. Mudança de cidade, mudança de vida. Assimilação da falta. Quanto às perdas, toda mudança traz perdas, sim, eu sei... perdi até a mim mesma em alguns momentos. Mas em toda perda, um ganho sempre pega carona. É aquela velha história, parece *cliché*, mas essa é a vida. A minha, a tua e a de todo mundo. Na matemática das minhas perdas e ganhos desses dois anos, não tenho dúvidas do quanto o saldo final é positivo. Positivo por me trazer mais aprendizado, mas principalmente, mais pessoas. Ou talvez por me fazer enxergar com mais clareza ainda a importância dessas pessoas na minha vida. Algumas são aquelas de sempre, outras de não tão sempre assim, algumas mais recentes, mas todas, hoje, fazem parte da minha vida!

Obrigada aos meus pais pelo carinho incondicional, pela presença, pelo apoio, pela segurança, pela paciência, pelo incentivo, pelo cuidado, mas principalmente pelo amor de sempre.

Obrigado ao Guto por ser quem é.

Um agradecimento super especial à pessoa que entrou na minha vida há pouco tempo, mas que ocupa um espaço privilegiado no meu coração, a amiga Simone.

Também entram na lista dos meus agradecimentos pessoais pessoas não menos importantes: minha avó, meus tios, primos e amigos.

No meio acadêmico, agradeço de coração à Simone Sarmento e à Maitê Gil por terem me socorrido nos momentos de desespero. E por além dos conhecimentos teóricos, terem sempre um ombro amigo disponível e estarem dispostas a me escutar. Vocês são especiais

Obrigada também à Aline Pacheco por ter gentilmente permitido que eu trabalhasse com seu precioso BELC.

Obrigada ao Prof. Dr. Tony Berber Sardinha pelos sábios conselhos e pela humildade em compartilhar seu infinito saber.

À Cristina e ao Augusto por terem dividido a responsabilidade de me orientar no desenvolvimento deste trabalho.

Agradeço também aos professores e colegas do PPGL da PUCRS e ao CNPq pela bolsa auxílio.

Enfim, agradeço a todos que estiveram ao meu lado durante o período do mestrado e que, de alguma forma, me ajudaram a concretizar e concluir esta pesquisa.

RESUMO

Este trabalho reúne referenciais teóricos da Linguística de Corpus e de correntes teóricas da metáfora e tem o objetivo de verificar a variação na produção de metáforas por aprendizes brasileiros de inglês, através de uma abordagem baseada em corpus. O corpus utilizado na investigação foi o *Brazilian English Leaner Corpus* (BELC) (PACHECO, 2010), o qual é composto por quatro níveis de proficiência: (i) Beginner, (ii) Pre-Intermediate, (iii) Intermediate e (iv) Advanced; e três tarefas: (i) tarefa 1 – texto descritivo com informações pessoais em 1ª pessoa, (ii) tarefa 2 – texto descritivo com informações pessoais em 3ª pessoa e (iii) tarefa 3 – texto narrativo sobre uma viagem. O corpus foi anotado manualmente, com base nos procedimentos de Cameron (2003) e do Grupo Pragglejaz (2007). A frequência de metáforas foi extraída com a ferramenta *Concord* do *WordSmith Tools* (SCOTT, 2012). A pesquisa foi organizada nas seguintes fases: (i) anotação manual de metáforas no corpus; (ii) extração da frequência de metáforas no BELC e em seus subcorpora; (iii) comparação das frequências de uso de metáforas linguísticas nos quatro subcorpora de níveis de proficiência; (iv) comparação das frequências de uso de metáforas linguísticas nos três subcorpora de tipos textuais; (v) comparação das frequências de uso de metáforas linguísticas nos doze subcorpora individuais. Os níveis de significância das variações de frequência foram calculados com o teste estatístico *Log Likelihood.* Os resultados mostram que a produção de metáforas aumenta em cada nível de proficiência e varia de acordo com tipos textuais diferentes: textos com informações pessoais em 3ª pessoa tendem a apresentar frequência mais alta de itens metafóricos em comparação com narrativas pessoais em 1ª pessoa, o que corrobora resultados de estudos anteriores (BERBER SARDINHA, 2012).

Palavras-chave: Linguística de Corpus. Corpus de aprendiz. Produção metafórica. Língua estrangeira.

ABSTRACT

This study brings together theoretical assumptions from Corpus Linguistics and metaphor theories and aims at verifying variation in metaphor production by Brazilian English learners following a corpus-based approach. The corpus used for the investigation was the *Brazilian English Leaner Corpus* (BELC) (PACHECO, 2010), which is composed of four proficiency levels: (i) Beginner, (ii) Pre-Intermediate, (iii) Intermediate, and (iv) Advanced; and three types of task: (i) task 1 (descriptive text with personal information in 1st person), (ii) task 2 (descriptive text with personal information in 3rd person), and (iii) task 3 (narrative text about a trip). The corpus was manually annotated based on the procedures established by Cameron (2003) and by the Pragglejaz Group (2007). The metaphor frequency was later calculated using the software WordSmith tools (SCOTT, 2012), more specifically, the Concord tool. The research was organized in the following stages: (i) manual annotation of metaphor occurrences; (ii) extraction of metaphor frequency in the whole corpus and in its subcorpora; (iii) comparison of frequencies of use of linguistic metaphors in the three proficiency levels subcorpora; (iv) comparison of frequencies of use of linguistic metaphors in the four textual types subcorpora; (v) comparison of frequencies of use of linguistic metaphors in the twelve individual subcorpora. The significance levels of the frequency variations were calculated with the statistical test Log Likelihood. The results show that metaphor production increases at each proficiency level and varies according to the different text types: texts containing personal information in 3rd person tend to present higher frequency of metaphorical items when compared to personal narratives in 1st person, what corroborates findings from previous studies (BERBER SARDINHA, 2012).

Keywords: Corpus Linguistics. Learner corpus. Metaphor production. Foreign language.

LISTA DE ABREVIATURAS

A: *Advanced*

B: *Beginner*

BELC: *Brazilian English Learner Corpus*

BNC: *British National Corpus*

BoE: *Bank of English*

CEPRIL: Centro de Pesquisa, Recursos e Informação em Linguagem

COCA: *Corpus of Contemporary American English*

COHA: *Corpus of Historical American English*

CoMAprend: Corpus Multilíngue de Aprendizes

COMET: Corpus Multilíngue para Ensino e Tradução

CorTrad: Corpus de Tradução

EBRALC: Escola Brasileira de Linguística Computacional

ELC: Encontro de Linguística de Corpus

HKUST: *Hong Kong University of Science and Technology Learner Corpus*

I: *Intermediate*

ICLE: *International Corpus of Learner English*

KWIC: *Key-Word-In-Context*

LdC: Linguística de Corpus

LE: Língua Estrangeira

LILE: Corpus de Linguística e Literatura

LL: *Log Likelihood*

LLC: *Longman Learners' Corpus*

L1: Primeira Língua

L2: Segunda Língua

MCI: *Metaphor Candidate Identifier*

MIP: *Metaphor Identification Procedure*

OULC: *Oxford University Learning Center*

P: *Pre-Intermediate*

PB: Português Brasileiro

TTR: *type/token ratio*

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

SUMÁRIO

1 CONSIDERAÇÕES INICIAIS

A Linguística de Corpus (LdC) é uma abordagem empírica para o estudo da língua e serve como uma fonte de dados que reflete a língua como é usada em contextos reais. A palavra corpus, originalmente utilizada para designar um conjunto de dados sobre um determinado tema, adquiriu um novo sentido na LdC. Nessa área, um corpus é uma coleção de textos autênticos (orais ou escritos) coletados de acordo com critérios específicos, representativos de uma língua, variedade linguística ou linguagem especializada e armazenados em formato eletrônico. O objetivo principal de um corpus é servir como referência do que é típico na língua, sendo assim utilizado em pesquisas linguísticas. Através do distanciamento de exemplos artificiais, o uso da LdC confere plausibilidade às pesquisas linguísticas de natureza quantitativa e qualitativa de descrição da língua.

Os corpora podem ser de diversos tipos, sendo que cada um deles cumpre seu papel na investigação de aspectos linguísticos. Um corpus diacrônico é usado para descrever o desenvolvimento e as mudanças de uma língua ao longo dos anos. Já um corpus especializado visa a representar certo tipo de linguagem e contém textos específicos de uma determinada área de conhecimento, como artigos médicos sobre cardiologia, por exemplo. Outro tipo de corpus é o corpus de aprendiz. Um corpus de aprendiz é formado por textos autênticos (ver nota de rodapé 4 sobre autenticidade) produzidos por falantes de uma LE[1] em contextos de aprendizagem.

Corpora de aprendizes proporcionam o acesso a produções autênticas de aprendizes, oferecendo uma base empírica não disponível às pesquisas sobre aquisição[2] de línguas antes do surgimento da LdC. Grandes quantidades de textos desse tipo, organizados de acordo com critérios rigorosos de compilação, oportunizam a identificação de dificuldades enfrentadas ao longo do processo de aprendizagem e proporcionam evidências para investigações de caráter descritivo, visando melhor entender a linguagem de aprendizes.

[1] Nesta pesquisa, os termos L2 (segunda língua) e LE (língua estrangeira) são usados indistintamente para fazer referência a uma língua que não a materna.
[2] Neste trabalho, os termos aquisição e aprendizagem são utilizados indistintamente.

Por acreditar no poder da combinação do uso de ferramentas computacionais com os dados de corpora em pesquisas de natureza linguística, a LdC é o pilar principal desta pesquisa quantitativa de análise de dados que trata da produção metafórica por aprendizes brasileiros de inglês como LE, falantes de PB como L1, no *Brazilian English Learner Corpus* (BELC) (PACHECO, 2010). O objetivo desta pesquisa é verificar a variação na produção de metáforas com relação ao nível de proficiência e ao tipo de tarefa. O corpus conta com produções escritas de 424 aprendizes, sendo eles classificados em quatro níveis de proficiência *(Beginner*, *Pre-Intermediate*, *Intermediate* e *Advanced).* As tarefas produzidas são de três tipos: tarefa 1 – texto descritivo com informações pessoais em 1ª pessoa; tarefa 2 – texto descritivo com informações pessoais em 3ª pessoa; e tarefa 3 – texto narrativo sobre uma viagem. Dentre os objetivos deste trabalho, destaco também o preenchimento de lacunas no que se refere aos estudos sobre produção metafórica em LE. Tanto a metáfora quanto a aquisição de línguas têm sido aspectos amplamente abordados em pesquisas linguísticas de natureza aplicada. No campo dos estudos da metáfora, o fenômeno tem sido abordado sob diversas perspectivas e em diversas áreas do conhecimento. Apesar disso, poucos são os estudos que trabalham na interface entre a metáfora e a aquisição de LEs, tanto que Cameron (1999) aponta para a pouca atenção dispensada à metáfora pela Linguística Aplicada. Há também uma carência no que se refere ao uso de corpora de aprendizes nesse tipo de pesquisa.

Considero importante ressaltar que apesar de ter a metáfora como um de seus pilares, este trabalho tem como foco principal a descrição da linguagem do aprendiz, com base na LdC. A metáfora aqui é o aspecto investigado dentro desse escopo. Sendo que sob essa mesma perspectiva poderiam ser estudados fatores diversos, como aspectos semânticos e pragmáticos na aprendizagem de uma LE.

Os objetivos deste trabalho estão fundamentados nas seguintes questões de pesquisa:

1. Aprendizes brasileiros de inglês como LE, falantes de PB como L1, como evidenciado pelo BELC, produzem metáforas?
2. Há variação na frequência da produção metafórica no corpus de estudo com relação ao nível de proficiência linguística em LE?
3. Há variação na produção de metáforas no corpus de estudo com relação ao tipo de tarefa?

As hipóteses que norteiam este trabalho são:

1. Aprendizes brasileiros de inglês como LE, falantes de PB como L1, produzem metáforas.

2. Há variação na produção metafórica com relação aos níveis de proficiência linguística, sendo que quanto mais avançado o nível, maior o número de ocorrências metafóricas.

3. Há variação na produção metafórica com relação ao tipo de tarefa, sendo que probabilidades de uso da linguagem metafórica variam de acordo com tipos textuais específicos.

Estudos que abordam a metáfora baseados em corpora apresentam muitas vantagens. Uma delas é conseguir mostrar através de porções de linguagem extraídas de contextos reais de uso que a metáfora na língua real é muito diferente da metáfora investigada com base em exemplos introspectivos. Prova disso é que a forma A é B *(Ela é uma flor)*, apesar de típica, é pouco frequente na língua (CAMERON, 2003, DEIGNAN, 2005). Entretanto, a metáfora impõe certos desafios para a LdC (BERBER SARDINHA, 2007b). O principal desafio da pesquisa em metáfora baseada em corpus é de natureza metodológica. Berber Sardinha *(Ibidem)* apresenta quatro métodos básicos para encontrar metáforas: (i) pela introspeção do linguista; (ii) pela leitura do corpus; (iii) pelo uso do concordanciador; e (iv) pelo uso de programa computacional identificador de metáforas. Os métodos (i) e (ii) são essencialmente manuais, enquanto que o (iii) e o (iv) são assistidos por computador, mas não eximem o pesquisador de uma análise manual cuidadosa. O método utilizado neste trabalho será a leitura do corpus. Apesar de um método muito antigo e popular, e leitura é subjetiva e requer que a anotação (ver item 2.5.4 sobre anotação) do pesquisador seja validada de alguma forma.

Diante dos desafios metodológicos e da inexistência de um modelo específico para a investigação de metáforas em corpora de aprendizes, procurou-se criar um procedimento criterioso e o menos subjetivo possível para a identificação de metáforas linguísticas. Para isso, o método utilizado será a leitura e anotação manual de metáforas no corpus através dos procedimentos de Cameron (2003) e do Grupo Pragglejaz (2007). No que se refere ao *Metaphor Indentifier Procedure* (MIP), criado pelo Grupo Pragglejaz (2007), o procedimento visa a identificação de metáforas no discurso naturalmente produzido, ou seja, na língua em

uso. Porém, o método foi desenvolvido para análise de língua materna. Esse ponto impõe desafios e limitações à anotação do BELC, pois o procedimento não prevê a existência de desvios da língua padrão e de transferências da L1 para a L2, por exemplo, comuns no processo de aprendizagem de uma LE.

A análise de metáforas nesta investigação é de cunho *bottom-up*, e tem o objetivo de levantar todas as ocorrências metafóricas do corpus. Ou seja, não se parte de uma lista de metáforas preestabelecida, são considerados todos os itens linguísticos do corpus. Com o intuito de não se limitar à simples definição do que é metáfora, o ponto de partida para a anotação de metáforas no corpus será a metáfora linguística. São utilizados pressupostos de Aristóteles, Lakoff e Johnson e Lynne Cameron. O recorte utilizado na descrição das correntes da metáfora parte da hipótese de que o principal ponto divergente entre as teorias metafóricas é o seu *lócus* (VEREZA, 2010): na visão tradicional, a metáfora ocorre na linguagem (ARISTÓTELES, 1997, [séc. IV a.C.]); na visão cognitivista, o *lócus* da metáfora é o pensamento (LAKOFF e JOHNSON, 1980); e na abordagem da metáfora sistemática (CAMERON, 2003), o discurso.

Após a anotação de metáforas no corpus, serão extraídas as frequências de ocorrências metafóricas no BELC, em seus subcorpora de níveis de proficiência, de tipos de tarefa e individuais[3]. Os números obtidos serão contrastados com o objetivo de analisar a variação do uso de metáforas nos subcorpora.

Após a breve introdução desta seção, o próximo capítulo (capítulo 2) procura situar o leitor no contexto da Linguística de Corpus. O capítulo oferece um painel histórico da LdC e apresenta os conceitos principais da área. Unindo teoria e prática, são abordadas ferramentas para análise linguística e examinadas suas contribuições para esta pesquisa, assim como para outros campos da linguística.

O terceiro capítulo objetiva apresentar os corpora de aprendizes e ressaltar a possibilidade de se investigar a língua do aprendiz através deles. Além disso, neste capítulo, descrevo o BELC (PACHECO, 2010), base empírica desta pesquisa.

[3] Um subcorpus individual corresponde a um texto específico produzido em um determinado nível. O nível avançado, por exemplo, foi transformado em três subcorpora: um subcorpus correspondente ao texto 1, um ao texto 2 e outro ao texto 3. Dessa forma divididos, serão observadas as frequências em 12 subcorpora. A identificação dos mesmos foi feita através de uma letra correspondente ao nível (B, P, I, A) e um número correspondente ao texto (1, 2, 3). O código B1, por exemplo, corresponde ao texto 1 produzido no nível *Beginner*.

No quarto capítulo são apresentadas correntes de estudo da metáfora. Além disso, discuto a maneira como a metáfora pode ser estudada sob a perspectiva da LdC, relacionando de forma mais direta a LdC aos estudos da metáfora.

No quinto capítulo, passo a abordar esta investigação propriamente dita. Apresento o método da pesquisa, alguns desafios enfrentados na sua escolha e descrevo as etapas seguidas no processo de anotação e análise dos dados do corpus.

Em seguida, no capítulo 6, são analisadas e discutidas as frequências de metáforas no BELC e em seus subcorpora de níveis de proficiência, tipos de tarefa e subcorpora individuais.

Este trabalho encerra-se com considerações finais a respeito dos dados obtidos, no capítulo 7. São retomados alguns pontos principais apresentados no decorrer do trabalho, assim como os objetivos e as questões de pesquisa, numa tentativa de respondê-las.

2 LINGUÍSTICA DE CORPUS

A partir dos anos 60, uma nova área da linguística, chamada Linguística de Corpus (LdC), conferiu um novo sentido à palavra *corpus.* Na LdC, um corpus é uma coleção de textos produzidos naturalmente na língua (em contraposição a textos induzidos e à língua da máquina), armazenados em formato eletrônico e com o intuito de serem alvo de investigações linguísticas. Através da utilização de coleções de textos naturais, a LdC cresceu consideravelmente nos últimos anos e vem impactando diversas áreas de pesquisa em linguística. Seu crescimento se deve não apenas ao seu caráter essencialmente empírico, mas à sua capacidade de gerar evidências inéditas sobre a língua, tais como frequência de palavras e palavras que tendem a co-ocorrer umas com as outras.

O objetivo deste capítulo é oferecer um painel histórico da LdC, apresentar conceitos principais e ferramentas para exploração e análise de material linguístico. Através da explicitação de teorias e práticas, pretende-se examinar as contribuições da área para investigações linguísticas, em especial para este trabalho.

2.1 LINGUÍSTICA DE CORPUS: COMO TUDO COMEÇOU

A LdC se ocupa da coleta criteriosa de porções de linguagem armazenadas em formato eletrônico com o propósito de servirem para investigações linguísticas. Assim, é uma forma empírica de estudo da língua. Por depender do uso de computadores, o surgimento do primeiro corpus eletrônico aconteceu em um contexto histórico pouco favorável aos seus avanços. Os entraves tecnológicos existentes eram muitos e as ferramentas computacionais limitadas. Como na linguística moderna, o termo corpus é quase sinônimo do termo corpus em formato eletrônico (MCENERY e WILSON, 2004 [1996]), pode-se imaginar as dificuldades enfrentadas na época para digitalizar os corpora e acessá-los através de computadores.

O primeiro corpus linguístico eletrônico, o *Brown University Standard Corpus of Present-day American English*, lançado nos anos 60, foi o marco do início dos trabalhos com

corpora. Pouco menos de dez anos antes do lançamento do corpus *Brown*, Chomsky havia lançado seu livro *Syntatic Structures*, no qual divulgava o gerativismo e defendia uma visão racionalista da linguagem em oposição à abordagem empírica da LdC. Para Chomsky, o que interessava era o estudo da competência (as normas internalizadas que o falante sabe sobre a língua) e, segundo ele, os dados necessários para tal análise provinham da intuição do linguista que os buscava em sua mente por meio da introspecção (MCENERY e WILSON, 2004 [1996], BERBER SARDINHA, 2000). Dados empíricos seriam úteis apenas para a investigação do desempenho (o uso que os falantes fazem da língua) dos usuários da língua. A compilação desse corpus, um tanto quanto desafiadora para a época, e a mudança de paradigmas linguísticos ocorrida na época (o racionalismo predominando em relação ao empirismo) foram fatores determinantes que vieram a tornar o *Brown* uma referência na LdC. A partir de então, o desenvolvimento e aprimoramento de computadores e ferramentas utilizadas para a análise de corpora vêm permitindo e possibilitando progressos na área.

Os progressos alcançados na área nos últimos anos se dão devido a fatores como a alta capacidade que os computadores atuais têm de armazenar dados e o desenvolvimento de ferramentas capazes de manipular corpora com acurácia. A Internet foi também um fator importante na história da LdC, à medida que através dela os textos não precisam mais ser digitalizados. Se textos disponíveis *online* atenderem às necessidades do linguista, podem ser facilmente retirados do ambiente virtual e armazenados em computadores para a compilação de corpora. Além de facilitar a coleta dos dados para a compilação de um corpus, a Internet proporciona ao linguista de corpus o acesso a uma variada gama de textos das mais diferentes fontes que abrange desde livros, jornais, revistas e periódicos de áreas específicas do conhecimento a bate-papos informais, por exemplo. Os progressos da LdC trazem consigo um aumento de interesse na área, o qual pode ser sentido nos eventos anuais ELC (Encontro de Linguística de Corpus) e EBRALC (Escola Brasileira de Linguística Computacional), que terão em 2013 suas 12ª e 7ª edição, respectivamente. Apesar de incipientes, esses eventos já contam com um número considerável de participantes todos os anos e de apresentações de trabalhos pertencentes a áreas como descrição de linguagens especializadas, tradução, aquisição de línguas, sociolinguística, entre outras.

2.2 LINGUÍSTICA DE CORPUS: DEFINIÇÃO E CARACTERÍSTICAS

A LdC ocupa-se da coleta criteriosa de textos autênticos (orais ou escritos) com a finalidade de serem utilizados e explorados em análises linguísticas. A LdC pode, portanto, ser descrita como uma abordagem empírica para o estudo da língua (TOGNINI-BONELLI, 2001). Ao invés de investigar o que é teoricamente possível na língua, tem como foco a investigação do uso e da maneira como os usuários utilizam os recursos de linguagem disponíveis, através da observação de material autêntico[4] (BIBER et al., 1998). Assim como a Linguística Sistêmico-Funcional, a LdC trabalha com a noção de língua enquanto sistema probabilístico. De acordo com Berber Sardinha (2000, p. 350), "a visão da linguagem enquanto sistema probabilístico pressupõe que embora muitos traços linguísticos sejam possíveis teoricamente, eles não ocorrem com a mesma frequência". A noção probabilística, proposta por Michael Halliday (1991)[5], pressupõe a existência de probabilidades que regulam as escolhas feitas pelos usuários da língua, o que significa que as escolhas dos usuários não são aleatórias, mas reguladas pela probabilidade de ocorrência de padrões possíveis na língua.

Há uma forte discussão na área no que diz respeito ao estatuto da LdC: se a LdC é uma metodologia ou se pode ser considerada uma disciplina independente. Alguns autores definem a LdC como uma disciplina independente, outros, como Granger (2002), consideram-na uma metodologia linguística. A autora é uma das linguistas que descreve a LdC como uma metodologia que tem o potencial de mudar perspectivas sobre a língua através da utilização de coletâneas de textos autênticos, produzidos em contextos reais de uso. Ainda segundo Granger (Ibidem), a LdC não é nem um novo ramo da linguística e nem uma nova teoria sobre a língua, mas uma metodologia poderosa no acesso à língua em uso.

---

[4] Neste contexto, material autêntico diz respeito a textos em linguagem natural, produzidos por humanos, em contraposição à linguagem da máquina. A principal característica da autenticidade na LdC está associada ao pressuposto básico de que os textos que compõem um corpus não podem ter sido produzidos para fins de pesquisa. O BELC, base empírica desta pesquisa, por exemplo, foi produzido para fins de investigação da aquisição de morfemas em inglês como LE. Entretanto, Pacheco (2010), durante a compilação do corpus, não induziu a produção de porções de linguagem que revelassem itens os quais estava investigando. Da mesma forma, ao compilar o corpus, a autora jamais imaginou que os textos coletados seriam alvo de uma pesquisa sobre produção metafórica. Nesse sentido, a autenticidade do material do BELC é garantida.

[5] HALLIDAY, Michael. A. K. Corpus studies and probabilistic grammar. In: AIJMER, Karin; ALTENBERG, Bengt (Orgs.). **English corpus linguistics:** Studies in honour of Jan Svartvik. London: Longman, 1991.

McEnery et al. (2007[2006]) também consideram a LdC uma metodologia. Diferentemente da sintaxe, por exemplo, a LdC não explica fatos sobre a língua, mas pode ser utilizada como forma de explorar áreas como sintaxe, semântica, pragmática; portanto, não é uma disciplina. Sarmento (2008) enfatiza que "a LdC é uma metodologia que pode ser aplicada a uma grande variedade de estudos linguísticos, ou ainda ao ensino de línguas, ou seja, é uma das várias maneiras de fazer linguística" (p. 24). Assim, o termo "de corpus" pode ser atrelado a diversas áreas da Linguística gerando expressões como "Pragmática de Corpus", por exemplo, em oposição à Pragmática não baseada em corpora (MCENERY e WILSON, 2004 [1996]).

O ponto de vista adotado neste trabalho vai ao encontro do defendido por Shepherd (2009) e Oliveira (2009) de que a LdC não é uma disciplina e nem uma metodologia de análise, mas uma abordagem, uma perspectiva para se chegar à língua empiricamente. Tal abordagem empírica é capaz de revelar novas concepções teóricas e descrições sobre a linguagem e possibilita que se reescreva "descrições existentes para a linguagem de forma mais clara" (SHEPHERD, 2009, p. 167). A LdC será aqui considerada, então, uma abordagem para o estudo da língua que se ocupa da coleta e investigação de corpora.

Um corpus, como mencionado anteriormente, é um conjunto de textos autênticos coletados de acordo com critérios específicos e armazenados em formato eletrônico para servirem de objeto a investigações linguísticas. As características básicas e importantes de um corpus são: (i) representatividade; (ii) amostragem; (iii) formato eletrônico; e (iv) autenticidade. A (i) representatividade é a particularidade que distingue um corpus de uma coleção de textos aleatórios (MCENERY et al., 2007 [2006]) e está associada ao seu tamanho (REPPEN, 2010). Na compilação de corpora para a produção de dicionários, por exemplo, o corpus precisa conter milhões de palavras a fim de incluir as mais diferentes palavras existentes na língua, assim como os diferentes sentidos de palavras polissêmicas (BIBER, 1990, REPPEN, 2010). Ou seja, é necessário que os resultados encontrados nas pesquisas baseadas em um determinado corpus possam ser generalizados para a variedade linguística como um todo (LEECH, 1991[6] apud MCENERY et al., 2007 [2006]).

[6] LEECH, Geoffrey. The state of art in corpus linguistics. In: AIJMER, Karin.; ALTENBERG, Bengt. (Ed.). **English Corpus Linguistics**. London: Longman, 1991. p. 8-29.

Entretanto, a compilação de um corpus representativo não é tarefa simples. Biber (1993) discute questões importantes na compilação de corpora representativos e salienta que a representatividade é uma característica que depende, em primeiro lugar, de uma definição da população a qual se deseja representar. É só a partir desta definição que se pode estabelecer uma base de amostragem adequada e determinar que textos serão incluídos no corpus, o número aproximado de palavras de cada texto, os gêneros aos quais esses textos pertencerão e o número de textos pertencentes a cada gênero, por exemplo. Reppen (2010) salienta que em algumas situações, a língua sendo estudada permite que o investigador compile um corpus que a represente em sua completude. Um corpus de falas dos personagens de um determinado seriado, por exemplo, tem a possibilidade de incluir todas as falas, atingindo assim representação completa.

A (ii) amostragem é a propriedade que os corpora têm de, através de uma amostra, representar uma variedade linguística. A representação oferecida pelo corpus deve mostrar as mesmas peculiaridades e suas devidas proporções encontradas na língua como um todo, em situações reais de uso. A palavra “manga”, por exemplo, apresenta dois significados na língua portuguesa: manga da camisa e a fruta manga. Um corpus geral de língua portuguesa deve conter uma amostra da língua que dentre suas ocorrências, inclua os dois sentidos de “manga”. O (iii) formato eletrônico é outra característica importante de um corpus, tanto que atualmente o termo corpus é quase sinônimo da expressão corpus digital (MCENERY e WILSON, 2004 [1996]). A formatação eletrônica dos corpora permite que os dados sejam lidos e processados por computadores rapidamente, facilitando sua manipulação por parte do pesquisador e gerando, assim, resultados consistentes e confiáveis em razão da precisa habilidade que a máquina tem de processar dados de corpora. A (iv) autenticidade dos textos sugere que textos coletados para a compilação de um corpus devem ser em linguagem natural, não de máquina, e não produzidos com o intuito de serem utilizados em investigações linguísticas (BERBER SARDINHA, 2004) (ver nota de rodapé 4 sobre autenticidade).

As quatro características acima descritas são importantes na compilação de um corpus e devem ser levadas em consideração nas pesquisas que envolvem quaisquer tipos de corpora. Estas características asseguram a qualidade do material coletado e, consequentemente, dos resultados das pesquisas baseadas nesses materiais. Se consideradas tais características, a combinação do uso de ferramentas computacionais com os dados de corpora tem a possibilidade de gerar resultados quantitativos e qualitativos confiáveis que podem revelar

fenômenos desconhecidos sobre a língua. Resultados quantitativos são estatísticos e mostram, por exemplo, a frequência com que a palavra de busca aparece em um determinado contexto. Resultados qualitativos, por outro lado, vão além dos números e exibem a maneira como palavras ou conjuntos de palavras são usados em contexto, permitindo a observação das ocorrências do termo de busca, seus contextos e formas de uso.

2.3 TIPOS DE CORPORA

Os corpora podem ser classificados de acordo com seus tamanhos, finalidades e forma como são compilados. A classificação aqui adotada foi proposta por Sarmento (2009) com base em Sinclair[7] (1995) e Hunston[8] (2002).

- Corpus Geral: Um corpus que contém muitos tipos de textos, os quais podem ser representativos da linguagem oral, escrita ou ambas. Por ser representativo de língua geral e por em diversas situações ser usado como contraste em relação aos corpora mais específicos, deve ser significativamente maior que um específico. O *Corpus of Contemporary American English*[9] (COCA) é um exemplo de corpus de língua geral representativo do inglês americano. Outro exemplo de corpus geral é o *British National Corpus*[10] (BNC), um corpus de cerca de 100 milhões de palavras, considerado uma amostra representativa do inglês britânico.
- Corpus Monitor: Tem o intuito de verificar mudanças em uma língua. Por esse motivo, novos textos são inseridos no corpus anualmente, mensalmente ou até diariamente. Segundo Berber Sardinha (2000, p. 340), "a composição é reciclada para refletir o estado atual de uma língua". Diversas obras sobre LdC indicam o *Bank of English* (BoE) como exemplo de corpus monitor. Além disso, autores sugerem seu uso na investigação de variações ocorridas no inglês, pois durante o processo de compilação, textos eram adicionados ao corpus quase que diariamente, refletindo assim as mudanças ocorridas na língua ao longo dos anos.

[7] SINCLAIR, John. **Paper presented at IX Encontro da Associação Portuguesa de Linguística.** Lisboa, 1995.
[8] HUNSTON, Susan. **Corpora in Applied Linguistics.** London: Cambridge University Press, 2002.
[9] http://corpus.byu.edu/coca/
[10] http://www.natcorp.ox.ac.uk/

- Corpus Comparável: São dois ou mais corpora representativos de duas línguas diferentes ou de diferentes variedades de uma mesma língua, os quais são usados para identificar diferenças e equivalências em cada língua. Portanto, precisam seguir as mesmas diretrizes de compilação. O corpus de Linguística e Literatura (Corpus LILE), compilado pela professora Simone Sarmento, na Universidade Federal do Rio Grande do Sul, é um corpus comparável de resumos de trabalhos de conclusão de cursos de graduação, dissertações de mestrado, teses de doutorado e artigos de revistas nacionais e internacionais das áreas de linguística e literatura, em inglês e português.
- Corpus Paralelo: Dois ou mais corpora paralelos contêm textos em uma determinada língua (L1) e suas respectivas traduções (L2). Um corpus é considerado bidirecional quando os textos das duas línguas estão alinhados em duas direções de tradução, português → respectivas traduções em inglês e inglês → respectivas traduções em português, por exemplo. Pesquisas em corpora paralelos permitem, por exemplo, que se identifique como uma determinada palavra em português foi traduzida para o inglês em diferentes contextos. Um exemplo de corpus paralelo é o CorTrad[11], um corpus paralelo bidirecional composto de textos originais em português e em inglês e suas respectivas traduções. O CorTrad é um dos subcorpora do projeto COMET[12] (Corpus Multilíngue para Ensino e Tradução), desenvolvido na Universidade de São Paulo (USP).
- Corpus de Aprendiz: Os corpora de aprendizes são coleções de textos autênticos (escritos ou orais) produzidos por falantes de uma LE em uma situação de aprendizagem. O *International Corpus of Learner English* (ICLE) é um corpus de aprendizes de inglês, falantes de diversas línguas maternas, dirigido por Sylviane Granger, na Universidade de Louvain, na Bélgica. Já no Brasil, pode-se citar o CoMAprend[13] (Corpus Multilingue de Aprendizes), um corpus de aprendizes brasileiros de diversas línguas. Com o intuito de analisar a produção de metáforas em LE, esta pesquisa tem como base empírica um corpus de aprendizes brasileiros de inglês como LE, falantes de PB como L1, o BELC, produzido durante a tese de doutorado de Pacheco (2010). Como parte fundamental deste trabalho, os corpora de aprendizes serão abordados detalhadamente no próximo capítulo.

[11] http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html
[12] http://www.fflch.usp.br/dlm/comet/
[13] http://www.fflch.usp.br/dlm/comet/comaprend.html

- Corpus Pedagógico: Constituído de livros didáticos ou gravações, um corpus pedagógico representa a linguagem à qual aprendizes são expostos e destina-se ao ensino de línguas e à pesquisas pedagógicas.
- Corpus Histórico ou Diacrônico: Formado por textos produzidos em uma determinada língua em diversos períodos de tempo, um corpus Histórico ou Diacrônico visa a identificar o desenvolvimento de uma língua através dos tempos. Através de uma pesquisa em um corpus diacrônico, é possível observar características como a mudança de significado de palavras e mudanças estilísticas que aconteceram em uma língua com o passar dos anos. Um exemplo de corpus histórico é o *Corpus of Historical American English*[14] (COHA), um corpus de 400 milhões de palavras que contém textos produzidos entre os anos 1810 e 2009.
- Corpus Especializado: Corpora especializados são corpora contendo textos específicos de uma determinada área de conhecimento, gênero, etc. Utilizados para representar certo tipo de texto ou linguagem, um corpus especializado pode conter desde bulas de remédio, manuais de eletrodomésticos, até sentenças judiciais. Entretanto, a compilação de um corpus especializado, seja ele composto por bulas de remédio ou manuais de eletrodomésticos, deve seguir determinadas diretrizes a fim de representar com exatidão a linguagem que se deseja investigar. Sarmento (2008), por exemplo, compilou um corpus de manuais de aviação a fim de descrever o uso dos verbos modais neste tipo de linguagem especializada.

2.4 ANÁLISE DE CORPORA

Como já mencionado, um corpus é uma coletânea de textos autênticos armazenados e acessados através de computadores. O conteúdo dos corpora só pode ser acessado através de ferramentas computacionais especializadas para tal tarefa. Alguns corpora estão disponíveis na Internet e dispõem de seus próprios recursos de pesquisa *online*, por exemplo o *Corpus of Contemporary American English*[15] (COCA), um corpus de língua geral representativo do inglês americano, que tem cerca de 425 milhões de palavras, foi compilado entre os anos 1990

[14] http://corpus.byu.edu/coha/
[15] http://corpus.byu.edu/coca/

e 2012 e é subdividido em corpora menores de diferentes gêneros: fala, ficção, revistas populares, jornais e textos acadêmicos. Nos casos em que o pesquisador opta pelo uso de um corpus não disponível online, há a necessidade de utilizar programas computacionais desenvolvidos especialmente para realizar o processamento dos dados do corpus. Um desses programas é o *Wordsmith Tools* (SCOTT, 2012). Independentemente da maneira pela qual os corpora são acessados, os recursos mais utilizados nas pesquisas linguísticas são: (i) concordanciador; (ii) lista de frequência; e (iii) lista de colocados.

O (i) concordanciador é uma ferramenta muito utilizada para processar as informações de um corpus e permite a observação, em contexto, das ocorrências do termo de busca. Inserindo uma palavra ou frase no campo de busca, o concordanciador gera resultados qualitativos, apresentando todas as ocorrências daquela palavra ou frase (palavra nódulo, *node*) no centro da tela acompanhada de algumas das palavras que se encontram imediatamente à sua esquerda ou à sua direita (o co-texto da palavra ou frase de busca). Tais informações são dispostas em uma tela gerada pelo programa utilizado na pesquisa, a qual é chamada de KWIC *(Key-Word-In-Context).* Cada uma das linhas retrata um uso diferente da palavra nódulo, empregada por um falante diferente, em tempo e contextos também distintos, conforme a figura 1. Simplificadamente, o concordanciador é uma "ferramenta básida da LdC e significa utilizar um programa de computador para encontrar todas as ocorrências de uma determinada palavra ou frase no corpus"[16] (O'KEEFFE et al., 2007, p. 8).

Figura 1: Linhas de concordância extraídas com o concordanciador do COCA

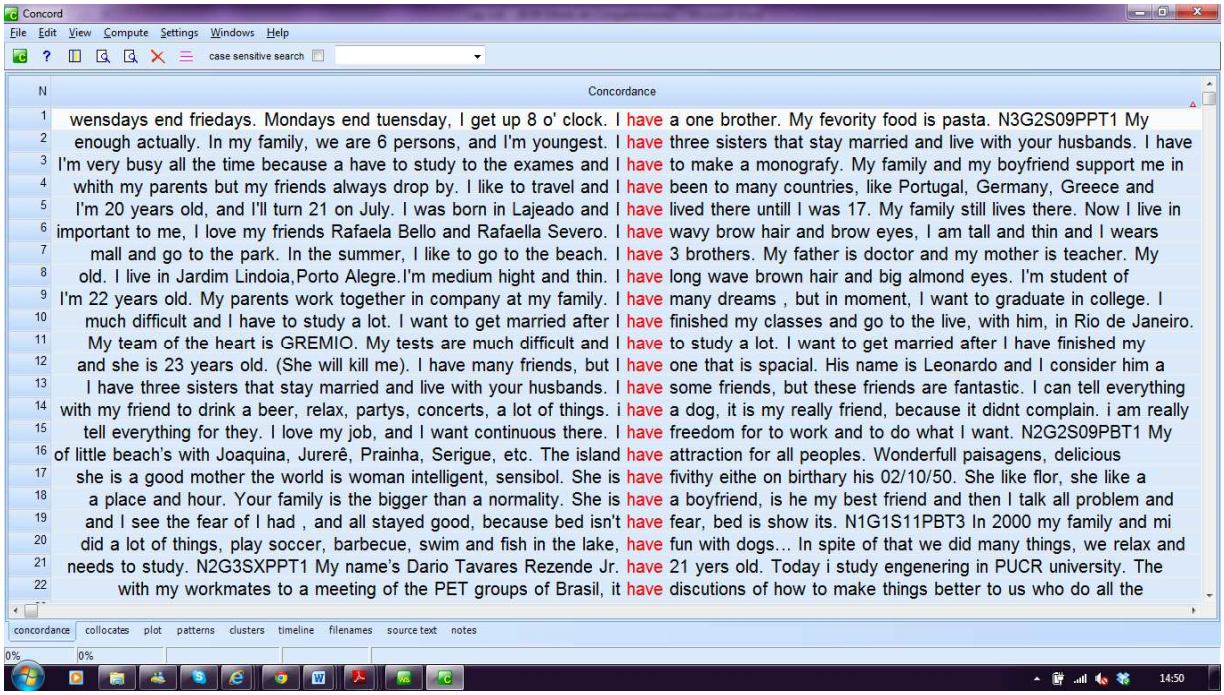
CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST --------- CREATE NEW LIST [?]

1	2011	ACAD	AmerScholar	A	B	C	review of Disaster Movie, a story about a western Massachusetts yoga center. " **Would** you be supporting yourself with this job? " he asked casually. Without thinking
2	2011	ACAD	AmerScholar	A	B	C	the building's needs admirably until about four in the afternoon, when the taps **would** suddenly run dry and the toilets refused to flush. # The staff included a
3	2011	ACAD	AmerScholar	A	B	C	for the Leader. Of course, I was American. Any attack on me **would** have received international publicity. At worst the government might revoke my tourist visa,
4	2011	ACAD	AmerScholar	A	B	C	morning, when Frederica assigned stories, and Friday afternoon, when the entire office **would** go into a mad scramble to meet deadline, little work appeared to get done
5	2011	ACAD	AmerScholar	A	B	C	devoted to arguing over what kind of cake to order as to deciding which stories **would** make the front page. When international reporters showed up to interview Frederica or Lal
6	2011	ACAD	AmerScholar	A	B	C	. # DESPITE APPEARANCES, the Leader was an influential paper in Sri Lanka and **would** play a pivotal role in the January 2010 presidential election, the first national election
7	2011	ACAD	AmerScholar	A	B	C	national election since the government's victory over the LTTE. Everyone assumed the election **would** be a cakewalk for incumbent President Mahinda Rajapaksa. Billboards througl
8	2011	ACAD	AmerScholar	A	B	C	into an empty teacup. # " If Fonseka was involved, the defense secretary **would** have arrested him to draw suspicion away from himself, " Lal said. "
9	2011	ACAD	AmerScholar	A	B	C	Frederica was interviewing Fonseka at his office when the candidate made a stunning accusation that **would** change the course of the election. According to Fonseka, in the final blo
10	2011	ACAD	AmerScholar	A	B	C	Defense Secretary Gotabaya Rajapaksa, the president's brother. # Worried that government thugs **would** once again burn down the Leader's press to stop the story's publication,
11	2011	ACAD	AmerScholar	A	B	C	, I asked Frederica about the danger the story could put her in. How **would** Defense Secretary Rajapaksa react to an accusation of war crimes? Frederica scoffed at the
12	2011	ACAD	AmerScholar	A	B	C	Sri Lanka's sclerotic justice system. A decision in favor of the defense secretary **would** almost certainly shut the Leader down for good. # I stayed at the paper
13	2011	ACAD	AmerScholar	A	B	C	. # On Friday nights after the paper was put to bed, the staff **would** sometimes gather in Frederica's office to snack on Sri Lankan staples like fried vadai
14	2011	ACAD	AmerScholar	A	B	C	Sri Lankan staples like fried vadai and coconut sambol. Cigarette in hand, Lal **would** regale everyone with the exploits of his charismatic brother, who had seemed to know
15	2011	ACAD	AmerScholar	A	B	C	not take that commitment for granted. # Lasantha's posthumous prediction that President Rajapaksa **would** thwart a full investigation into his murder proved all too accurate. Four
16	2011	ACAD	AmerScholar	A	B	C	was a struggle as dramatic -- and perhaps even as decisive -- as any that **would** play out at Shiloh or Chattanooga. It was also a struggle unlike What most
17	2011	ACAD	AmerScholar	A	B	C	. # Throughout the winter and early spring of 1861, the Union revolutionaries who **would** soon fight the battle for Missouri were preparing for the war in hidden corners of
18	2011	ACAD	AmerScholar	A	B	C	they threw caution aside and sang out. Just a few of the older men **would** begin, then more and more men joined in until dozens swelled the chorus,
19	2011	ACAD	AmerScholar	A	B	C	or rather the German dialect, everywhere, " one Landsmann enthused. Certainly you **would** hear it in places like Tony Niederwiesser's Tivoli beer garden on Third Street,
20	2011	ACAD	AmerScholar	A	B	C	or Vogel's orchestra played waltzes and sentimental tunes from the old country. You **would** hear it in the St. Louis Opera House on Market Street, where the house

[16] Tradução minha. Texto original: *"concordancing is a core tool in corpus linguistics and it simply means using corpus software to find every occurrence of a particular word or phrase"*.

No caso do *WordSmith Tools*, o pesquisador precisa realizar o *upload* de um corpus no *software* para obter as informações desejadas. A figura 2 ilustra a extração das linhas de concordância da palavra *have* no BELC[17], base de dados desta pesquisa.

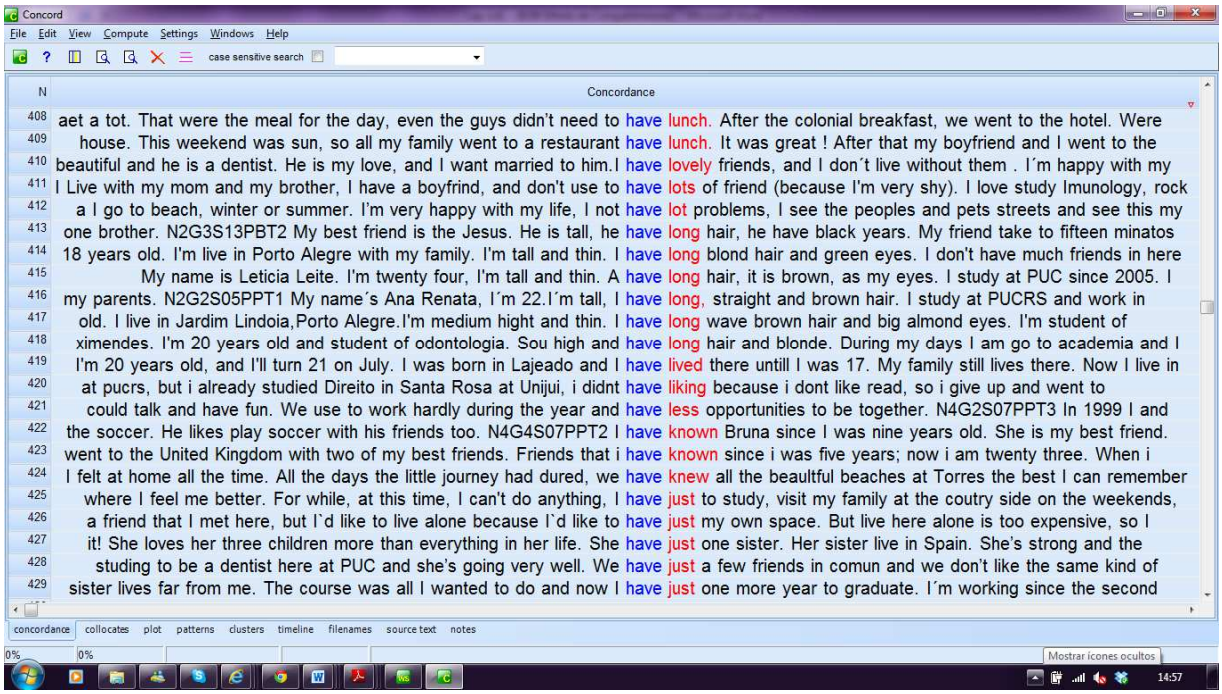
Figura 2: Linhas de concordância extraídas com o concordanciador do *WordSmith Tools*



A palavra de busca *(have)* é apresentada no centro da tela. Existe também a possibilidade de solicitar que o programa apresente os resultados na ordem em que aparecem no corpus ou em ordem alfabética das palavras do co-texto à direita ou à esquerda da palavra nódulo. Na figura 3, as linhas de concordância foram organizadas em ordem alfabética, de acordo com a primeira letra da palavra imediatamente à direita de *have*.

[17] As linhas de concordância de *have* extraídas do BELC apresentam desvios da língua padrão, comuns no processo de aprendizagem de um LE.

Figura 3: Linhas de concordância organizadas de acordo com a primeira letra da palavra imediatamente à direita da palavra nódulo



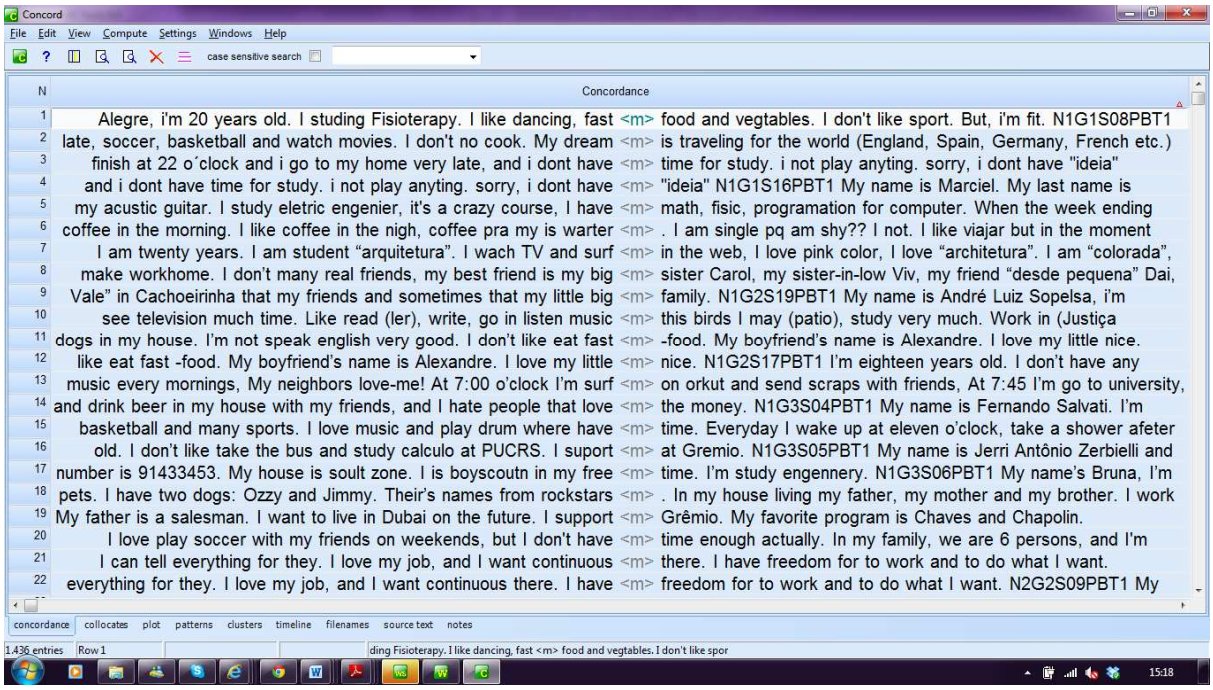
Através da observação das concordâncias da figura 3, pode-se identificar usos metafóricos da palavra de busca, como na linha de concordância abaixo (quadro 1).

Quadro 1: Linha de concordância de *have* extraída do BELC

use to work hardly during the year and **have <m>** less opportunities to be together. N4

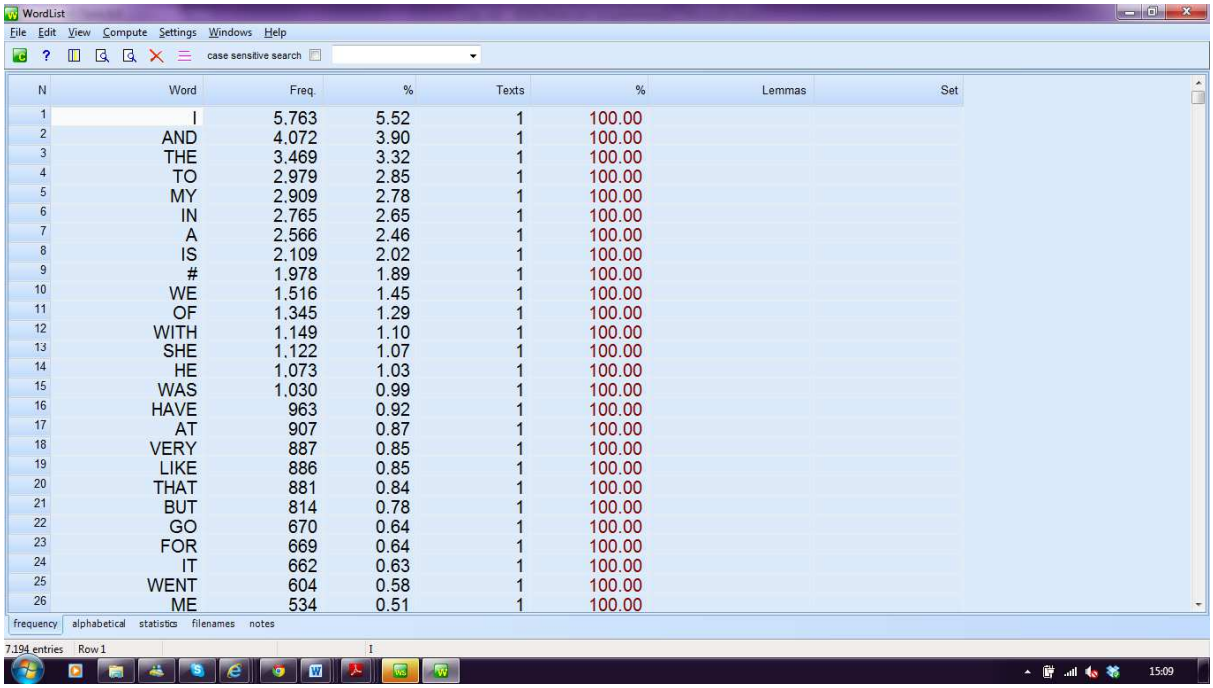
Pode-se também fazer uma busca de etiquetas (ver item 2.5.4 sobre anotação) inseridas nas palavras do corpus. No caso desta pesquisa, as ocorrências metafóricas do BELC são identificadas com a etiqueta <m>. Inserindo <m> no campo de busca, o programa busca todas as palavras anotadas com esse código (figura 4).

Figura 4: Linhas de concordância anotadas com o código <m> extraídas com o concordanciador do *WordSmith Tools*



Outro recurso é a (ii) lista de frequência de palavras que, diferentemente do concordanciador que gera resultados qualitativos, apresenta resultados quantitativos do termo de busca, permitindo o acesso e a identificação do que é comum e raro no uso da língua. Além de possibilitar o acesso à frequência de todas as palavras do corpus, tal ferramenta também possibilita a busca de palavras específicas que sejam do interesse do pesquisador. Quando se busca a frequência de todas as palavras do corpus, a ferramenta apresenta uma lista das palavras com seus respectivos números de ocorrências. Essa lista pode ser tanto ordenada a partir da palavra mais frequente até a mais rara, quanto organizada alfabeticamente. A tela abaixo (figura 5) mostra a lista das palavras mais frequentes no BELC, organizada em ordem decrescente de frequência.

Figura 5: Lista de frequência das palavras do BELC extraída com o *WordSmith Tools*



Ainda sobre as listas de frequência, na tabela 1, estão dispostas as 10 palavras mais frequentes do inglês americano, segundo a interface do COCA. Os itens são essencialmente gramaticais e cumprem papel funcional no discurso. A presença de preposições ressalta o padrão *noun + preposition + noun,* comum no uso da língua (*the side of the car*, por exemplo) (O'KEEFFE *et al.*, 2007).

Tabela 1: Lista das 10 palavras mais frequentes do COCA[18]

Ordem de frequência	**Palavra**
1	*The*
2	*Be*
3	*And*
4	*Of*
5	*A*
6	*In*
7	*To*
8	*Have*
9	*To*
10	*It*

[18] Disponível em: http://www.wordfrequency.info

Outra possibilidade no uso de listas de frequência é a comparação do número de ocorrências em diferentes corpora ou nos subcorpora de um mesmo corpus. Granger (2002) afirma que as evidências oferecidas por corpora complementam tantas outras. Entretanto, salienta que, com relação à frequência, a LdC é a única fonte confiável de evidências desta natureza.

A (iii) lista de colocados de uma determinada palavra ou frase permite a identificação das combinações de palavras com alta frequência de uso. Portanto, essa ferramenta permite a identificação das palavras que tendem a co-ocorrer com o termo de busca. A palavra *toy* (brinquedo), por exemplo, co-ocorre com frequência com *children* (crianças), ao contrário de *men* (homens) ou *women* (mulheres) que acompanham *toy* menos frequentemente (HUNSTON, 2002). É uma ferramenta útil para encontrar palavras que tendem a ocorrer perto de verbos que não tem um significado próprio, mas adquirem significado quando usados ao lado de outras palavras, como é o caso dos verbos *have, get,* e *make* no inglês (MCCARTEN, 2007), por exemplo. A tabela 2 mostra colocados de *make,* revelando padrões de uso como *make sure, make sense* e *make difference.*

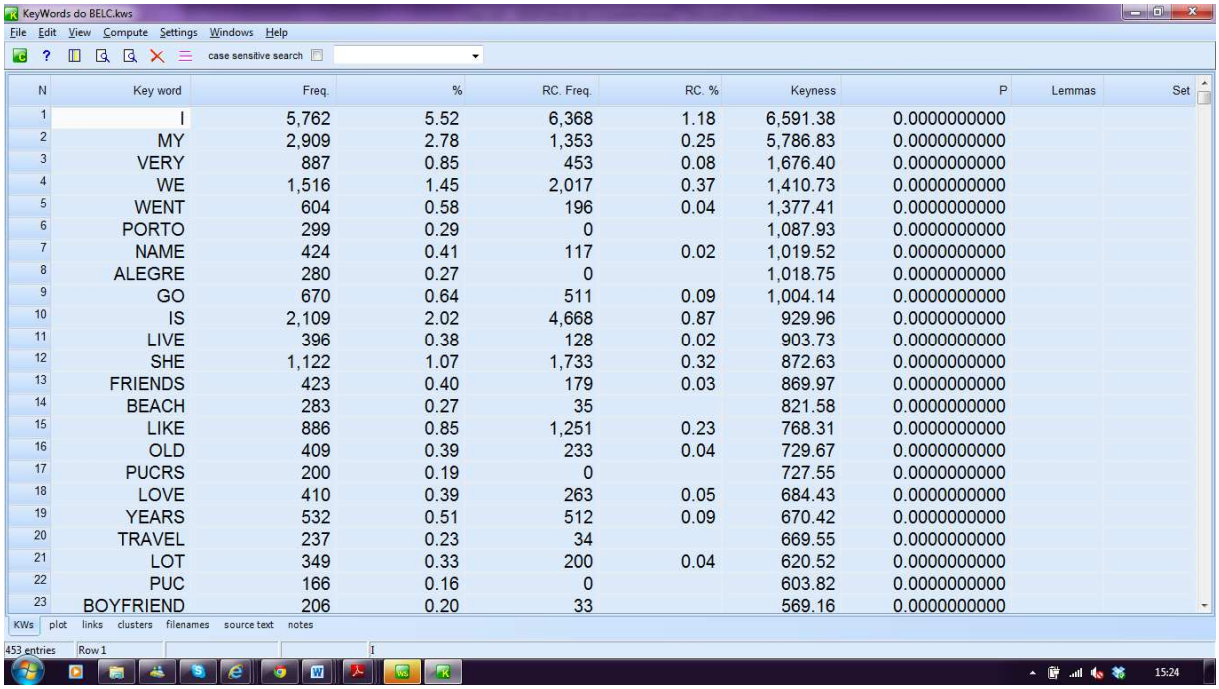
Tabela 2: Os 10 colocados de *make* mais frequentes no COCA

Make	**Colocados**
1	*Sure*
2	*Sense*
3	*Difference*
4	*Money*
5	*Decisions*
6	*Feel*
7	*Decision*
8	*Clear*
9	*Mistake*
10	*Changes*

Alguns *software* disponibilizam uma ferramenta não tão amplamente utilizada como as anteriores, mas muito útil dependendo do tipo de investigação conduzida. Esse recurso, chamado de *Keywords*, extrai as palavras-chave do corpus de estudo. Na LdC, palavras-chave são palavras particularmente características do gênero o qual se está investigando. Uma lista de palavras-chave é gerada através da comparação de listas de palavras de dois corpora, um

corpus de estudo e um corpus de referência. O corpus de estudo é o foco da análise, o corpus que se quer descrever. Neste caso, por exemplo, o corpus de estudo é o BELC. O corpus de referência é utilizado apenas para fins de comparação. Após a comparação, a ferramenta apresenta uma lista de palavras estatisticamente peculiares ao corpus de estudo. A extração de palavras-chave, portanto, requer um corpus de estudo, um corpus de referência e uma ferramenta capaz de gerar a análise estatística da comparação entre as listas de palavras dos dois corpora. A lista é organizada de acordo com a ordem de *"keyness"*, ou seja, da palavra mais significativa até a menos significativa, como na tela abaixo (figura 6), que apresenta a lista das palavras mais características do BELC, em oposição a um corpus de inglês geral.

Figura 6: Lista de palavras-chave do BELC extraída com o *WordSmith Tools*


KeyWords do BELC.kws

File Edit View Compute Settings Windows Help

case sensitive search

N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P	Lemmas	Set
1	I	5,762	5.52	6,368	1.18	6,591.38	0.0000000000		
2	MY	2,909	2.78	1,353	0.25	5,786.83	0.0000000000		
3	VERY	887	0.85	453	0.08	1,676.40	0.0000000000		
4	WE	1,516	1.45	2,017	0.37	1,410.73	0.0000000000		
5	WENT	604	0.58	196	0.04	1,377.41	0.0000000000		
6	PORTO	299	0.29	0		1,087.93	0.0000000000		
7	NAME	424	0.41	117	0.02	1,019.52	0.0000000000		
8	ALEGRE	280	0.27	0		1,018.75	0.0000000000		
9	GO	670	0.64	511	0.09	1,004.14	0.0000000000		
10	IS	2,109	2.02	4,668	0.87	929.96	0.0000000000		
11	LIVE	396	0.38	128	0.02	903.73	0.0000000000		
12	SHE	1,122	1.07	1,733	0.32	872.63	0.0000000000		
13	FRIENDS	423	0.40	179	0.03	869.97	0.0000000000		
14	BEACH	283	0.27	35		821.58	0.0000000000		
15	LIKE	886	0.85	1,251	0.23	768.31	0.0000000000		
16	OLD	409	0.39	233	0.04	729.67	0.0000000000		
17	PUCRS	200	0.19	0		727.55	0.0000000000		
18	LOVE	410	0.39	263	0.05	684.43	0.0000000000		
19	YEARS	532	0.51	512	0.09	670.42	0.0000000000		
20	TRAVEL	237	0.23	34		669.55	0.0000000000		
21	LOT	349	0.33	200	0.04	620.52	0.0000000000		
22	PUC	166	0.16	0		603.82	0.0000000000		
23	BOYFRIEND	206	0.20	33		569.16	0.0000000000		

KWs plot links clusters filenames source text notes

453 entries Row 1

15:24

O *WordSmith Tools* é um dos *software* para análise de corpora que possui a ferramenta *KeyWords*. Quanto à dimensão do corpus de referência, Berber Sardinha (2005) salienta que é uma das características que pode influenciar a extração de palavras-chave. O autor (*Ibidem*) investigou a influência do tamanho do corpus de referência na obtenção de palavras-chave através do *WordSmith Tools* e concluiu que a confiabilidade dos resultados é garantida quando o corpus de referência é cinco vezes maior que o corpus de estudo, pois corpora

maiores não interfeririam no número de palavras-chave. No que diz respeito à composição do corpus, sugere-se incluir textos de diversos gêneros, pois as particularidades de cada gênero exercem influência sobre as palavras que podem vir a se tornar chave (BERBER SARDINHA, 2004).

Se os corpora forem compilados seguindo as características básicas já mencionadas ((i) representatividade; (ii) amostragem; (iii) formato eletrônico; e (iv) autenticidade), os recursos acima descritos são capazes de gerar informações valiosas às mais diversas abordagens linguísticas. Biber *et al.* (1998) argumentam que uma abordagem baseada em corpus proporciona ferramentas e métodos eficazes que podem ser aplicados a quase todas as áreas da linguística.

A ampla variedade de aplicação, acurácia e consequente riqueza de informações oferecida pelo estudo da língua através de exemplos reais de uso proporciona acesso ao que de fato ocorre natural e autenticamente em situações de utilização da língua. As evidências empíricas provenientes do uso de corpora fornecem ao pesquisador informações confiáveis às quais a introspecção sozinha não seria capaz de chegar. A LdC possibilita então que, ao invés de observar o que é teoricamente possível em uma língua, o pesquisador acesse o que ocorre naturalmente em situações de uso e perceba as escolhas que os usuários fazem ao utilizar a língua (BIBER *et al.*, 1998).

2.5 TERMOS DA LINGUÍSTICA DE CORPUS

Nesta subseção, descrevo alguns termos utilizados na LdC, os quais serão utilizados ao longo desta pesquisa.

2.5.1 ***Token***

Como forma de ilustrar o que é *token*, utilizarei o texto abaixo (quadro 2), retirado do BELC.

Quadro 2: Exemplo de texto do BELC

Filipe, my boyfriend, is very important in my life. We are together for two tears and four months. He is a person very mature, inteligent and sensible. This are the characteristics that make me like him. He is the same age of me and he works and study too. I love tell to people how we met. He was one of the musics that played with me in the band of old songs. He is a very good music. However, he has a little strong head and it's dificult, because he always thinks he's right. We love go to the cinema, restaurants and go to travel. I hope to stay with him for a long time, but he told me once that he wants a lot to has an experience in foreign.

O texto apresenta 132 *tokens*, ou seja, 132 palavras separadas por espaço ou pontuação, incluindo as repetições de uma mesma palavra. A palavra *is*, por exemplo, aparece quatro vezes no texto. Essas quatro ocorrências são incluídas no número total de palavras do texto. Isso significa que dentre os 132 *tokens* do texto, quatro são ocorrências de *is*.

2.5.2 *Type*

O número de *types* corresponde ao número de formas distintas existentes no texto. Tomando como exemplo o mesmo texto do item anterior (quadro 2), pode-se dizer que há 82 formas distintas no texto. Isso significa que dentre essas 82 formas, nenhuma é igual a outra, ou seja, as repetições não são consideradas. A forma *is*, por exemplo, considerada quatro vezes na soma do número de *tokens* é considerada uma única vez na quantidade de *types* do texto.

2.5.3 *Type-token ratio*

O valor *type-token ratio* (TTR) corresponde à divisão do número de *types* pelo número de *tokens* e ilustra a variação lexical de um corpus. Ou seja, quanto maior o valor TTR, maior a riqueza de vocabulário dos textos. O valor TTR do exemplo (quadro 2) é 62,12. Esse número indica que 62,12% das palavras do texto ocorrem apenas uma vez no texto e que 37,88% delas repetem-se pelo menos uma vez. No caso de um corpus de aprendizes, este dado auxiliar a quantificar a evolução da qualidade da escrita do aprendiz. Além disso, é uma forma de monitorar a aquisição de vocabulário e o uso de formas novas ao longo do processo de aprendizagem.

2.5.4 Anotação

Na LdC, anotação é a codificação das informações linguísticas de um corpus para que as informações anotadas possam, em uma etapa subsequente, serem extraídas por programas computacionais especializados. Ou seja, a anotação se refere à inserção de etiquetas no corpo dos textos do corpus. As etiquetas inseridas no corpus são demarcadas por símbolos específicos, como, por exemplo, a etiqueta <Autor=Dimenstein> indicadora da autoria do texto e a etiqueta *casa-v*, em que *–v* indica a classe gramatical de *casa*, verbo (BERBER SARDINHA, 2004). O tipo mais comum de anotação é a anotação morfossintática, também chamada de POS *(part-of-speech) tagging* que consiste em etiquetar as palavras do corpus conforme suas classes gramaticais (adjetivo, verbo, substantivo, por exemplo) (*Ibidem*).

O tipo de informação adicionada pelas etiquetas dependerá dos objetivos da análise do pesquisador. No caso desta pesquisa, as ocorrências metafóricas do BELC serão identificadas com a etiqueta <m>. A anotação, portanto, agrega valor ao corpus e torna explícita a análise linguística do pesquisador (MCENERY *et al.*, 2007 [2006]). Contudo, a anotação linguística em qualquer nível, seja ele sintático, semântico ou discursivo, explicita a análise individual e pessoal do pesquisador e afilia o trabalho a um paradigma de pesquisa. Por conta disso, o anotador deve deixar claro os instrumentos e fundamentos subjacentes à anotação.

2.6 ABORDAGEM BASEADA EM CORPUS *(CORPUS-BASED)* E ABORDAGEM DIRECIONADA PELO CORPUS *(CORPUS-DRIVEN)*

Em termos metodológicos, existem duas abordagens principais para as pesquisas em LdC: abordagem baseada em corpus *(corpus-based)* e abordagem direcionada pelo corpus *(corpus-driven).* Uma das principais dicotomias entre as duas abordagens é que enquanto a primeira é de natureza confirmatória, a segunda é de natureza exploratória (KAUFFMANN, 2005).

Em um estudo *corpus-based*, o corpus é utilizado como fonte de exemplos e como forma de explicitar e testar conceitos, categorias, hipóteses e teorias pré-existentes. O

pesquisador costuma partir de teorias preestabelecidas e utiliza o corpus como fonte de exemplos para corroborar ou não a teoria com a qual está trabalhando. Um dos pontos positivos de uma abordagem *corpus-based* é que os exemplos utilizados na investigação são autênticos e conferem maior confiabilidade à pesquisa. Por outro lado, argumenta-se que este tipo de análise linguística não dá conta da riqueza de dados que o corpus oferece ao pesquisador (TOGNINI-BONELLI, 2001). Em oposição, uma pesquisa dirigida pelo corpus *(corpus-driven)* considera o corpus como um todo. Os dados para análise emergem do corpus e as afirmações teóricas devem refletir diretamente as evidências fornecidas (*Ibidem*). Nesse sentido, os dados e evidências que emergem do corpus durante sua manipulação são o fio condutor da análise e indicam a direção e o caminho a serem percorridos na pesquisa.

A posição adotada nesta pesquisa está em consonância com a adotada por McEnery *et al.* (2007 [2006]). Não será adotada uma posição rígida em relação às duas abordagens. O termo *corpus-based* será aqui utilizado em sentido amplo abrangendo ambas as abordagens *(corpus-based* e *corpus driven).* A escolha se justifica pela utilização de ambas as vertentes. Primeiramente, utilizo uma metodologia *corpus-driven*, pois não parto de metáforas específicas, mas considero todas as palavras do corpus na anotação. Entretanto, utilizo uma abordagem baseada em corpus quando uso os procedimentos propostos por Cameron (2003) e pelo Grupo Pragglejaz (2007) no julgamento da metaforicidade das ocorrências. Após a identificação de metáforas, os dados e evidências que emergirem do corpus durante seu processamento serão o fio condutor da análise.

3 CORPORA DE APRENDIZES

Nos anos subsequentes ao seu surgimento, a LdC foi crescendo, ganhando espaço e assumindo sua posição no campo da linguística. O amadurecimento da LdC em conjunto com dificuldades empíricas da pesquisa em aquisição de segunda língua despertou o interesse na área e foi, aos poucos, revelando a possibilidade de se estudar a língua do aprendiz através da observação de grandes quantidades de textos produzidos em contextos de aprendizagem de línguas, os corpora de aprendizes. O acesso a produções (orais ou escritas) de aprendizes oferece uma base empírica nunca antes disponível às pesquisas sobre aquisição de LE. Por esse motivo, oportunizam a identificação das dificuldades dos aprendizes e têm grande potencial de proporcionar evidências, descrições e percepções valiosas aos estudos sobre aquisição de línguas, superando algumas dificuldades até então enfrentadas em suas investigações.

3.1 CORPORA DE APRENDIZES: COMO TUDO COMEÇOU

O surgimento de corpora eletrônicos e a fácil, rápida e precisa maneira de acessá-los propiciada pelo uso do computador e pelo desenvolvimento de programas especializados para isso, fizeram surgir uma nova maneira de fazer linguística. Mas em conjunto com os avanços da LdC, foram surgindo também alguns desafios. O primeiro corpus linguístico eletrônico, o corpus *Brown*, surgiu nos anos 60. Entretanto, até o início dos anos 90, nenhum esforço havia sido feito na tentativa de compilar um corpus de linguagem autêntica de aprendizes de inglês[19]; isso representava uma lacuna no conhecimento sobre a produção desses aprendizes, dada a quantidade de aprendizes de inglês no mundo todo (GRANGER, 1998, 2003).

Em meados dos anos 90, acadêmicos passaram a reconhecer o valor dos corpora de aprendizes e das evidências que eles poderiam gerar para a descrição e o melhor entendimento da linguagem de aprendizes de línguas. Projetos foram então lançados com o intuito de

[19] Granger (1998) faz menção à compilação de um corpus de aprendizes de inglês especificamente, pois, segundo ela, a língua inglesa foi a língua mais estudada sob a perspectiva da LdC e o primeiro corpus linguístico eletrônico, o corpus *Brown*, é um corpus de inglês. Portanto, se a inexistência de corpora de aprendizes já representava uma lacuna na LdC, a falta de um corpus de aprendizes de inglês representava uma lacuna ainda maior.

preencher tal lacuna. O destaque foi o processo de compilação de três corpora: o *International Corpus of Learner English* (ICLE)[20]; o *Longman Learners' Corpus* (LLC), sendo ambos corpora de aprendizes de inglês falantes de diversas línguas maternas; e *o Hong Kong University of Science and Technology* (HKUST) *Learner Corpus*, um corpus de aprendizes chineses de inglês (GRANGER, 1998). A partir de então, grande atenção passou a ser dedicada a esse tipo de corpora, principalmente através dos trabalhos de acadêmicos e pesquisadores como Sylviane Granger, Fanny Meunier, Silvia Bernardini, Guy Aston, entre outros.

No Brasil, um projeto está sendo desenvolvido e conduzido pela professora Stella Tagnin, na USP (Universidade de São Paulo): a compilação do CoMAprend[21] (Corpus Multilingue de Aprendizes). O CoMAprend é um corpus multilíngue de aprendizes brasileiros, constituído de textos em diversas línguas (alemão, espanhol, francês, inglês e italiano) produzidos por falantes de uma única língua materna, o português brasileiro (TAGNIN e FROMM, 2008).

Pode-se perceber, nesta seção, que apesar de ser uma área ainda incipiente (tem uma história de cerca de 20 anos), o interesse de pesquisadores e acadêmicos nos corpora de aprendizes fez surgir projetos de destaque nacional e internacional. Portanto, apesar de Granger (2009) argumentar que esta é uma área que ainda está longe de ter atingido maturidade, acredito que existe um futuro promissor no que diz respeito à compilação e disponibilização de outros corpora de aprendizes proeminentes tanto no cenário brasileiro quanto no cenário internacional. Este campo da LdC coloca-se como uma nova perspectiva na abordagem de questões referentes à aquisição e aprendizagem de línguas, exercendo, através de suas descrições, impacto em áreas subjacentes como o ensino de LEs e a produção de material didático.

[20] http://www.uclouvain.be/en-cecl-icle.html
[21] http://www.fflch.usp.br/dlm/comet/comaprend.html

3.2 CORPORA DE APRENDIZES: DEFINIÇÃO

Um corpus de aprendiz é uma coletânea de textos autênticos[22] (escritos ou orais) produzidos por aprendizes de uma LE/L2 destinado a servir de base empírica às pesquisas sobre aquisição e ensino de línguas (GRANGER, 1998, 2002, 2009). As produções de aprendizes coletadas para a compilação de corpora podem ser tanto na língua materna dos informantes quanto em uma segunda língua. Se compilados em língua não nativa, os corpora podem ser de dois tipos: LE e L2 (GRANGER, 2002). As frases e orações que compõem um corpus de aprendiz não podem ser escolhidas aleatoriamente para fazerem parte do corpus, mas devem ser autênticas no sentido de não terem sido induzidas e nem passado por qualquer tipo de correção. Produções autênticas de aprendizes contêm erros e, para diversos analistas, são justamente os erros que tornam os corpora ricos para a realização de investigações e análises linguísticas. Com relação ao que foi mencionado, Granger (*Ibidem*) salienta que "não se pode utilizar o termo *corpus* para referir-se a uma coletânea de frases erradas extraídas de textos de aprendizes. Corpora de aprendizes são constituídos de extensões de discurso, as quais contêm tanto o uso correto quanto errôneo da língua"[23] (*Ibidem*, p. 9).

Existem diversos tipos de corpora. Além de poderem conter textos orais ou escritos, os corpora podem ser bilíngues ou monolíngues; compostos de textos pertencentes à língua geral ou de textos específicos de uma variedade linguística; podem retratar o uso da língua em um período específico de tempo ou o uso da língua ao longo dos anos, por exemplo. Considerando as características citadas, os corpora de aprendizes são, em sua maioria, monolíngues (apresentam apenas textos na língua alvo dos aprendizes) e compostos de textos específicos (no sentido de serem produzidos em um contexto de aprendizagem de uma L2). Outra característica das coletâneas de textos de aprendizes é que dadas as dificuldades de se compilar corpora de linguagem oral, elas são, em sua grande maioria, amostras de linguagem

[22] Como explicitado no capítulo sobre Linguística de Corpus desta dissertação, a autenticidade dos materiais de corpora tem as seguintes características: (i) os textos são em linguagem natural, produzidos por humanos, em contraposição à linguagem de computadores; (ii) os textos não podem ser produzidos para fins de pesquisa. O BELC, base empírica desta pesquisa, por exemplo, foi compilado com o intuito de ser alvo da investigação da aquisição de morfemas em inglês como LE. Apesar de a autora do corpus ser a mesma autora do trabalho, a produção dos textos que compõem o corpus não foi induzida de forma a revelar itens pelos quais procurava. Com relação ao uso do BELC nesta investigação, os textos que compõem o corpus assumem um caráter ainda mais autêntico, pois quando produzidos, não imaginou-se em nenhum momento que serviriam de base empírica a uma dissertação sobre produção metafórica em LE.

[23] Tradução minha. Texto original: *One cannot use the term 'corpus' to refer to a collection of erroneous sentences extracted from learner texts. Learner corpora are made up of continuous stretches of discourse which contain both erroneous and correct use of the language.*

escrita. O caráter sincrônico é outra particularidade. Visto que a compilação de corpora longitudinais representa um desafio para a LdC por exigir que uma população de aprendizes seja acompanhada por muito tempo, corpora de aprendizes quase sempre representam a língua de aprendizes em um período específico de tempo (GRANGER, 2002).

Por se tratar de uma variedade de língua muito heterogênea e contar com diversos tipos de aprendizes e contextos de aprendizagem, o projeto de compilação de um corpus de aprendiz precisa ser muito bem definido e exige que se estabeleçam critérios rigorosos, a fim de controlar as possíveis variáveis existentes e bem representar a língua do aprendiz em questão. Dentre as variáveis envolvidas para uma representação consistente da linguagem autêntica de aprendizes estão o contexto de aprendizagem, a língua materna dos informantes, seus níveis de proficiência e a tarefa utilizada na compilação do corpus (GRANGER, 2002). O controle das variáveis é uma tarefa trabalhosa, meticulosa e demorada. Além da coleta propriamente dita e do controle das especificidades do aprendiz em questão, é necessário organizar as informações. Os textos coletados devem ser devidamente identificados conforme o nível de proficiência dos informantes, por exemplo.

Uma dificuldade enfrentada na compilação de corpora de aprendizes diz respeito à coleta propriamente dita, principalmente quando o pesquisador não é o professor dos informantes. Berber Sardinha (2010, p. 337) salienta que é “muito complicado conseguir a colaboração de professores, coordenadores e diretores de escola, e efetivamente coletar os textos”. O autor atribui as dificuldades mencionadas a diversas razões, algumas delas estão relacionadas à falta de tempo do professor e à dificuldade de cumprir os conteúdos estabelecidos, não tendo, assim, tempo para dedicar à produção e coleta de textos em sala de aula. Outras dificuldades refletem o desinteresse na pesquisa acadêmica.

3.3 CORPORA DE APRENDIZES E AQUISIÇÃO DE SEGUNDA LÍNGUA

Os estudos sobre aquisição de L2 se concentram em explicar e trazer à tona aspectos até então desconhecidos e gerar novas percepções sobre a maneira como se dá a aquisição de uma língua que não seja a materna. Dentro do escopo das investigações sobre aquisição de L2 se encaixam questões e focos de pesquisa (i) que dizem respeito ao modo como aprendizes criam um novo sistema linguístico, mesmo com exposição limitada à língua alvo; (ii) que

tentam entender por que alguns aprendizes conseguem atingir um nível de proficiência e outros não; (iii) que objetivam compreender o motivo pelo qual a maioria dos aprendizes não consegue atingir o mesmo nível de proficiência da língua nativa; (iv) que almejam entender a natureza das hipóteses levantadas pelos aprendizes com relação às regras da língua alvo, suas relações e semelhanças com as regras da L1, por exemplo. Com base nos focos de pesquisa acima citados, pode-se perceber as diversas áreas com as quais a aquisição de L2 dialoga: psicologia, linguística, sociolinguística, análise da conversa, entre outras (GASS e SELINKER, 2008). As áreas citadas e a aquisição de línguas se ajudam mutuamente no desenvolvimento de investigações. Entretanto, até 2008, ano de lançamento da terceira edição da obra *Second Language Acquisition – An Introductory Course*, de Gass e Selinker, quando no capítulo 3 da obra, intitulado *Second and Foreign Language Data*, os autores abordam a natureza dos dados utilizados em investigações da área, nenhuma menção é feita à LdC e aos corpora de aprendizes[24].

Nessa época, a compilação de corpora e os estudos envolvendo corpora de aprendizes já eram diversos, visto que nos anos 90, haviam sido compilados os primeiros corpora de aprendizes (o ICLE, o LLC e o HKUST *Learner Corpus*) (GRANGER, 1998). Mas apesar disso, os estudiosos de aquisição de L2, especificamente, pouco se utilizavam da LdC. Até então, a maioria das investigações sobre aquisição de línguas se valia de dados experimentais e introspectivos, os quais eram, em sua maioria, os seguintes: (i) dados sobre as tentativas dos aprendizes ao utilizar a língua, através da produção ou compreensão em L2; (ii) dados intuitivos alcançados através do julgamento do aprendiz sobre a gramaticalidade de sentenças, por exemplo; e (iii) dados adquiridos através de questionários ou tarefas em que sujeitos informavam sobre suas estratégias de aprendizagem (*Ibidem*). Porém, por possuírem variáveis difíceis de serem controladas e pelas dificuldades operacionais na coleta dos textos, os dados se limitavam a quantidades relativamente baixas e provenientes de um número também baixo de informantes. O difícil controle sobre as variáveis e a quantidade limitada de dados levantam questões sobre a generalização dos resultados alcançados (GRANGER, 2002).

[24] Na obra *The Handbook of Second Language Acquisition* (DOUGHTY e LONG, 2005), pude identificar a menção a estudos baseados em corpora de aprendizes e a referência a pesquisadores de LdC, como Douglas Biber e Sylviane Granger. Portanto, não se pode generalizar quando fala-se no distanciamento entre a aquisição de L2 e a LdC.

Mark[25] (1998 apud GRANGER, 2002) faz as mesmas observações que Granger (1998, 2002), porém do ponto de vista pedagógico, salientando que alguns fatores que cumprem papel importante tanto no ensino quanto na aprendizagem de línguas foram deixados para trás em detrimento de outros. Convencionalmente, eram investigados assuntos como motivação, estilos de aprendizagem (todos relacionados a variáveis intrínsecas aos aprendizes), assuntos relacionados à língua alvo e ao aprendiz. Não desmereço a importância dos estudos acima citados para o desenvolvimento da área da qual estamos falando, mas ressalto que até pouco tempo atrás, pouco se sabia sobre a produção do aprendiz (GRANGER, 2002). Nesse sentido, a LdC e em especial os corpora de aprendizes têm aplicações diversas tanto ao estudo da aprendizagem de línguas em si, quanto ao lado pedagógico da área, que abarca tanto o ensino de línguas quanto a produção de material didático.

Desde o advento da área da LdC denominada corpus de aprendiz, muitos estudos vêm sendo desenvolvidos. Entre esses estudos, podemos citar o nome de Sylviane Granger como um expoente na área. A pesquisadora é autora de artigos diversos e livros sobre o tema, nos quais além de divulgar seu trabalho, promove os benefícios do uso de corpora de aprendizes, estimulando pesquisas na área. Em 1998, Granger lançou o livro *Learner English on Computer*. A obra apresenta uma visão global da área e aborda estudos que descrevem a linguagem do aprendiz de inglês e as aplicações pedagógicas de corpora de aprendizes. Em 2002, em conjunto com Hung e Petch-Tyson, Granger editou a obra *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Como o próprio título já deixa claro, o livro discute as aplicações e implicações do uso de corpus de aprendiz através dos estudos de diversos pesquisadores.

A seguir, descrevo o BELC, base empírica deste trabalho.

3.4 BELC – *BRAZILIAN ENGLISH LEARNER CORPUS*

Os dados utilizados nesta pesquisa são provenientes de um corpus de aprendiz, o BELC, compilado por Pacheco (2010), na tentativa de preencher uma lacuna até então existente na área: a inexistência de estudos baseados em produções autênticas, desde o nível

[25] MARK, Kevin L. **The significance of learner corpus data in relation to the problems of language teaching**. Bulletin of general education, 312, p. 77-90.

inicial, de aprendizes de inglês como LE, falantes de PB como L1. O BELC conta com produções escritas de 424 alunos de inglês geral, graduandos e graduados das mais diferentes áreas, da Pontifícia Universidade Católica do Rio Grande do Sul. Na época da coleta dos dados, os informantes realizavam o curso de inglês ou como disciplina eletiva ou como parte de um curso regular de línguas composto de oito níveis. Cada nível contou com um número de informantes que variou de 36 até 86, conforme a tabela 3.

Tabela 3: Sujeitos da pesquisa por nível do curso de inglês geral

Nível	**Informantes**
1	42
2	61
3	86
4	62
5	40
6	38
7	59
8	36

No total, os dados foram coletados em 24 turmas.

O instrumento de coleta desenvolvido por Pacheco é composto de quatros partes: (i) apresentação da pesquisa e identificação dos informantes; (ii) nivelamento; (iii) textos; e (iv) coleta de dados propriamente dita.

Na primeira parte da pesquisa, os alunos foram informados sobre a pesquisa na qual seus textos seriam posteriormente utilizados, assinaram um documento consentindo a utilização dos dados e preencheram uma ficha de identificações e informações pessoais, as quais poderiam vir a ser relevantes no processo de análise dos dados.

Anteriormente à aplicação do instrumento, os 424 informantes realizaram um teste de proficiência linguística em língua inglesa com o intuito de atenuar as diferenças na análise da produção escrita dos sujeitos, as quais podem surgir a partir de diferentes níveis de proficiência. O teste escolhido pela autora foi o *Placement Test* da *Oxford University Learning Center* (OULC), o qual é composto de 50 questões e tem como instrução a disponibilidade de 30 a 40 minutos para sua realização. Originalmente, a classificação de proficiência fornecida pelo OULC é a seguinte:

Tabela 4: Classificação de proficiência segundo o OULC

Score	**Classificação do OULC**
0-30	*Too low*
31-40	*English for Social or Academic Purposes*
41-50	*Advanced*

Entretanto, para os devidos fins de sua pesquisa, a autora mudou a classificação do OULC e classificou seus informantes da seguinte forma:

Tabela 5: Classificação de proficiência segundo a pesquisa de Pacheco (2010)

Score	**Classificação**
0-20	*Beginner (B)*
21-30	*Pre-Intermediate (P)*
31-40	*Intermediate (I)*
41-50	*Advanced (A)*

Os informantes foram então classificados da seguinte forma: (i) *Beginner* (iniciante) (B); (ii) *Pre-Intermediate* (pré-intermediário) (P); (iii) *Intermediate* (intermediário) (I); e (iv) *Advanced* (avançado) (A). O BELC informa tanto o nível do curso (1 a 8) em que cada informante se encontrava na época da coleta dos dados quanto o nível de proficiência (B, P, I ou A) dos sujeitos.

Para a identificação do nível do curso e do nível de proficiência dos sujeitos, a seguinte codificação foi utilizada no cabeçalho de cada texto, como no exemplo do quadro 3:

- letra N (nível do curso) + X (número do nível do curso, podendo variar de 1 a 8)
- letra P (nível de proficiência) + B, P, I ou A (de acordo com o resultado do teste de proficiência)

Quadro 3: Texto do corpus devidamente identificado

N2G1S10**PP**T3 I haven't gone far away, but I go in the last year to Parana with my family. We goes to plane and I like too much. My mother has fair, because she never goes so far, then she was nervous. We stay in a hotel, but it has a beautiful and confortable places to go and a great swimming pool. Then, I was every day in the water with my brother and sister. It was very cool! My parents goes to visit the florests and parks in the city and they loved the nature and the tranquil place. The trip was very nice, because I was with my family.

A terceira parte do instrumento são as produções que compõem o corpus. Os alunos foram orientados quanto às normas de coleta adotadas na produção dos textos. Dentre as normas a serem seguidas estavam a restrição de tempo e a não utilização de ferramentas de pesquisa (dicionários, entre outros) ou corretores digitais. Por conta disso, os informantes redigiram os textos no corpo de seus e-mails e não no *Word*. A utilização do *Word* corrigiria e indicaria automaticamente os erros cometidos pelos alunos, quando a intenção era justamente manter a autenticidade do material produzido.

O corpus é composto de três tipos de tarefa produzidos por cada informante sobre os seguintes temas e com os seguintes números aproximados de palavras:

Tabela 6: Descrição dos tipos de tarefa do BELC

Tipo de tarefa	**Tema**	**Número de palavras**
Tarefa 1	Texto descritivo com informações pessoais em 1ª pessoa	100 palavras
Tarefa 2	Texto descritivo com informações pessoais em 3ª pessoa	100 palavras
Tarefa 3	Texto narrativo sobre uma viagem que o sujeito tenha realizado	200 palavras

Depois das três primeiras partes, Pacheco passou para a coleta dos dados propriamente dita, a qual foi realizada durante o ano de 2008. Após, os dados coletados foram organizados em um banco de dados digital dividido em oito subcorpora correspondentes aos níveis do curso (1 a 8). Entretanto, para os fins desta pesquisa, o corpus foi reorganizado e dividido em subcorpora de acordo com o nível de proficiência dos informantes (B, P, I, A) e com o tipo de texto produzido (1, 2 e 3). A opção pelas duas classificações se justifica porque ambas são necessárias para a investigação da variação da produção metafórica entre níveis de

proficiência e entre tipos de tarefa. Ademais, a reorganização do corpus de acordo com o nível de proficiência dos informantes (B, P, I, A) confere mais credibilidade à pesquisa. Considerando que em cursos de inglês geral, os alunos são, em algumas situações, encaixados em níveis do curso que não correspondem fielmente às suas proficiências linguísticas, julga-se mais adequado e mais confiável classificar os textos de acordo com o nível de proficiência dos informantes que os produziram.

Em linhas gerais, o BELC é um corpus de aprendizes brasileiros de inglês como LE, falantes de português brasileiro como L1, composto de 103.593 palavras.

A compilação do BELC representa uma contribuição inquestionável, possibilitando que seus dados possam ser explorados das mais variadas formas e sob diferentes perspectivas no desenvolvimento de novas pesquisas sobre aquisição e aprendizagem de inglês como L2 no Brasil (PACHECO, 2010). Além de considerar a LdC uma abordagem adequada para o estudo da língua, acredito na acurácia de suas ferramentas e considero de extrema importância a autenticidade dos dados provenientes de corpora. Acredito também que a plausabilidade da pesquisa é atestada quando os resultados são provenientes de linguagem produzida em contextos reais de uso. O BELC foi então escolhido como fonte de dados para esta pesquisa por ter sido compilado seguindo rigorosos critérios na coleta e organização dos dados e por oferecer evidências empíricas do processo evolutivo da aprendizagem de inglês como LE, por falantes brasileiros de português como L1. Outro fator determinante para a escolha do BELC foi o caráter autêntico do material. Além de ter sido minuciosamente compilado, autora nunca cogitou a possibilidade de o corpus ser utilizado como base empírica de uma pesquisa sobre produção metafórica. Esse fator atesta ainda mais a autenticidade dos textos.

4 METÁFORA

A história dos estudos da metáfora tem um longo caminho percorrido. Este capítulo apresenta um recorte que prioriza as principais correntes teóricas. O recorte escolhido para tratar dos estudos metafóricos é baseado em Vereza (2010) e diz respeito ao *lócus* da metáfora. A autora parte da hipótese de que o principal ponto divergente entre as teorias metafóricas é justamente o *lócus*: na visão tradicional, a metáfora ocorre na linguagem; na visão cognitivista, o *lócus* da metáfora é o pensamento; e na abordagem da metáfora sistemática, o discurso.

A primeira noção de metáfora que se tem conhecimento data da Grécia Antiga. Aristóteles definiu-a como um fenômeno atrelado ao campo da linguagem: um ornamento, um artifício para embelezar a linguagem. Durante anos e anos, a visão da metáfora se resumiu a essa concepção, conhecida como tradicional. Após Aristóteles, surgiram outras vertentes de estudo da metáfora, mas todas atrelavam o fenômeno ao âmbito da linguagem. Foi só nos anos 80 que o foco mudou da linguagem para o pensamento. Essa virada paradigmática (VEREZA, 2010) é conhecida como virada cognitiva e iniciou-se com o lançamento do livro *Metaphors We Live By*, de Lakoff e Johnson (1980). Nesse livro foi apresentada a teoria da metáfora conceptual. A partir de então, a metáfora passou a ser vista como um fenômeno do pensamento. A linguagem, nessa perspectiva é apenas o espaço para a realização de um fenômeno cognitivo. Ou seja, a metáfora conceptual está na mente e licencia a metáfora linguística. Mais recentemente, sem deixar de lado os pressupostos da abordagem cognitivista, a metáfora passou a ser estudada por uma perspectiva discursiva no âmbito da linguística aplicada. A abordagem discursiva da metáfora, também chamada de metáfora sistemática, representou uma nova mudança de foco na metaforologia, do pensamento para o discurso.

Nas próximas seções serão abordadas as visões acima mencionadas. Além das teorias da metáfora, considero importante abordar a relação dessa área com a LdC.

4.1 METÁFORA NA LINGUAGEM

A primeira definição de metáfora da qual se tem conhecimento data da Grécia antiga. Foi Aristóteles que, no século IV a.C., em *Arte Poética*, primeiro abordou a noção de metáfora. Tendo como foco a linguagem, a visão aristotélica da metáfora define o fenômeno como “a transferência dum nome alheio do gênero para a espécie, da espécie para o gênero, duma espécie para outra, ou por via de analogia (ARISTÓTELES, 1997, [séc. IV a.C.], XXI, p. 42) e está presente no pensamento ocidental até os dias de hoje. Segundo Aristóteles, existem quatro tipos de metáfora: (i) transferência de gênero para a espécie; (ii) transferência de espécie para gênero; (iii) transferência de uma espécie para outra; e (iv) analogia.

No primeiro caso, transferência de gênero para a espécie, Aristóteles exemplifica com a frase “Meu barco está parado ali”. Gil (2012) relata que, hoje, esse é considerado um caso de sinonímia em que um termo é mais técnico que o outro, pois o verbo *parar* está substituindo o verbo *fundear* (mais técnico), que significa ancorar o barco. O segundo tipo, transferência de espécie para gênero, o exemplo dado por Aristóteles é “Palavra! Odisseu praticou milhares de belas ações!”. Segundo ele, há uma transposição da espécie para o gênero, pois a palavra *milhares* não está sendo utilizada em seu sentido literal, mas é empregada no lugar da palavra *muitas.* Quando utilizada na transposição de uma espécie para outra, o filósofo utiliza como exemplos “extraiu a vida com o bronze” e “talhou com o incansável bronze”. Em ambos os casos, os verbos *extrair* e *talhar* equivalem a *tirar*. O último tipo de metáfora, analogia, Aristóteles exemplifica através da frase “a velhice está para a vida como a tarde para o dia”, exemplo esse em que há uma analogia entre a velhice e uma parte do dia. Este último tipo de metáfora é o que mais se assemelha às noções de metáfora encontradas nas definições de gramáticas contemporâneas (BERBER SARDINHA, 2007b, GIL, 2012). Em todos os casos, há uma transposição do sentido literal de uma palavra. A transposição do sentido literal de um termo exige que outro termo seja utilizado em seu lugar. Assim, o significado do termo substituinte será “emprestado” ao sentido do termo substituído. Entretanto, Vereza (2010) sugere que Aristóteles não propôs uma noção clara e sistemática de metáfora e que apenas o terceiro tipo (de uma espécie para outra) pode ser vista como um caso de metáfora. Os outros tipos se relacionam a outras figuras de linguagem, como a metonímia e a hiperonímia.

4.2 METÁFORA NO PENSAMENTO

A Teoria da Metáfora Conceptual, formulada por George Lakoff e Mark Johnson e apresentada no livro *Metaphors We Live By* (1980), dá enfoque cognitivo à descrição da metáfora e a define como uma maneira de conceptualizar o mundo inerente ao pensamento humano. De acordo com Berber Sardinha (2007b), o título da obra já deixa claro o ponto principal da teoria: as metáforas são onipresentes em nossa cultura. Ao fazermos parte de uma sociedade, ao interagirmos com o mundo, ao nos expressarmos, ao entendermos e sermos entendidos, somos guiados e obedecemos *('live by')* às metáforas que fazem parte de nossa cultura.

Segundo Lakoff e Johnson (1980), a metáfora conceptual é uma forma de conceptualizar um domínio de experiência (geralmente, abstrato) em termos de um domínio mais concreto. O domínio que se deseja conceptualizar é chamado de domínio-alvo. O domínio em termos do qual o primeiro é definido é chamado de domínio-fonte. Conceptualizamos, por exemplo, ideias em termos de alimento *(Não engoli a desculpa dela)*, amor em termos de guerra *(Ele está fazendo de tudo para conquistá-la)* e vida em termos de viagem *(Carregamos uma bagagem de experiências ao longo da vida).* Os mapeamentos (relações) entre os dois domínios se estabelecem, em sua maioria, de forma inconsciente. A forma inconsciente através da qual lidamos com esse fenômeno se dá pela convencionalidade e sistematicidade no uso.

Conceitos entendidos metaforicamente são tão sistemáticos e tão "impregnados" em nossa cultura que acabamos perdendo a noção de seu caráter metafórico; ou seja, é como se a descrição (conceptualização) oferecida pela metáfora conceptual fosse objetiva. TEMPO É DINHEIRO[26] é um exemplo: acabamos perdendo de vista o que TEMPO realmente é. A conceptualização metafórica (TEMPO É DINHEIRO) parece se fazer necessária para melhor compreendermos o conceito do domínio-alvo TEMPO (LAKOFF E JOHNSON, 1980). Ou seja, "metáforas conceptuais são convencionais, quer dizer, são inconscientes... Elas não se parecem metáforas, no sentido tradicional... assim, elas se confundem com o senso comum" (BERBER SARDINHA, 2007b, p. 33). Pensar que o tempo é como dinheiro é tão natural em

[26] Está convencionalizado o uso de caixa alta em metáforas conceptuais.

nossa cultura que utilizamos expressões corriqueiras do tipo *Perdi um tempão naquela fila imensa* sem nos darmos conta de que trata-se de uma metáfora.

Segundo Lakoff e Johnson, as metáforas conceptuais estão intimamente ligadas à formação e estruturação de conceitos. De acordo com os autores, os conceitos que construímos ao longo da vida estruturam nosso pensamento, a maneira como percebemos o mundo e a maneira como nos referimos e relacionamos com outras pessoas. Ou seja, a metáfora é inerente ao nosso sistema conceptual e à maneira como percebemos, sentimos e vivenciamos o que está ao nosso redor. A metáfora como fenômeno de pensamento, e não como figura de linguagem, passa a ser um recurso cognitivo, segundo Vereza (2010), "usado, não só para se "referir" a algo por meio de outro termo mais indireto, mas, de fato, construir esse algo cognitivamente, a partir da interação com um outro domínio da experiência". Nesse sentindo, a metáfora conceptual se caracteriza como um fenômeno cognitivo que encontra na linguagem o espaço para sua realização. A metáfora conceptual subjaz à metáfora linguística.

A teoria de Lakoff e Johnson (1980) trabalha com os seguintes conceitos e definições:

- Metáfora Conceptual: Uma metáfora conceptual é uma forma de conceptualizar um domínio de experiência (geralmente, abstrato) em termos de um domínio mais concreto. Por exemplo, O AMOR É UMA VIAGEM.
- Domínio-fonte: O domínio em termos do qual o domínio-alvo é definido. Por exemplo, VIAGEM é o domínio-fonte da metáfora conceptual O AMOR É UMA VIAGEM.
- Domínio-alvo: O domínio que se deseja conceptualizar. Na metáfora conceptual o AMOR É UMA VIAGEM, o domínio-alvo é o amor.
- Expressão metafórica: a realização linguística de uma metáfora conceptual. *Nossos destinos se cruzaram* é um exemplo de expressão metafórica da metáfora conceptual O AMOR É UMA VIAGEM.
- Mapeamentos: as relações estabelecidas entre os domínios fonte e alvo. Se O AMOR É UMA VIAGEM, um dos mapeamentos possíveis é: viajantes → amantes.

4.3 METÁFORA NO DISCURSO

A metáfora no discurso, também chamada de metáfora sistemática ou metáfora em uso, é uma abordagem discursiva para o estudo da metáfora que teve início com Lynne Cameron, por volta do ano 2000. Nessa abordagem ocorre a união da linguagem, do pensamento e do uso na emergência de metáforas relativamente estáveis. Seu surgimento se deu por duas razões principais: (i) como contraponto à teoria cognitiva da metáfora e (ii) devido ao acesso à grandes quantidades de dados autênticos de língua em formato eletrônico e à programas computacionais capazes de identificar padrões sistemáticos de uso, disponibilizados pela LdC (BERBER SARDINHA, 2007b).

O surgimento da metáfora sistemática − em parte como contraponto à teoria da metáfora conceptual − tem relação com o que Cameron e Deignan (2006) apontam sobre uma abordagem cognitiva da metáfora: não se leva em conta a experiência prévia dos indivíduos com a língua, mas se dá primazia à representação mental dos indivíduos que fizeram o uso da metáfora. A metáfora sistemática, pelo contrário, coloca em primeiro lugar o uso, a recorrência e a sistematicidade das metáforas em um contexto de uso da língua, mas sem contraposição à teoria da metáfora conceptual. Cameron (2003) justifica sua proposta argumentando que o "falar" e o "pensar" não podem ser vistos como fenômenos que acontecem separadamente, mas que são conectados um ao outro e construídos em conjunto. A partir da visão discursiva da metáfora, o fenômeno passou a ser visto pela perspectiva do uso, mas nunca desconsiderando sua importância na construção de significados no âmbito do discurso. Seria, portanto, o surgimento de uma abordagem que reúne o pensamento e a linguagem (*Ibidem*). Sobre a abordagem discursiva da metáfora, Vereza (2010) argumenta que seu surgimento não constitui o retorno a uma visão essencialmente linguística da metáfora e não pode, por esse motivo, ser vista como um retrocesso na metaforologia. Em consonância com Vereza, Gil (2012, p. 66) diz que a abordagem da metáfora sistemática não nega a teoria da metáfora conceptual, "porém defende que as suposições sobre o processamento mental dos falantes são secundárias e só podem ser feitas se houver, para isso, dados relacionados ao momento discursivo em questão".

Com relação à segunda razão do surgimento desta abordagem apontada por Sardinha (2007b), a disponibilidade do uso de corpora em pesquisas, para Vereza (2007) tem superado

muitas limitações metodológicas até então enfrentadas na área. Tais limitações advinham de fatores como o uso de exemplos inventados nos estudos até então desenvolvidos, incluindo pesquisas de Lakoff e Johnson (1980). Segundo a autora, a utilização de exemplos autênticos, representativos da língua em uso, como objeto de estudo nas investigações sobre metáfora garante a legitimidade das evidências e das descrições ou explicações de algum aspecto da língua. Nesse sentido, a abordagem da metáfora sistemática encontra na LdC um aporte metodológico importante na busca por itens recorrentes e sistemáticos, fornecendo grande quantidades de dados que permitem a identificação de padrões de uso da língua.

Nas palavras de Cameron (2005[27], p. 1, *apud* BERBER SARDINHA, 2007b, p. 38), uma metáfora sistemática é "um grupo de termos ligados semanticamente (em conjunto com seus sentidos e seu afeto) de um domínio de Veículo, que são usados para falar sobre um conjunto conexo de ideias de Tópico durante um evento discursivo". As metáforas sistemáticas são sistemáticas em um determinado contexto de uso e se constroem no desenrolar do evento discursivo. Diante disso, Gil (2012) enfatiza que para identificar a sistematicidade de uma metáfora no discurso, seria necessária a observação de porções autênticas de língua em uso, que mostrassem a repetição da mesma metáfora em outras interações e comprovassem seu caráter sistemático na conexão entre expressões linguísticas e metáforas sistemáticas, não limitadas a um contexto específico.

Cameron e Deignan (2006) também salientam que fatores pragmáticos e afetivos no uso de metáforas são inseparáveis de sua forma léxico-gramatical. Portanto, é da interação entre fatores pragmáticos, afetivos e léxico-gramaticais no uso que emergem os chamados metaforemas. Um metaforema é então um conjunto de padrões relativamente estáveis de uso da língua que combina fatores linguísticos, cognitivos, afetivos e socioculturais (*Ibidem*). De acordo com Berber Sardinha (2007b, p. 41), "metaforema é uma metáfora linguística que possui uma forma estável e recorrente e se associa regularmente com um sentido semântico pragmático". Cameron e Deignan (2006) exemplificam a emergência *online* do metaforema *lollipop trees* no discurso de uma sala de aula, enquanto a professora observava desenhos de árvores de seus alunos de 9 a 11 anos. Observando círculos desenhados no topo de linhas verticais que um aluno havia desenhado, a professora comenta que não se pareciam com árvores reais, mas tinham aparência de *lollipop trees.* A partir daquele momento, a turma

[27] CAMERON, Lynne. **Metaphor Course Handout.** São Paulo: Pontifícia Universidade Católica de São Paulo, 2005.

passou a utilizar a expressão metafórica, o que demonstra o caráter estável da metáfora no contexto em questão, pelo menos naquele momento. Naquela sala de aula, *lollipop trees* adquiriu além de uma forma léxico-gramatical estável, estabilidade em relação a fatores cognitivos, afetivos e socioculturais.

A terminologia utilizada para análise da metáfora em uso parte dos termos criados por I. A. Richards[28] (1936 *apud* BERBER SARDINHA, 2007b) e é a seguinte:

- Veículo: em uma metáfora linguística, o Veículo é a parte usada em sentido metafórico naquele contexto.

- Tópico: Tópico é a parte da metáfora linguística à qual o Veículo se refere. É a parte não metafórica de uma metáfora linguística.

- Domínio de Veículo/de Tópico: áreas de conhecimento referentes ao Veículo/Tópico. O domínio de Veículo corresponde ao domínio-alvo da metáfora conceptual e o de Tópico, ao domínio-fonte.

Em *A gente precisa vestir a camiseta da empresa*, o Veículo é a porção metafórica da metáfora linguística: *vestir a camiseta*, uma vez que não está sendo usada no sentido literal de realmente vestir uma peça de roupa, mas sim de lutar e engajar-se pela empresa. Já o restante da sentença é a parte não metafórica que se refere ao Veículo, o Tópico: *a gente precisa.* Nesse caso, o Tópico diz respeito aos funcionários da empresa.

Apresento agora um quadro proposto por Berber Sardinha (2007b, p. 44) que contrasta os principais pontos da metáfora conceptual e da metáfora sistemática.

[28] RICHARDS, Ivor A. **The Philosophy of Rethoric**. New York-London: Oxford University Press, 1936.

Quadro 4: Pontos contrastantes entre a teoria da metáfora conceptual e a abordagem da metáfora sistemática

Teoria da metáfora conceptual	**Abordagem da metáfora sistemática**
O termo ‘metáfora’ significa ‘metáfora conceptual’, que é mental e abstrata.	O termo ‘metáfora’ representa ‘metáfora em uso’, que é verbal e concreta.
Ênfase no individual, idealizado.	Ênfase no sociocultural, coletivo, concreto.
Foco na cognição humana.	Foco no uso linguístico.
Interface com a linguística cognitiva, a psicolinguística e a filosofia.	Interface com a análise do discurso, linguística aplicada e linguística de corpus.
Linguagem idealizada. Exemplos inventados ou colecionados. Dados linguísticos são secundários.	Linguagem em uso. Exemplos retirados de corpora autênticos. Dados linguísticos são centrais.
Os critérios para identificação da metáfora na linguagem não são claros.	Critérios para identificação de metáfora na linguagem são claramente definidos.
Busca de validação psicológica por meio de experimentos controlados em laboratório.	Realidade psicológica é suposta por meio da evidência do uso linguístico.
Tendência generalizante: as metáforas conceptuais são formuladas de modo genérico (em ‘o amor é uma viagem’, não especificamos o tipo de amor nem o tipo de viagem).	Tendência particularizante: as metáforas sistemáticas são formuladas de modo particular, de acordo com as evidências de uso (dependendo dos participantes e dos usos metafóricos feitos por eles, poderíamos especificar o tipo de viagem e o tipo de amor: ‘amor entre marido e mulher é uma viagem sem volta’).
Interesse pelo universal. Tentativa de entendimento de características universais do ser humano ou do comportamento de grandes grupos humanos (cultura ‘americana’, ‘ocidental’, ‘humana’, etc.)	Interesse pelo local. Tentativa de entendimento do comportamento de grupos ou indivíduos específicos (pessoas ou comunidades em contextos determinados) ou de tipos de discurso específicos.
Mapeamentos entre domínios são estáveis e previsíveis.	Mapeamentos são emergentes, não previsíveis, construídos em contextos específicos.
Pensamento tem precedência sobre o uso. A linguagem é secundária, pois é apenas uma manifestação do pensamento. Pensamos metaforicamente, portanto falamos metaforicamente.	Uso tem precedência sobre pensamento. Inferências sobre o pensamento devem ser cuidadosas. Há ainda muitas questões abertas sobre o uso de metáforas; por isso, é muito problemático fazer asserções sobre o pensamento a partir das metáforas na linguagem.

4.4 METÁFORA E LINGUÍSTICA DE CORPUS

Nas seções anteriores, ficou evidente o longo caminho percorrido nos estudos da metáfora desde a primeira noção, proposta por Aristóteles, na Grécia antiga. Primeiramente caracterizada como uma figura de linguagem com papel meramente decorativo no discurso, passou, nos anos 80, a ser vista como parte fundamental da cognição, guiando a maneira como vivemos, construímos conceitos e aprendemos ao longo da vida. Por volta do ano 2000, o foco passou do pensamento para o discurso. A partir de então, a metáfora deixou de ser apenas um fenômeno do pensamento e começou a ser estudada por um viés discursivo. Todas as noções até então desenvolvidas, apesar de seus pontos divergentes, ressaltam o caráter onipresente da metáfora.

Se o discurso é permeado por metáforas, um corpus não seria diferente. Entretanto, os estudos de metáfora em corpora são recentes e iniciaram com Alice Deignan, por volta do ano 2000. Até então, pesquisas sobre metáforas não baseadas em corpora costumavam utilizar experimentos ainda hoje usados em algumas pesquisas. Nesse tipo de experimento é solicitado que os participantes interpretem textos. São normalmente apresentados dois textos, um metafórico e outro que transmite a mesma ideia em sentido literal. Os tempos de reação dos participantes aos dois textos são medidos e, em cima desses dados, as hipóteses do pesquisador são avaliadas. Entretanto, os textos utilizados são geralmente inventados e não refletem itens que são de fato frequentes na língua (DEIGNAN, 2008). Deignan (2005) coloca o uso de exemplos inventados como um dos pontos negativos nos estudos sobre a teoria cognitiva da metáfora. A grande maioria das metáforas linguísticas utilizadas nesses estudos é informada por sujeitos, os quais tendem a produzir exemplos raros em situações naturais de uso da língua. Dados coletados através de testes e experimentos são fontes valiosas de contribuições para o avanço do conhecimento sobre metáfora, mas somente os dados de corpora são capazes de gerar percepções inatingíveis pela intuição. Diante disso, Deignan (2008) coloca que o pressuposto principal de sua obra *Metaphor and Corpus Linguistics* é o de que dados linguísticos naturais, não produzidos com o intuito de exemplificar determinada teoria, são preferíveis em relação a dados intuitivos.

Ainda sobre o uso de dados de corpora e de dados intuitivos, Deignan (2005) salienta que a memória humana apresenta algumas limitações em oposição à memória do computador

e de ferramentas computacionais. Além desse fator, o fato de seres humanos não conseguirem ser precisos ao descreverem seu próprio desempenho na língua também se coloca como um ponto a favor do uso de corpora. Deignan inclusive cita a experiência de colegas linguistas de corpus e lexicógrafos que relatam ter encontrado em seus estudos usos e padrões que jamais teriam intuído. Dessa forma, a LdC auxilia os pesquisadores a chegar a análises menos subjetivas. Deignan (2008) salienta a importância da LdC no acesso à fatos da língua que de outra maneira permaneceriam escondidos. A autora aborda a possibilidade de corpora gerarem novas percepções sobre a língua e a maneira como eles estão auxiliando para um melhor entendimento da metáfora.

Como forma de ilustrar isso, Deignan (2008) cita alguns estudos conduzidos por Gibbs[29] (1994) e Lakoff[30] (1987), os quais se propuseram a investigar o uso metáforas de temperatura para conceptualizar sentimentos, principalmente a raiva em termos da pressão de um fluído em um container aquecido. Suas pesquisas concluíram que essa metáfora é utilizada para descrever comportamentos e experiências individuais na conceptualização metafórica de raiva. Entretanto, posteriormente, pesquisas sobre a mesma metáfora investigada por Gibbs e Lakoff, baseadas no uso de corpora, demonstraram que a conclusão dos pesquisadores não corresponde ao que de fato ocorre em situações autênticas de uso da língua, mas que metáforas de calor são normalmente utilizadas em contextos em que a raiva é experienciada coletivamente.

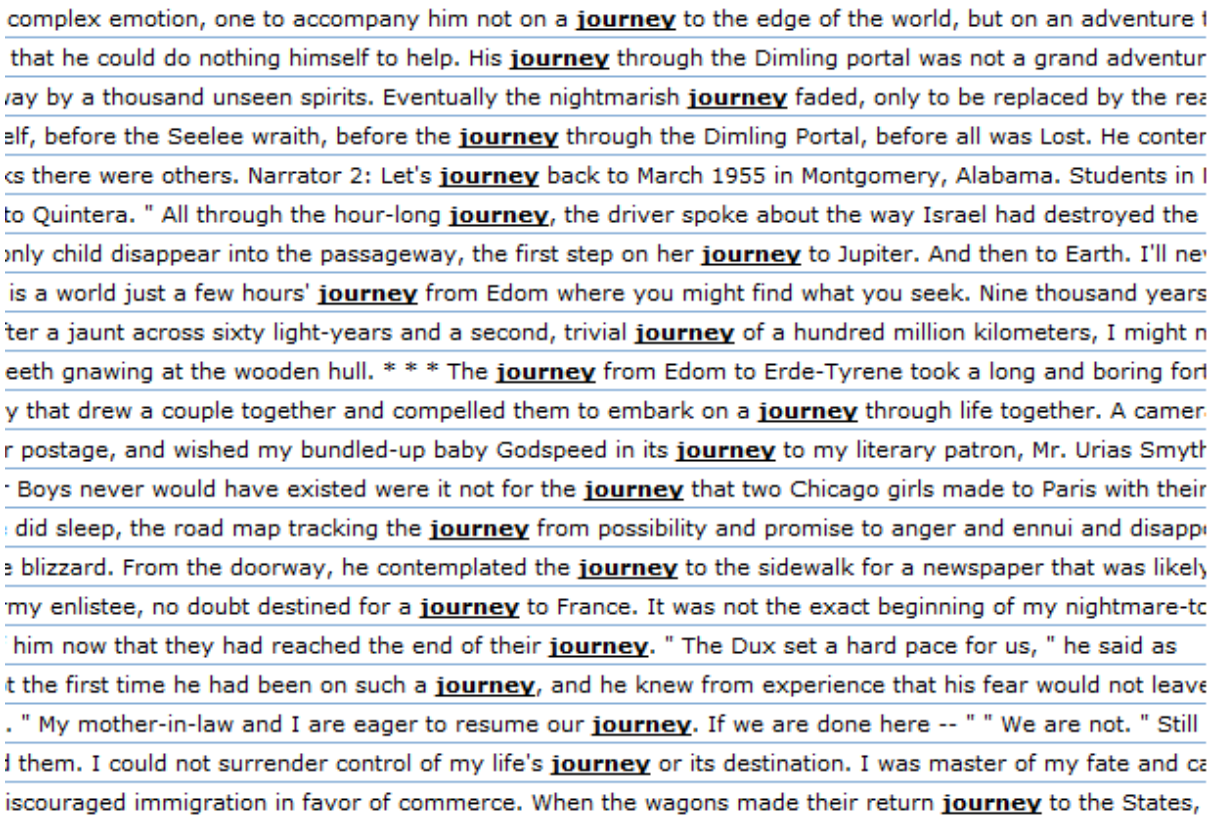
Segundo Berber Sardinha (2007a), tanto o uso de corpora eletrônicos quanto a teoria cognitiva da metáfora (LAKOFF e JOHNSON, 1980) mudaram consideravelmente o contexto dos estudos sobre metáforas. A teoria de Lakoff e Johnson, por seu caráter inovador, mudou radicalmente a concepção de metáfora até então concebida. De uma ferramenta poética e retórica passou a um fenômeno cognitivo de conceptualização do mundo inerente ao pensamento humano. Já o surgimento de corpora eletrônicos foi determinante por proporcionar outras maneiras de analisar metáforas em grandes corpora. Entretanto, as metáforas conceptuais, enquanto processos cognitivos, são fenômenos abstratos e, portanto, tornam-se um desafio para a linguística de corpus. A busca por metáforas em corpora eletrônicos se dá, então, através das expressões metafóricas resultantes dos mapeamentos

[29] GIBBS, Raymond W. **The poetics of mind**: figurative thought, language, and understanding. Cambridge: Cambridge University Press, 1994.
[30] LAKOFF, George. **Women, fire, and dangerous things:** What categories reveal about the mind. Chicago: Chicago University Press, 1987.

entre diferentes domínios. Outro fato que torna a busca por metáforas através de ferramentas computacionais possível é o fato de haver convencionalidade e recorrência no uso, características importantes na LdC. Entretanto, ainda resta ao analista julgar a metaforicidade das ocorrências. Há também casos em que o pesquisador não parte de uma lista preestabelecida de metáforas, mas inicia sua anotação sem metáfora alguma em mente. Pesquisas dessa natureza são, normalmente, de abordagem *corpus-driven*, o que é o caso desta investigação. O levantamento de todas as ocorrências metafóricas de um corpus pode ser realizado com o auxílio de programas especializados.

Os *softwares* disponíveis para a análise de corpora estão em um estágio de desenvolvimento já avançado. Contudo, apesar desse alto grau de inteligência, as ferramentas utilizadas na manipulação de corpora não são humanas e conforme Berber Sardinha (2007b, p. 12), “a metáfora é um recurso tão humano que talvez seja a última coisa que os robôs do futuro entendam”. É evidente que concordanciadores, listas de colocados e programas identificadores de metáforas têm muito a contribuir com os estudos da metáfora, mas a análise e a manipulação do corpus sempre terão um aporte humano no sentido de julgar a metaforicidade das ocorrências. Como se sabe, concordanciadores e listas de colocados buscam por formas de palavras específicas escolhidas pelo pesquisador. Se digitarmos a palavra *journey* no concordanciador do corpus geral do COCA, por exemplo, teremos acesso à seguinte figura:

Figura 7: Linhas de concordância de *journey* extraídas do COCA


complex emotion, one to accompany him not on a **journey** to the edge of the world, but on an adventure t
that he could do nothing himself to help. His **journey** through the Dimling portal was not a grand adventur
vay by a thousand unseen spirits. Eventually the nightmarish **journey** faded, only to be replaced by the rea
elf, before the Seelee wraith, before the **journey** through the Dimling Portal, before all was Lost. He conter
cs there were others. Narrator 2: Let's **journey** back to March 1955 in Montgomery, Alabama. Students in I
to Quintera. " All through the hour-long **journey**, the driver spoke about the way Israel had destroyed the
nly child disappear into the passageway, the first step on her **journey** to Jupiter. And then to Earth. I'll ne
is a world just a few hours' **journey** from Edom where you might find what you seek. Nine thousand years
ter a jaunt across sixty light-years and a second, trivial **journey** of a hundred million kilometers, I might n
eeth gnawing at the wooden hull. * * * The **journey** from Edom to Erde-Tyrene took a long and boring fort
y that drew a couple together and compelled them to embark on a **journey** through life together. A camer
r postage, and wished my bundled-up baby Godspeed in its **journey** to my literary patron, Mr. Urias Smyth
Boys never would have existed were it not for the **journey** that two Chicago girls made to Paris with their
did sleep, the road map tracking the **journey** from possibility and promise to anger and ennui and disapp
e blizzard. From the doorway, he contemplated the **journey** to the sidewalk for a newspaper that was likely
my enlistee, no doubt destined for a **journey** to France. It was not the exact beginning of my nightmare-to
him now that they had reached the end of their **journey**. " The Dux set a hard pace for us, " he said as
t the first time he had been on such a **journey**, and he knew from experience that his fear would not leave
. " My mother-in-law and I are eager to resume our **journey**. If we are done here -- " " We are not. " Still
l them. I could not surrender control of my life's **journey** or its destination. I was master of my fate and ca
iscouraged immigration in favor of commerce. When the wagons made their return **journey** to the States,

Entretanto, não existe uma ferramenta avançada de busca que possibilite a restrição da pesquisa a ocorrências metafóricas. Voltando ao exemplo de *journey*, a tela gerada pelo concordanciador do COCA apresenta ocorrências da palavra no corpus, sejam elas metafóricas ou literais. A metaforicidade das ocorrências será julgada pelo pesquisador através da observação do co-texto da palavra nódulo. Nas linhas de concordância da figura 1, por exemplo, identificou-se uma ocorrência metafórica na linha de concordância *that drew a couple together and compelled them to embark on a **journey** through life together*, em que a palavra *journey* é utilizada como forma de falar sobre a vida. Um uso literal foi identificado em *immigration in favor of commerce. When the wagons made their return **journey** to the States*, pois aqui a palavra está sendo empregada no sentido de viagem.

Da mesma forma que os dados de corpora podem ser utilizados nos estudos de metáforas através de linhas de concordância, os dados também podem ser analisados através dos colocados à direita ou à esquerda da palavra de busca. Sobre as colocações da palavra nódulo, Deignan (2005) salienta que são um ótimo ponto de partida para a análise de ocorrências metafóricas, mas que requerem uma análise manual cuidadosa, já que os mesmos

padrões de colocação podem ocorrer tanto no sentido metafórico quanto no sentido literal da palavra. Outra ferramenta de corpus muito útil na busca por metáforas é o programa identificador de metáforas. Diferentemente do concordanciador e dos colocados, o programa é destinado especificamente à análise de metáforas. O identificador de metáforas trabalha com a probabilidade de uso metafórico. O programa analisa todas as palavras do corpus colocando uma etiqueta em cada uma delas. A informação apresentada na etiqueta é um número que varia de 0,0001 (0,01%) a 1 (100%) e indica a probabilidade da palavra etiquetada ser ou não uma metáfora. Da mesma forma que o uso do concordanciador não dispensa a análise manual do pesquisador, o programa identificador de metáforas também não exime o analista do julgamento da metaforicidade das palavras etiquetadas, visto que o programa trabalha com probabilidades e não com certezas. Berber Sardinha (2012) analisa a versão 4 do *Metaphor Candidate Identifier* (MCI) e identifica pontos positivos e negativos na sua utilização. Se por um lado, o programa permite a exploração de grandes corpora[31], por outro lado, não exclui a análise qualitativa do analista humano que precisa julgar a metaforicidade das ocorrências de acordo com seus contextos de uso.

Os exemplos corroboram a ideia de que o significado das palavras, segundo a LdC, se desenvolve nos seus usos e que a observação do contexto das ocorrências na tela do concordanciador auxilia a detectar o significado das palavras pelo analista. Entretanto, para o julgamento da metaforicidade dos itens, são necessários métodos confiáveis que excluam a arbitrariedade das decisões do analista.

Já se sabe o quanto a LdC é pertinente aos estudos sobre metáfora. Programas concordanciadores e listas de frequência são ferramentas úteis na busca por metáforas, pois evidenciam padrões reais de uso e revelam ocorrências nem sempre lembradas pelo pesquisador. Porém, conforme Deignan (2008), os pesquisadores de metáforas em corpora lidam com a dificuldade de estabelecer um ponto de partida para suas pesquisas. Enquanto as ferramentas computacionais utilizadas na manipulação dos dados de corpora requerem um termo de busca para dar início à investigação, os estudiosos de metáforas não se interessam por expressões fixas, mas por padrões de língua de forma mais ampla e profunda. Como forma de superar esses desafios, Deignan (*Ibidem*) oferece algumas alternativas: (i) ler e identificar os itens linguísticos de interesse, em um corpus pequeno; (ii) partir de palavras do

[31] Segundo Berber Sardinha (2012), na análise de metáforas, um corpus com mais de 100 mil palavras já pode ser considerado grande, devido ao trabalho manual envolvido em sua análise qualitativa.

campo semântico da(s) metáfora(s) que se está investigando; (iii) tomar como ponto de partida expressões metafóricas subjacentes à metáfora conceptual de interesse listadas na literatura e, em seguida, fazer a busca de seus colocados.

4.5 VARIAÇÃO DE USO DA METÁFORA

A variação é um dos aspectos linguísticos abordados após o início dos trabalhos com corpora. O estudo da variação entre linguagens produzidas em diferentes situações abarca comparações de diferenças linguísticas entre gêneros ou registros. Através desses estudos, identifica-se como a língua de fato acontece em cada gênero/registro e de como os padrões de uso se alteram em linguagens produzidas em diferentes situações. No que diz respeito à variação de uso de metáfora, a frequência é um dos parâmetros mais abordados. Probabilidades de uso de linguagem metafórica variam de acordo com gêneros/registros específicos. Da mesma forma, variedades especializadas apresentam probabilidades diferentes de ocorrências de metáforas em relação à língua geral.

Berber Sardinha (2011a) discute aspectos da pesquisa sobre metáfora do ponto de vista da LdC. Segundo o autor, há poucas pesquisas sendo desenvolvidas na área. Berber Sardinha (*Ibidem*) salienta que as teorias da metáfora assim como pesquisas já desenvolvidas chamam atenção para a ubiquidade da metáfora na linguagem, mas que tais postulados precisam ser provados com base em corpora. O pesquisador tem desenvolvido investigações na área, as quais sugerem que metáforas não são uniformemente distribuídas em tipos textuais diferentes e que certas metáforas são mais particulares de gêneros/registros específicos do que da língua como um todo. Cameron (2003) investigou a metáfora no discurso educacional e encontrou probabilidade de ocorrência de uma metáfora a cada 37 palavras. Berber Sardinha (2012) analisou o uso metafórico em narrativas autobiográficas, a pesquisa indicou o uso de uma metáfora a cada 115 palavras. Os resultados das pesquisas mencionadas sugerem que narrativas pessoais tendem a apresentar menor probabilidade de ocorrência metafórica.

No que se refere a esta investigação, pretende-se apresentar uma análise que mostre resultados consistentes sobre a frequência da produção metafórica por aprendizes brasileiros de inglês como LE, falantes de PB como L1, numa abordagem baseada em corpus. Tendo em

vista que estudos já realizados revelam variação considerável no uso de metáforas de acordo com o tipo de língua e o contexto em que são produzidos, acredita-se que durante o processo evolutivo de aprendizagem de uma LE, a produção metafórica se configura num nível crescente de frequência. Da mesma forma, espera-se que tipos textuais diferentes apresentem frequências peculiares a cada um.

5 METODOLOGIA

5.1 ESCOPO, OBJETIVOS E QUESTÕES DE PESQUISA

Esta é uma pesquisa quantitativa de análise de dados que pretende investigar a produção metafórica por aprendizes brasileiros de inglês como LE, falantes de português brasileiro como L1. A produção metafórica será investigada, no BELC, numa abordagem baseada em corpus. O objetivo desta pesquisa é verificar se há variação na produção de metáforas em LE com relação ao nível de proficiência e ao tipo de tarefa. Espera-se que as contribuições do presente trabalho proporcionem um melhor entendimento sobre o processo de produção metafórica em inglês como LE em diferentes níveis de proficiência e tipos de tarefa. Espera-se também que auxiliem a suprir a carência de pesquisas no que se refere à produção de metáforas em LE e ao uso de corpora de aprendizes nas pesquisas sobre produção metafórica em LE.

As perguntas de pesquisa deste trabalho são:

1. Aprendizes brasileiros de inglês como LE, falantes de PB como L1, como evidenciado pelo BELC, produzem metáforas?
2. Há variação na frequência da produção metafórica no corpus de estudo com relação ao nível de proficiência linguística em LE?
3. Há variação na produção de metáforas no corpus de estudo com relação ao tipo de tarefa?

As hipóteses que norteiam este trabalho são:

1. Aprendizes brasileiros de inglês como LE, falantes de PB como L1, produzem metáforas.

2. Há variação na produção metafórica com relação aos níveis de proficiência linguística, sendo que quanto mais avançado o nível, maior o número de ocorrências metafóricas.

3. Há variação na produção metafórica com relação ao tipo de tarefa, sendo que probabilidades de uso da linguagem metafórica variam de acordo com tipos textuais específicos.

5.2 DELIMITAÇÃO DA UNIDADE DE ANÁLISE

Em um primeiro momento, a unidade de análise no corpus será a metáfora linguística[32]. A identificação de metáforas conceptuais subjacentes (LAKOFF e JOHNSON, 1980) ou metáforas sistemáticas no discurso (CAMERON, 2003) é secundária. Os dados serão analisados de acordo com as evidências proporcionadas durante a manipulação do corpus. De acordo com Berber Sardinha (2007b, p. 148), "os critérios de reconhecimento de metáforas vão sendo criados a partir dos próprios dados", ou seja, esses critérios são dinâmicos e informados pelos próprios dados. Sendo a busca por metáforas linguísticas o ponto de partida da análise, a retomada dos estudos em metaforologia, abordados no capítulo 4, se justifica por esta pesquisa não se limitar à simples visão do que é metáfora. A noção de metáfora como símile não daria conta da riqueza de evidências oferecida pela língua em uso. Deignan (2005), por exemplo, mostra que a concepção de metáfora como predicação (A é B) presente em muitas teorias é pouco frequente em porções reais de linguagem. Portanto, não parto de teorias preestabelecidas, mas tomo conhecimento delas e coloco-as à disposição do leitor, a fim de enriquecer a análise dos dados que emergirem do corpus.

5.3 MÉTODOS BÁSICOS NA BUSCA POR METÁFORAS

Berber Sardinha (2007b) apresenta quatro métodos básicos para encontrar metáforas: (i) pela introspeção do linguista; (ii) pela leitura do corpus; (iii) pelo uso do concordanciador; e (iv) pelo uso de programa computacional identificador de metáforas. Os métodos citados

[32] Uma metáfora linguística é uma oração ou um enunciado que contém palavras usadas metaforicamente (BERBER SARDINHA, 2007b). O enunciado *O dólar caiu em relação ao euro* dito por um empresário em uma reunião de negócios, por exemplo, é um exemplo de metáfora linguística, pois contém palavras usadas metaforicamente *(dólar caiu)*. É relevante ressaltar que todas as teorias da metáfora apresentadas no capítulo 4 trabalham com o conceito de metáfora linguística.

têm seus pontos positivos e negativos e podem ser combinados de acordo com a necessidade e o objetivo do analista na busca por metáforas. Dentre os quatro métodos abordados, o (i) e o (ii) são essencialmente manuais, enquanto o (iii) e o (iv) são assistidos por computador, mas não dispensam análise manual do pesquisador. Além dessas características, pode-se dizer que os métodos (ii), (iii) e (iv) têm como foco o uso real da língua, enquanto que a (i) introspeção aceita exemplos inventados (*Ibidem*).

Um dos problemas enfrentados na identificação e na análise de metáforas em corpora é a dificuldade de lidar com grandes quantidades de textos. Essa dificuldade metodológica reside no fato de não existirem ferramentas avançadas de busca que possibilitem a identificação precisa de ocorrências metafóricas no corpus. Como já mencionado, existem métodos que, através de *softwares*, fazem um levantamento de palavras com probabilidade metafórica. Entretanto, esses procedimentos não eximem o analista de uma leitura cuidadosa. Quando o pesquisador opta por não utilizar métodos dessa natureza, a leitura do corpus como um todo é a alternativa. Dentre todos os métodos, a leitura pelo pesquisador ou analista é, evidentemente, o mais subjetivo e por isso, precisa ser feita mais de uma vez e, se possível, por mais de uma pessoa, a fim de garantir a confiabilidade da anotação (BERBER SARDINHA, 2007b). A leitura e releitura pelo analista e por outras pessoas são possíveis quando a quantidade de textos para análise é pequena. Se o corpus for extenso[33], a anotação torna-se um processo demorado que pode ser prejudicado pelo cansaço do pesquisador (*Ibidem*). Por mais atenta que a leitura seja, sempre apresentará alguma falha na identificação e anotação. Além disso, a análise de corpora extensos dificulta a releitura do corpus e a validação da anotação. Inicialmente, procurou-se por procedimentos que realizassem uma triagem inicial das palavras do corpus e que excluíssem a leitura do corpus como um todo; esses procedimentos foram buscados em razão da subjetividade da leitura e da anotação manual do corpus e, ainda, do BELC ser considerado um corpus extenso na análise de metáforas (BERBER SARDINHA, 2012).

[33] Segundo Berber Sardinha (2012), na análise de metáforas, um corpus com mais de 100 mil palavras já pode ser considerado extenso, devido ao trabalho manual envolvido em sua análise qualitativa.

5.4 A ESCOLHA DO MÉTODO: OBSTÁCULOS E DESAFIOS

Nos primeiros momentos de reflexão sobre o método básico (BERBER SARDINHA, 2007b) mais apropriado a ser utilizado na busca e anotação de metáforas no corpus de estudo, não se pensou na quantidade de obstáculos que apareceriam no caminho. O método utilizado foi a leitura e anotação manual de metáforas no corpus através dos procedimentos de Cameron (2003) e do Grupo Pragglejaz (2007). A seguir, relato os problemas enfrentados na escolha do método e descrevo o método propriamente dito.

Inicialmente, a ideia era utilizar o programa identificador de metáforas. O *software* etiquetaria as palavras do corpus de acordo com a probabilidade de cada palavra realizar um Veículo. Existe uma versão do programa identificador do Centro de Pesquisa, Recursos e Informação em Linguagem (CEPRIL)[34] disponível *online*[35]. Porém, no momento da anotação do corpus, o *software* apresentava problemas e, portanto, sua utilização não foi viável. Diante da indisponibilidade de uso do programa, foi preciso pensar em outra forma de busca. A ideia inicial era evitar a leitura do corpus, que ocasionaria outros problemas, como a quantidade de tempo dedicada ao processo em um corpus de cerca de 100.000 palavras e a maior subjetividade da anotação.

A saída encontrada para a triagem do corpus e seleção de palavras potencialmente metafóricas foi a utilização de uma metodologia de cunho *bottom-up/corpus-driven* que se baseia na identificação de palavras-chave do corpus e seus colocados e na extração das linhas de concordância, a fim de chegar-se a um conjunto de palavras com probabilidade de uso metafórico (BERBER SARDINHA, 2006, 2007c, 2011b). A opção pela metodologia mencionada se justificou por duas razões. Em primeiro lugar, o procedimento parecia atender à necessidade de evitar a leitura do corpus como um todo e auxiliar na seleção inicial de palavras com provável uso metafórico. A segunda razão estava associada ao fato desta dissertação estar sendo desenvolvida em consonância com os pressupostos da LdC. Diante disso e da gama de ferramentas existentes para a análise de corpora, considerei enriquecedor para a pesquisa me aproveitar das ferramentas disponibilizadas pela LdC.

---

[34] O CEPRIL é um centro de pesquisa ligado ao Programa de Pós-Graduação em Linguística Aplicada e Estudos da Linguagem (LAEL) da Pontifícia Universidade Católica de São Paulo (PUC-SP).
[35] http://www.corpuslg.org/tools/metaphor_tagger_2.1/

A extração das palavras do corpus de estudo foi feita através do programa computacional *WordSmith Tools* (SCOTT, 2012). O *software* foi criado em 1996, por Mike Scott, da Universidade de Liverpool, no reino Unido. É composto por diversas ferramentas e se destina ao processamento e à análise linguística baseada em corpora. As ferramentas do programa são: (i) *KeyWords*, que extrai palavras-chave do corpus; (ii) *WordList*, que apresenta os colocados da palavra de busca; e (iii) *Concord*, que mostra todas as linhas de concordância em que a palavra de busca ocorre. Para a extração das palavras-chave do BELC, foram utilizadas as ferramentas *Wordlist* e *KeyWords*. Entretanto, após a manipulação inicial do corpus e a análise das palavras levantadas, observou-se que, em contraposição ao que se havia imaginado, a lista das palavras mais características do corpus de estudo não revelou itens potencialmente metafóricos. Dentre as palavras selecionadas, *went, name, go, is* e *live* foram as cinco primeiras da lista de acordo com seus valores de *keyness.* Ao buscar pelos colocados e pelas linhas de concordância dessas palavras, percebeu-se que seus usos eram quase todos literais.

O insucesso desse método criou a necessidade de partir para o método que, no princípio, se desejava evitar: a leitura do corpus como um todo e anotação manual das metáforas encontradas.

5.5 LEITURA E ANOTAÇÃO MANUAL DO BELC

Diante da inexistência de um modelo específico para a identificação de metáforas em corpora de aprendizes de LEs, procurou-se por métodos criteriosos que, apesar de não específicos para a análise da língua de aprendizes, conferissem confiabilidade à pesquisa. O primeiro passo foi estabelecer os limites que seriam utilizados no momento de anotar ou não um item como metafórico, já que o julgamento da metaforicidade das ocorrências exige critérios bem delimitados e específicos de identificação. Em uma breve leitura sobre metáforas, observou-se que diversos teóricos vêm desenvolvendo, ao longo dos anos, métodos rigorosos de identificação de ocorrências metafóricas que garantam consistência na análise e evitem decisões arbitrárias (CAMERON, 2003, DEIGNAN, 2005, PRAGGLEJAZ, 2007, STEEN *et al.*, 2010).

Cameron (2003) discute critérios para a operacionalização do conceito de metáfora linguística. Segundo a pesquisadora, há dois elementos necessários para a identificação de metáforas:

- Existência de um termo metafórico (Veículo) semântica ou pragmaticamente incongruente em relação ao seu co-texto.
- Resolução da incongruência através de uma transferência de significado do Veículo para o Tópico.

Em um primeiro momento foram identificadas palavras que poderiam estar sendo usadas metaforicamente (Veículos) no discurso. A identificação se deu através da incongruência semântica ou pragmática da palavra em relação ao discurso à sua volta. Após a identificação dos Veículos, verificou-se se a incongruência poderia ser resolvida através da transferência de significado do Veículo para o Tópico.

Entretanto, essas duas condições não são suficientes para os objetivos deste trabalho, pois não são capazes de excluir alguns casos não metafóricos (CAMERON, 2003, BERBER SARDINHA, 2007b). Seguindo no objetivo de estabelecer critérios específicos e rigorosos para a identificação de metáforas, o procedimento proposto por Cameron (2003) foi aliado ao MIP *(Metaphor Identification Procedure)*, um método para a identificação de palavras usadas metaforicamente no discurso (GRUPO PRAGGLEJAZ, 2007). O MIP foi escolhido, pois é considerado um dos métodos mais confiáveis na identificação manual de metáforas. Steen *et. al* (2010) salientam que o método é resultado de seis anos de trabalho e que sua confiabilidade foi rigorosamente testada. Um dos aspectos do MIP considerado importante nesta pesquisa é o foco no discurso naturalmente produzido, o qual condiz com os pressupostos da LdC.

O MIP consiste em:

1. ler o texto para compreender seu sentido geral;
2. definir as unidades lexicais do texto;
3. a. determinar o significado de cada unidade lexical no contexto;

b. para cada unidade lexical, verificar se há um significado mais básico em outros contextos além do contexto em questão (significados mais básicos tendem a ser mais

concretos, relacionados ao funcionamento do corpo, mais precisos – em oposição a vagos – e historicamente mais antigos). Significados mais básicos não são necessariamente os mais frequentes da unidade lexical;

c. verificar se a unidade lexical tem um significado atual mais básico em outros contextos que não o contexto em questão e decidir se o significo contextual se opõe a ele, mas pode ser entendido em comparação a ele;

d. se sim, marcar a unidade lexical como metafórica.

Estes mesmos procedimentos foram utilizados por Gil (2012) na investigação da reflexão explícita sobre a metáfora em livros didáticos de Língua Portuguesa e da ocorrência da metáfora em livros de Matemática, Ciências, História e Língua Portuguesa. Segundo a autora, os procedimentos não sanam todas as dificuldades com as quais o pesquisador se depara na identificação de metáforas, mas possibilitam um processo de identificação de metáforas linguísticas mais criterioso.

5.6 MIP X CORPORA DE APRENDIZES

O MIP visa a identificação de metáforas no discurso naturalmente produzido. Porém, o método foi desenvolvido para análise de língua materna. Esse ponto impôs desafios e limitações à anotação do BELC, pois a metodologia não prevê a existência de desvios da língua padrão e de transferências da L1 para a LE, comuns na aprendizagem de uma LE. Um dos aspectos dessa natureza evidenciado pelo BELC foi a dificuldade enfrentada pelos aprendizes no emprego de preposições, as quais parecem ser utilizadas como unidades sintáticas desprovidas de conteúdo semântico. Diante dos pontos mencionados e da inexistência de um método específico para a identificação de metáforas em corpora de aprendizes, foi necessário estabelecer critérios que auxiliassem a lidar com as peculiaridades da língua de aprendizes brasileiros de inglês como LE, falantes de PB como L1.

Além disso, foi necessário tomar decisões concernentes à definição dos limites de uma unidade lexical, assim como à maneira como lidar com expressões idiomáticas, colocações e itens funcionais no discurso. Optou-se pelos seguintes critérios:

- Unidades lexicais: O critério utilizado para a definição de unidades lexicais foi o dicionário[36]. A palavra cabeçalho de um verbete foi considerada uma unidade lexical.

- Colocações: As palavras que compõem uma colocação foram analisadas individualmente. A colocação não foi anotada/analisada como uma única unidade lexical, exceto nos casos em que apareciam nos verbetes do dicionário.

- *Phrasal verbs:* O dicionário também foi utilizado na análise de *phrasal verbs.* Quando apresentados nos verbetes, foram analisados como uma unidade lexical única. Outro motivo para o estabelecimento deste critério foi o fato de que muitos não podem ser descompostos em unidades menores sem perda de significado.

- *Multiword units:* Quando apresentadas em conjunto no cabeçalho de um verbete, foram analisadas como uma unidade lexical única.

- *Poliwords:* Expressões como *of course, let alone, at least* e *all right*, foram consideradas uma unidade lexical única.

- Expressões idiomáticas: Quando apresentadas na seção *Idioms* de um verbete, foram analisadas como uma unidade lexical única.

- Palavras lexicais x palavras funcionais: como o MIP não é um método para a identificação de metáforas em corpora de aprendizes, muitas dificuldades foram enfrentadas na anotação. Os desvios da língua padrão e transferências de significado de uma língua para outra causaram problemas na análise. Por esse motivo, a decisão foi marcar palavras lexicais, as quais, segundo Berber Sardinha (2006, 2007a), apresentam maior probabilidade de realizar Veículos. Apenas verbos, substantivos e adjetivos foram considerados na anotação. A tabela abaixo (tabela 7) foi retirada de Berber Sardinha (2007a, p.189) e corresponde ao grau de metaforicidade das classes gramaticais. Essas informações foram obtidas pelo autor com base na anotação manual

[36] O *Oxford Advanced Leaners Dictionary* foi utilizado. **OXFORD Advanced Leaners Dictionary**. Oxford: Oxford University Press. [2011] Disponível em: <http://oald8.oxfordlearnersdictionaries.com/>. Acesso em: 7 nov. 2012.

de metáforas em corpora e mostram que itens lexicais apresentam maior probabilidade metafórica em relação a itens gramaticais.

Tabela 7: Probabilidade metafórica das classes de palavras

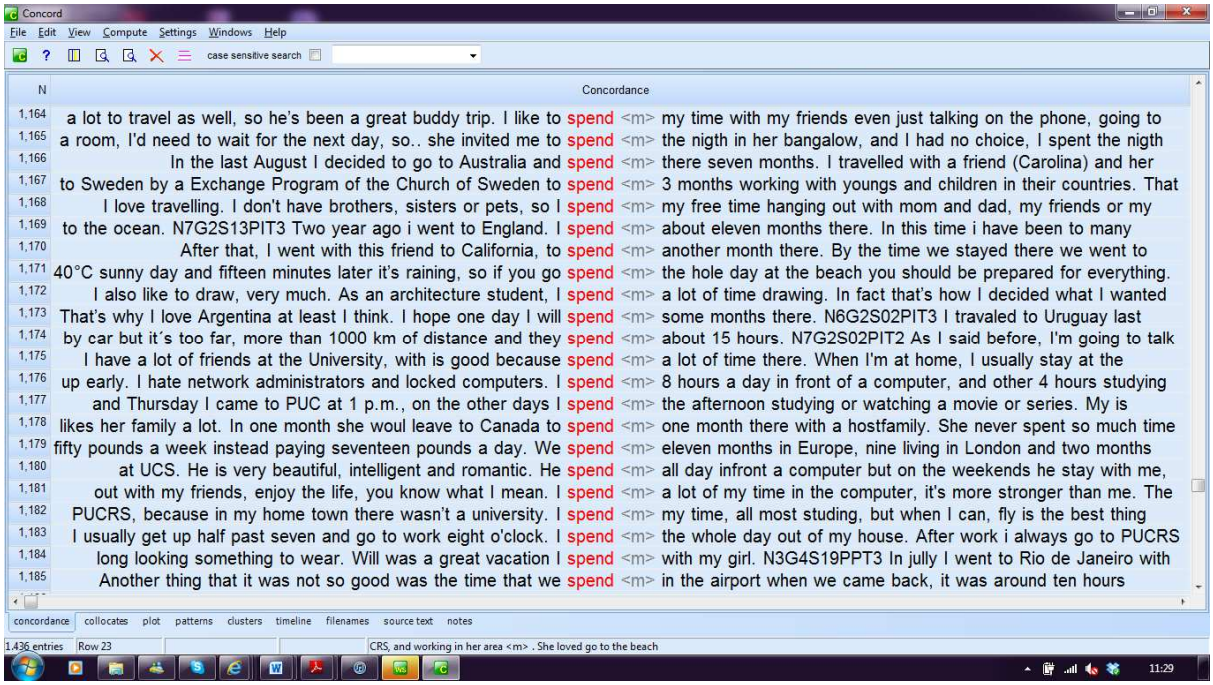
Classe de palavras	**Probabilidade metafórica**
Adjetivo	0,73
Advérbio	0,63
Artigo	0,00
Conjunção	0,00
Numeral	0,00
Pronome	0,00
Preposição	0,61
Substantivo	0,69
Verbo	0,70

5.7 ANOTAÇÃO E VALIDAÇÃO DA ANOTAÇÃO

Não há dúvidas de que a extensão de um corpus está diretamente ligada à maneira como se trabalha com ele. Na busca por metáforas através da leitura, por exemplo, quanto maior o corpus, maior a dificuldade do analista em anotá-lo. A leitura torna-se cansativa e o cansaço do pesquisador interfere na qualidade da anotação. Além disso, como salientado por Berber Sardinha (2007b), o tamanho do corpus interfere também na realização da validação da anotação. O BELC, por ter cerca de 100.000 palavras, é considerado um corpus extenso em pesquisas com metáfora (BERBER SARDINHA, 2012) em função do trabalho manual envolvido na análise.

Primeiramente, seguindo os critérios já especificados, o corpus foi anotado pelo pesquisador. As formas metafóricas foram identificadas com o código <m> para que, subsequentemente, o *WordSmith Tools* pudesse buscá-las através da inserção de <m> no campo de busca do concordanciador, conforme a figura 8.

Figura 8: Extração de ocorrências identificadas com o código <m>



Após concluída a anotação, foi realizada a releitura do corpus e revisão dos itens anotados e não anotados. Procurou-se conferir maior confiabilidade à anotação através da sua validação. Com relação à validação do corpus, alguns pontos foram discutidos com o professor Tony Berber Sardinha (informação verbal)[37] durante um curso sobre metáforas e tradução por ele ministrado, na PUCRS. O pesquisador sugeriu que a validação fosse realizada com uma amostra de 1.000 palavras (1,0% do corpus). Quando questionado sobre o baixo número de palavras e a provável pouca quantidade de metáforas anotadas na amostra, Berber Sardinha salientou que considerava o número suficiente, pois no momento da anotação, a decisão de não anotar um item é tão importante quanto a decisão de anotá-lo.

A validação foi realizada em conjunto com uma colega doutora em linguística com ampla experiência em estudos de LdC. Considerou-se importante o fato de a colega, assim como a analista, ter uma perspectiva menos teórica da metáfora e mais associada ao uso. Anteriormente à anotação da amostra do corpus pela colega, os critérios adotados pela pesquisadora na anotação foram explicitados. Em seguida, foi realizada uma seção de treino

[37] Informação recebida em 31 de agosto de 2012, durante comunicação pessoal em um curso intitulado *Corpora, registers, and metaphor: What every translator should know but was afraid to ask*, ministrado pelo professor Tony Berber Sardinha, na Pontifícia Universidade Católica do Rio Grande do Sul.

que consistiu na anotação de uma amostra do corpus pela pesquisadora e pela colega em conjunto. Após o treinamento, se deu a validação propriamente dita em que cada uma das linguistas, individualmente, anotou outra amostra do corpus. A comparação dos casos anotados mostrou que ambas as anotações foram quase que 100% concordantes. Os casos divergentes foram discutidos para que se chegasse a uma conclusão sobre incluí-los ou não na anotação, sendo que durante a discussão, houve concordância em todos os casos discutidos. Dessa forma, considerou-se válida a anotação da pesquisadora.

5.8 ANÁLISE QUANTITATIVA DOS DADOS

As fases metodológicas anteriores desta pesquisa foram estruturadas visando a identificar metáforas no corpus. Esta fase, por sua vez, se dedica à análise dos dados propriamente dita.

Nesta fase da pesquisa, os dados foram analisados quantitativamente através da extração da frequência de metáforas linguísticas no BELC, nos subcorpora correspondentes aos níveis de proficiência dos informantes (B, P, I, A), aos tipos da tarefa produzidos na coleta de dados (1, 2 e 3) e aos subcorpora individuais[38]. A análise se deu da seguinte forma:

- Verificação da frequência de metáforas linguísticas no BELC e em seus subcorpora.
- Contraste das frequências de metáforas linguísticas entre os subcorpora de níveis de proficiência observando a variação na produção metafórica.
- Contraste das frequências de metáforas linguísticas entre os subcorpora correspondentes aos tipos de tarefa observando a variação na produção metafórica.
- Contraste das frequências de metáforas linguísticas entre os subcorpora individuais observando a variação na produção metafórica.

[38] Um subcorpus individual corresponde a uma tarefa específica produzida em um determinado nível. Serão observadas, portanto, as frequências de todas as tarefas em todos os níveis. O nível *Beginner*, por exemplo, foi transformado em três subcorpora: um subcorpus correspondente à tarefa 1, um à tarefa 2 e outro à tarefa 3, os quais foram identificados como B1, B2 e B3.

A extração dos dados do corpus e das frequências de metáforas foi realizada através das ferramentas *WordList* e *Concord* do *Wordsmith Tools.* Para que as verificações e contrastes mencionados fossem realizados, os textos anotados foram transformados para arquivo .txt, pois o *software* só lê arquivos nesse formato.

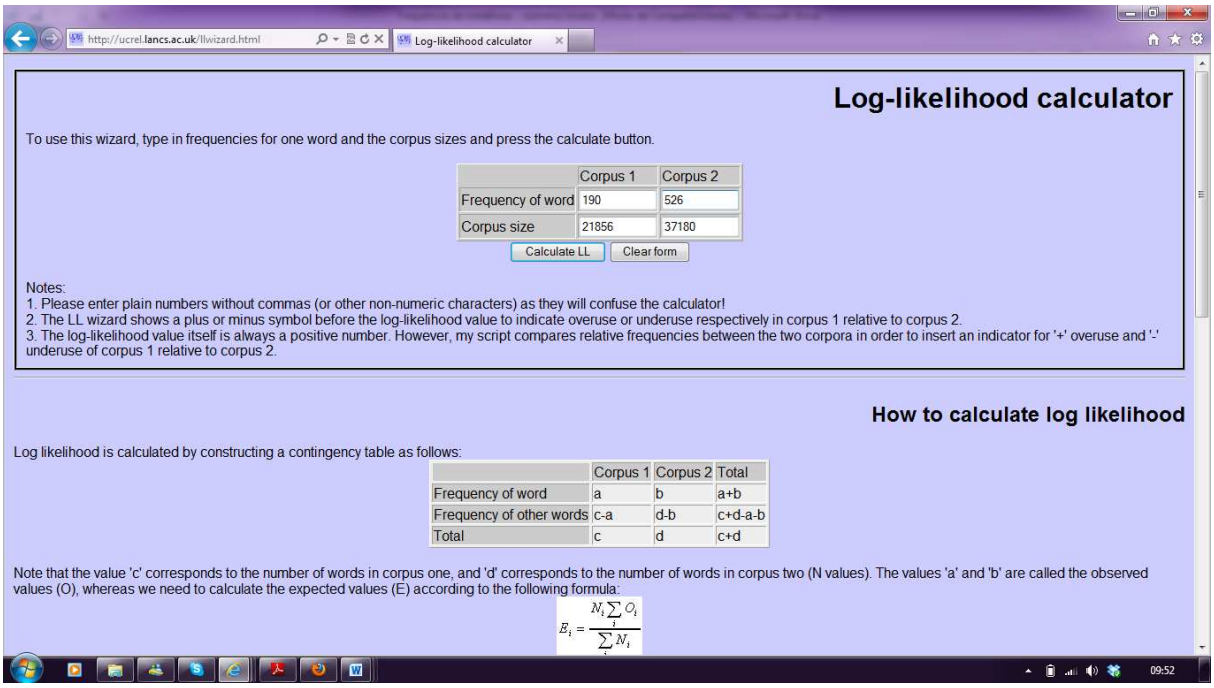
Como as dimensões dos subcorpora de níveis de proficiência, de tipos de tarefa e individuais são diferentes, para que as frequências pudessem ser comparadas, os resultados foram normalizados. O valor normalizado corresponde ao número de ocorrências de uma metáfora a cada 1.000 palavras. Para se chegar a esse número, divide-se o número total de ocorrências de metáforas pelo número total de *tokens*. O resultado da divisão é multiplicado por 1.000.

Com o objetivo de verificar se as variações de frequência encontradas são estatisticamente significativas, aplicou-se o teste estatístico *Log Likelihood* (LL). Segundo Rayson (2002), o LL é o teste estatístico com melhores resultados na comparação de frequências de itens entre dois corpora. O LL calcula a probabilidade de a diferença entre os dois corpora ser significativa ou aleatória. Se o resultado gerado for igual ou maior a 6,6, existe apenas 1,0% de chance de a diferença entre os corpora ser aleatória. Ou seja, 99,0% das chances indicam que a diferença não aconteceu aleatoriamente, mas por razões específicas. Esse resultado é normalmente expresso como $p<0,01$. A verificação estatística foi realizada com o *Log Likelihood Calculator*[39].

A tela inicial do *Log Likelihood Calculator* solicita que sejam inseridos o tamanho (número de *tokens)* dos dois corpora e o número de ocorrências do item sob investigação em cada corpus, conforme a tela abaixo (figura 9). O número de ocorrências não precisa ser normalizado, pois o LL considera o tamanho dos corpora.

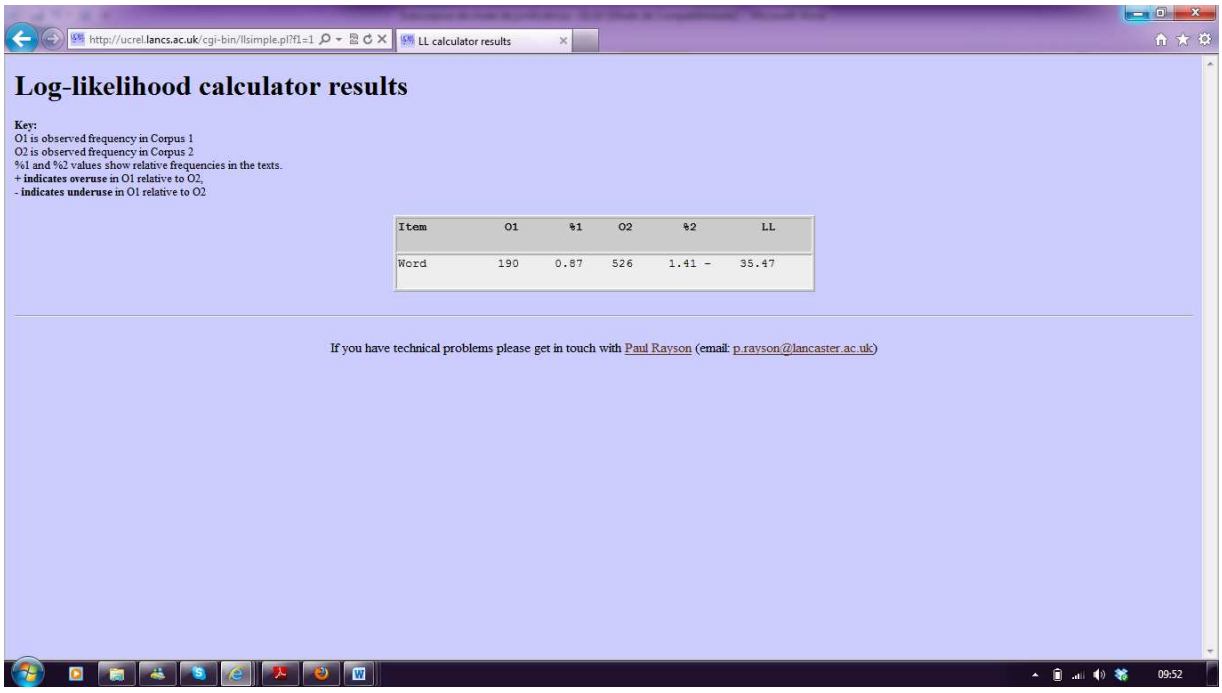
[39] O *Log Likelihood Calculator* está disponível em http://ucrel.lancs.ac.uk/llwizard.html

Figura 9: Tela inicial do *Log Likelihood Calculator*



Após, os resultados são apresentados na tela da seguinte forma (figura 10):

Figura 10: Apresentação dos resultados na tela do *Log Likelihood Calculator*



6 ANÁLISE E DISCUSSÃO DOS DADOS

Este capítulo tem o objetivo de analisar a variação da frequência das ocorrências metafóricas encontradas nos textos do BELC. Inicialmente, apresento as frequências observadas na totalidade do corpus. Em seguida, analiso o contraste das frequências entre os subcorpora de níveis de proficiência e de tarefas produzidas durante a coleta do corpus. Após, discuto as frequências encontradas nos doze subcorpora individuais do BELC. Ao final do capítulo, esboço algumas considerações de cunho qualitativo sobre pontos relevantes observados durante a anotação e ocorrências peculiares ao tipo de língua analisado.

Embora o corpus de estudo já tenha sido descrito anteriormente, considero relevante retomar a estrutura e descrição do BELC (tabela 8), visto que as próximas seções dizem respeito às suas dimensões, tarefas e níveis de proficiência linguística em LE.

Tabela 8: Estrutura e descrição do BELC

Descrição geral	O BELC (PACHECO, 2010) conta com produções autênticas, desde o nível inicial, de aprendizes brasileiros de inglês como LE, falantes de PB como L1.
Número de palavras	Cerca de 100.000 palavras.
Número de informantes	424 informantes.
Sobre os informantes	Alunos de inglês geral, graduandos e graduados das mais diferentes áreas, da PUCRS. Na época da coleta dos dados, os informantes realizavam o curso de inglês ou como disciplina eletiva ou como parte de um curso regular de línguas composto de oito níveis.
Níveis de proficiência dos informantes	*Beginner* (iniciante); *Pre-Intermediate* (pré-intermediário); *Intermediate* (intermediário); e *Advanced* (avançado).
Tarefas produzidas durante a coleta	O corpus é composto de 3 tipos de tarefa produzidos por cada informante sobre os seguintes temas e com os seguintes números aproximados de palavras[40]: Tarefa 1: Texto descritivo com informações pessoais em 1ª pessoa; cerca de 100 palavras. Tarefa 2: Texto descritivo com informações pessoais em 3ª pessoa; cerca de 100 palavras. Tarefa 3: Texto narrativo sobre uma viagem realizada pelo informante; cerca de 200 palavras.

6.1 BELC

Nesta seção, observarei a frequência de expressões metafóricas encontradas no BELC. Nesta análise, não serão considerados os níveis de proficiência dos informantes e nem os tipos de tarefa produzidos durante sua compilação. O corpus será visto como um único "arquivo", representativo da língua de aprendizes brasileiros de inglês como LE, falantes de PB como L1. Pretende-se através dos dados apresentados nesta seção, responder a primeira questão norteadora desta pesquisa: Aprendizes brasileiros de inglês como LE, falantes de PB como L1, como evidenciado pelo BELC, produzem metáforas?

[40] Os dados sobre o número aproximado de palavras de cada texto foram retirados de Pacheco (2010). Essas informações não foram verificadas nesta pesquisa.

Inicialmente, apresento uma tabela (tabela 9) na qual disponho os números de formas *(types),* itens *(tokens)* e da relação forma/item *(type/token ratio)* do BELC, pois alguns dos cálculos subsequentes serão feitos com base nesses números. Se contarmos o número total de palavras no corpus, chegaremos ao número 103.593 *(tokens).* Entretanto, dentre essas palavras existem várias que se repetem pelo menos uma vez. Como já mencionado, o número de *types* corresponde ao número de formas distintas existentes no texto, não considerando as repetições de uma mesma forma. Conforme a tabela, no BELC há 7.200 formas *(types).* A relação entre esses dois números, chamada de *type/token ratio* (TTR), corresponde à divisão do número de *types* pelo número de *tokens.* Nesse caso, o valor é 7,03. A interpretação desse número mostra que 7,03% das palavras do corpus ocorrem apenas uma vez. Ou seja, 92,97% das palavras repetem-se pelo menos uma vez no texto.

Tabela 9: Descrição do BELC em números

	Types	***Tokens***	***Type/token ratio***
BELC	7.200	103.593	7,03

Para verificar a frequência de metáforas produzidas no BELC, foi realizada uma busca geral de todas as ocorrências metafóricas do corpus, com o auxílio da ferramenta *Concord,* do *WordSmith Tools.* A frequência encontrada está disposta na tabela 10. Na primeira coluna, apresento o total bruto de ocorrências de <m> no corpus e na segunda, o valor normalizado. A tabela mostra que são produzidas cerca de 13 metáforas linguísticas por 1.000 palavras no corpus.

Tabela 10: Frequência de metáforas no BELC

Valor bruto de <m>	**Valor normalizado[41] de <m> (frequência por 1.000 palavras)**
1.436	13,86

[41] O valor normalizado corresponde ao número de ocorrências de uma metáfora a cada 1.000 palavras. Para se chegar a este número, divide-se o número total de ocorrências de metáforas pelo número total de palavras do corpus. O resultado da divisão é multiplicado por 1.000. No caso, 1.436/103.593 = 0,01386194 x 1.000 = 13,8619405.

Com base no número de ocorrências metafóricas, pode-se calcular a densidade metafórica no corpus (CAMERON, 2003), conforme a tabela 11. Esse cálculo indica que 1,38% dos 103.593 *tokens* do BELC são metafóricos.

Tabela 11: Densidade de metáforas no BELC

Metáforas	1.436
Palavras	103.593
Densidade	1,38

Essa densidade representa uma ocorrência de metáfora a cada 72 palavras[42], em média. Ou seja, a cada 72 palavras produzidas no BELC, uma é metáfora.

A frequência de linguagem metafórica pode ser confirmada através da análise de material autêntico (ver nota de rodapé 4 sobre autenticidade), o que é o caso desta pesquisa. Entretanto, considero importante salientar que o número de expressões metafóricas encontradas está atrelado à definição de metáfora estabelecida pelo pesquisador e ao método utilizado na identificação de itens metafóricos. Nesse sentido, torna-se difícil comparar os resultados encontrados por diferentes pesquisadores. Cameron (2003) encontrou probabilidade de ocorrência de uma metáfora a cada 37 palavras no discurso acadêmico. Berber Sardinha (2008), em uma investigação sobre metáforas de teleconferências de negócios, mostrou a ocorrência de uma metáfora a cada 22 palavras, em média. Berber Sardinha (2012) apresentou evidências de uso de uma metáfora a cada 115 palavras em narrativas autobiográficas.

A dificuldade de comparação entre resultados tem como consequência a impossibilidade de se estabelecer um parâmetro que permita categorizar a frequência de metáforas em corpora, seja na língua geral ou em tipos textuais específicos. De qualquer forma, apesar de não haver termos de comparação para classificar a frequência como alta/média/baixa, por exemplo, os resultados desta análise mostram que há produção metafórica durante a aprendizagem de uma LE. Acredita-se que a presença da metáfora no discurso do aprendiz se justifique tanto pela construção de sentido no texto (GEORGE e

[42] Esse valor é obtido através da divisão de mil pelo valor normalizado de <m> a cada mil palavras.

LAKOFF, 1980), quanto por razões de ornamentação da linguagem (ARISTÓTELES, 1997, [séc. IV a.C.]).

Os exemplos abaixo foram retirados do BELC e ilustram as teorias metafóricas acima mencionadas. É relevante destacar que por se tratar de um corpus de aprendiz, o corpus de estudo apresenta desvios da língua padrão, comuns na escrita de aprendizes durante o processo de aquisição, como no quadro 5, em que aparece o uso indevido do artigo indefinido *a*. No exemplo, o aprendiz utiliza *a* e não *an* como artigo indefinido para a palavra subsequente *angel.* A ocorrência metafórica apresentada no quadro 5 remete à primeira noção de metáfora, em que Aristóteles define o fenômeno como uma transferência de sentido. No exemplo, ao utilizar a palavra *angel*, o aprendiz estabelece uma relação de semelhança entre a pessoa a qual ele descreve e a figura de um anjo. No quadro 6, apresento a realização linguística de uma metáfora conceptual (LAKOFF e JOHNSON, 1980). No contexto da linha de concordância, a expressão linguística *warmful* pode ser considerada uma metáfora linguística subjacente à metáfora conceptual AFETO É CALOR. Da mesma forma, ao dizer que uma pessoa é fria, estamos utilizando uma expressão metafórica subjacente a essa mesma metáfora conceptual.

Quadro 5: Exemplo de metáfora linguística extraído do BELC

cefalia, is very, very good mannered, is a **Angel <m>** , I love she. N1G2S15PBT2 I spe

Quadro 6: Exemplo de metáfora linguística extraído do BELC

mployes and foreing students, were very **warmful <m>** . In that ocasion there were stu

Com relação à abordagem da metáfora no discurso, que teve início com Lynne Cameron, por volta do ano 2000, não foram identificadas ocorrências dessa natureza no corpus. Segundo Cameron e Deignan (2006), a metáfora discursiva ou metáfora sistemática coloca a recorrência e a sistematicidade contextual em primeiro plano. Berber Sardinha (2007b, p. 38) coloca pontos importantes sobre a abordagem, dentre os quais destaco a ideia de que o ponto de partida para o estudo de metáforas sistemáticas devem ser as metáforas

recorrentes, “que sistematicamente indiquem que os participantes de alguma interação estão ativando algum tipo de representação metafórica mental”. Essa ideia mostra que a sistematicidade só pode ser provada com base em evidências de uso que indiquem o uso sistemático de expressões metafóricas. Entretanto, para haver recorrência e sistematicidade no discurso, é necessário que o discurso tenha extensão o suficiente para dar espaço ao desenvolvimento e à construção de unidades recorrentes de sentido. Visto que os textos do BELC são curtos (cerca de 100 a 200 palavras), parece não haver espaço suficiente para a construção de metáforas sistemáticas no desenrolar do evento discursivo.

O resultado encontrado aponta para a ubiquidade da metáfora na língua em uso, indicando que o discurso de aprendizes de inglês como LE, falantes de PB como L1, também é permeado por metáforas. O fato de se chegar a esse resultado através de corpora corrobora postulados de teorias e estudiosos da metáfora (ARISTÓTELES, 1997, [séc. IV a.C.], LAKOFF e JOHNSON, 1980, CAMERON, 2003), assim como resultados encontrados em estudos anteriores (BERBER SARDINHA, 2008, 2012, CAMERON, 2003). Do ponto de vista da LdC, pode-se também analisar os números mais a fundo, a fim de verificar se essas 1.436 metáforas são uniformemente distribuídas entre os tipos textuais e entre os níveis de proficiência do BELC, ou se são mais características de um tipo de texto e mais produzidas durante um estágio específico do processo de aprendizagem da LE. A verificação da variação entre níveis de aprendizagem e entre tipos textuais encontra-se nos próximos itens.

6.2 SUBCORPORA DE NÍVEIS DE PROFICIÊNCIA

Esta seção aborda os níveis de proficiência linguística dos informantes do BELC. Antes de iniciar a análise de suas frequências de metáforas, considero importante apresentar os números dos subcorpora. Na tabela 12, estão dispostos os números de *types*, *tokens* e da relação *type/token* dos quatro corpora de níveis de proficiência. A relação *type/token* mostra que dos quatro níveis analisados, o nível avançado é o que apresenta maior variedade lexical.

Tabela 12: Descrição dos subcorpora de níveis de proficiência em números

	Types	***Tokens***	***Type/token ratio***
Beginner	2.840	21.856	13,13
Pre-Intermediate	3.690	37.180	10,03
Intermediate	3.930	39.504	10,05
Advanced	1.177	5.053	23,61

No nível inicial, 13,13% das palavras do corpus ocorrem apenas uma vez. Ao contrário do que se poderia esperar, ao invés da diversidade lexical aumentar do nível inicial ao nível pré-intermediário, há uma queda de cerca de três pontos. Entre os níveis pré-intermediário e intermediário esse número se mantém constante, cerca de 10,0% das palavras não se repetem nos textos. Já a comparação entre os níveis intermediário e avançado apresenta um aumento brusco de mais de 100%. Esse dado indica que no nível avançado, a repetição do mesmo léxico nos textos cai, sendo que o percentual de palavras não repetidas aumenta para 23,61% das palavras do texto. Esse dado sugere que nos níveis avançados há maior diversidade de vocabulário utilizado. O corpus, dessa forma, ajuda a quantificar a evolução da qualidade da escrita do aprendiz. Observa-se aqui que a extração do valor TTR é uma ferramenta útil para monitorar a aquisição de vocabulário e o uso de formas novas pelos aprendizes durante o processo de aprendizagem de uma LE.

Após algumas considerações sobre a diversidade lexical dos subcorpora, apresento a frequência de metáforas extraída de cada um deles. Disponho os resultados numa relação contrastiva entre os níveis de proficiência do BELC. Pretende-se através dos contrastes de frequência apresentados nesta seção, responder a segunda questão norteadora desta pesquisa: Há variação na frequência da produção metafórica no corpus de estudo com relação ao nível de proficiência linguística em LE?

A variação tem sido um aspecto bastante abordado em trabalhos baseados em corpora (BIBER, 1988, KAUFFMANN, 2005, BERBER SARDINHA, 2011a). Através desses estudos, identifica-se como a língua de fato acontece em cada gênero/registro e como os usos se alteram em linguagens produzidas em diferentes situações. No que diz respeito à variação de uso de metáfora, a frequência é um dos parâmetros mais abordados. Isso se dá pela facilidade e simplicidade da extração de frequência de itens em corpora. Probabilidades de

uso de linguagem metafórica variam de acordo com tipos textuais específicos. Da mesma forma, variedades especializadas apresentam probabilidades diferentes de ocorrências de metáforas em relação à língua geral.

A extração do número de metáforas produzidas em cada nível do BELC permite realizar comparações e verificar se a produção aumenta de acordo com o crescimento do processo evolutivo de aprendizagem. Os contrastes buscam verificar se existe variação na produção metafórica, de acordo com o nível de proficiência. Dessa forma, é possível perceber se a presença de metáforas é maior ou menor em cada nível e se pode ser caracterizada como mais peculiar de um dos níveis investigados. A frequência de metáforas em cada subcorpus de nível de proficiência *(Beginner*, *Pre-Intermediate*, *Intermediate* e *Advanced)* foi extraída com o concordanciador do *WordSmith Tools* e está disposta na tabela 13. Na primeira coluna, apresento o nível de proficiência linguística, na segunda, o total bruto de ocorrências de <m> em cada subcorpus e na terceira, seus valores normalizados. Os números dispostos na tabela abaixo indicam que quanto mais alto o nível de proficiência, maior a produção metafórica. Observa-se que desde o nível inicial até o nível avançado, o número de metáforas é crescente.

Tabela 13: Frequência de metáforas nos níveis de proficiência

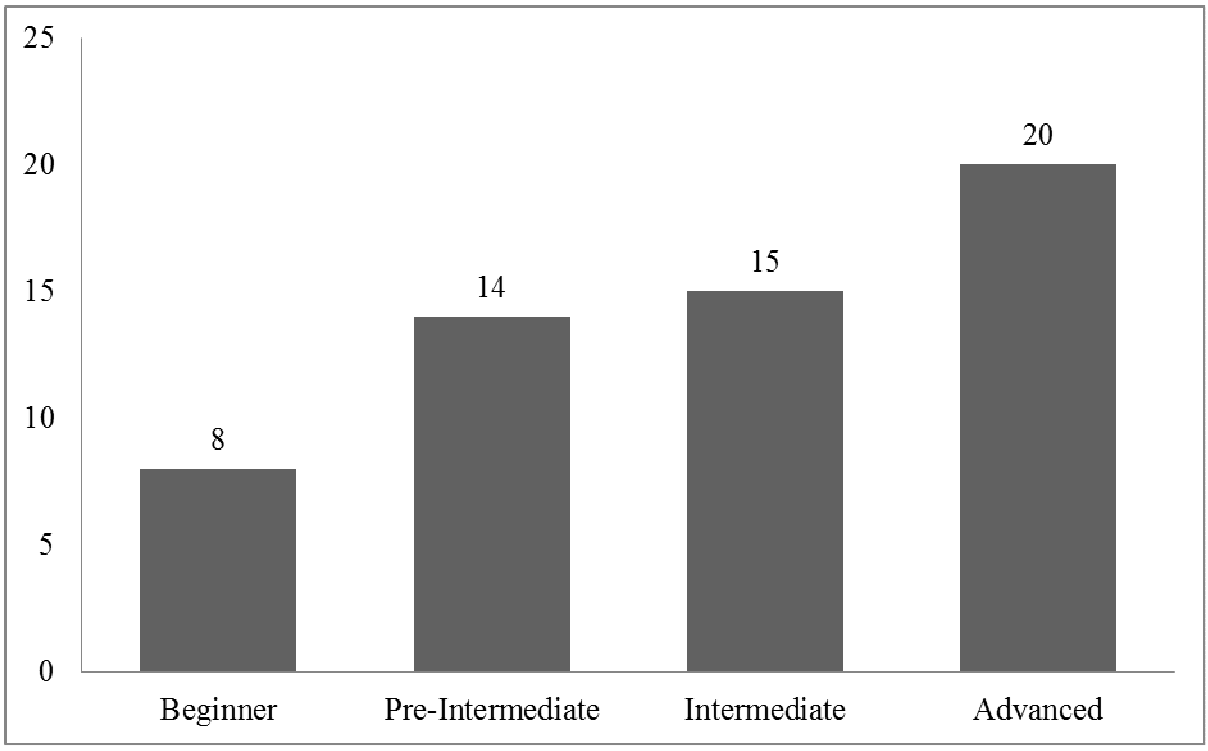
Nível de proficiência	**Valor bruto de <m>**	**Valor normalizado de <m> (frequência por 1.000 palavras)**
Beginner	190	8,69
Pre-Intermediate	526	14,14
Intermediate	617	15,61
Advanced	103	20,38

A comparação entre a produção de itens metafóricos em cada nível mostra que há variação de frequência. Entretanto, entre os níveis pré-intermediário e intermediário a variação é menor do que no contraste entre outros níveis, como o intermediário e o avançado, por exemplo.

A frequência disposta na tabela 13 pode ser melhor visualizada no gráfico 1. Entre os níveis *Beginner* e *Pre-Intermediate*; e *Intermediate* e *Advanced* os números de ocorrências apresentam variações de cerca de cinco metáforas por 1.000 palavras de um nível para outro.

No entanto, na comparação entre as frequências dos níveis *Pre-Intermediate* e *Intermediate* o valor se mantém quase que constante, apresentando variação de apenas uma metáfora por 1.000 palavras.

Gráfico 1: Frequência de metáforas por 1.000 palavras nos níveis de proficiência



Com base no número de ocorrências metafóricas e seus valores normalizados, pode-se calcular a cada quantas palavras uma metáfora é produzida, conforme a tabela 14. Se no *Beginner* uma metáfora é produzida a cada 115 palavras, no *Advanced* há uma ocorrência metafórica a cada 49 palavras, o que representa um aumento de mais de 100% do nível inicial ao nível final de proficiência no BELC. Esse dado sugere maior frequência de produção metafórica em níveis avançados, ou seja, com a evolução do nível de proficiência há também evolução na habilidade de utilizar palavras com sentido figurado pelo aprendiz.

Tabela 14: Razão de produção metafórica nos níveis de proficiência: produção de uma metáfora/palavras

Nível	**Produção de uma metáfora/palavras**
Beginner	115
Pre-Intermediate	70
Intermediate	64
Advanced	49

As frequências de metáforas apresentam variações diferentes na relação contrastiva entre os subcorpora do BELC. Primeiramente, considerando as frequências do nível inicial *(Beginner)* e do avançado *(Advanced)*, observa-se que a produção de itens metafóricos varia de forma crescente ao longo do processo de aprendizagem. Enquanto no nível inicial uma metáfora é produzida a cada 115 palavras, no nível final, uma metáfora ocorre a cada 49 palavras. Ao considerar o valor normalizado de <m> por 1.000 palavras, percebe-se que há uma diferença de quase 12 metáforas a mais produzidas no nível avançado (Básico: 8,69 e Avançado: 20,38 <m> por 1.000 palavras). Esses números revelam que a frequência de metáforas praticamente dobra do estágio inicial ao estágio final de aprendizagem. A interpretação dessa diferença sugere que ao longo dos quatro níveis, a frequência se dá de forma crescente.

Considerada a variação entre o primeiro e o último nível, analiso agora os contrastes entre as frequências de níveis imediatamente posteriores um ao outro. O contraste entre os níveis *Beginner* e *Pre-Intermediate* apresenta variação crescente. Se no primeiro, uma metáfora é produzida a cada 115 palavras, no segundo, há a ocorrência de uma a cada 70 palavras. Considerando o valor normalizado de ocorrências por 1.000 palavras em cada nível, observa-se que no nível pré-intermediário são produzidas quase seis metáforas a mais em comparação com o nível inicial. No contraste entre as frequências encontradas nos níveis pré-intermediário *(Pre-Intermediate)* e intermediário *(Intermediate)*, entretanto, observou-se diferença menor que na comparação anterior. Enquanto no primeiro são produzidas cerca de 14 metáforas a cada 1.000 palavras, no segundo, são produzidas 15. Isso corresponde a uma metáfora a cada 70 palavras no *Pre-Intermediate* e uma a cada 64 palavras no *Intermediate.* Essa variação sugere que entre esses níveis, há baixa evolução na produção metafórica na LE. Conforme a pontuação segundo a qual os aprendizes do BELC foram classificados (tabela abaixo) de acordo com seus níveis de proficiência em inglês como LE, esperava-se que a

diferença na produção metafórica de um nível para outro fosse uniformemente crescente. A classificação realizada foi baseada na pontuação disposta na tabela 15.

Tabela 15: Classificação de proficiência segundo a pesquisa de Pacheco (2010)

Score	**Classificação**
0-20	*Beginner* (B)
21-30	*Pre-Intermediate* (P)
31-40	*Intermediate* (I)
41-50	*Advanced* (A)

Observa-se, na tabela 15, que a diferença de pontos entre os níveis é a mesma em todos os níveis (10 pontos). Dessa forma, ao perceber que a produção aconteceu de forma crescente, esperava-se que essa diferença, assim como a diferença da pontuação entre um nível e outro, utilizada na classificação de Pacheco (2010), acontecesse de forma uniforme. Ou seja, que a produção aumentasse proporcionalmente de um nível para outro. Em oposição ao contraste entre os níveis pré-intermediário e intermediário, a comparação entre o intermediário e o avançado mostrou mais variação. No primeiro, como já mencionado, foram produzidas cerca de 14 metáforas por 1.000 palavras, o que equivale a uma metáfora a cada 64 palavras. No *Advanced,* observou-se a produção de cerca de cinco metáforas a mais do que no *Intermediate*, pouco mais de 20 metáforas por mil palavras. Esse número equivale a uma ocorrência metafórica a cada 49 palavras. Os números encontrados sugerem que o nível de proficiência exerce influência direta na quantidade de itens metafóricos.

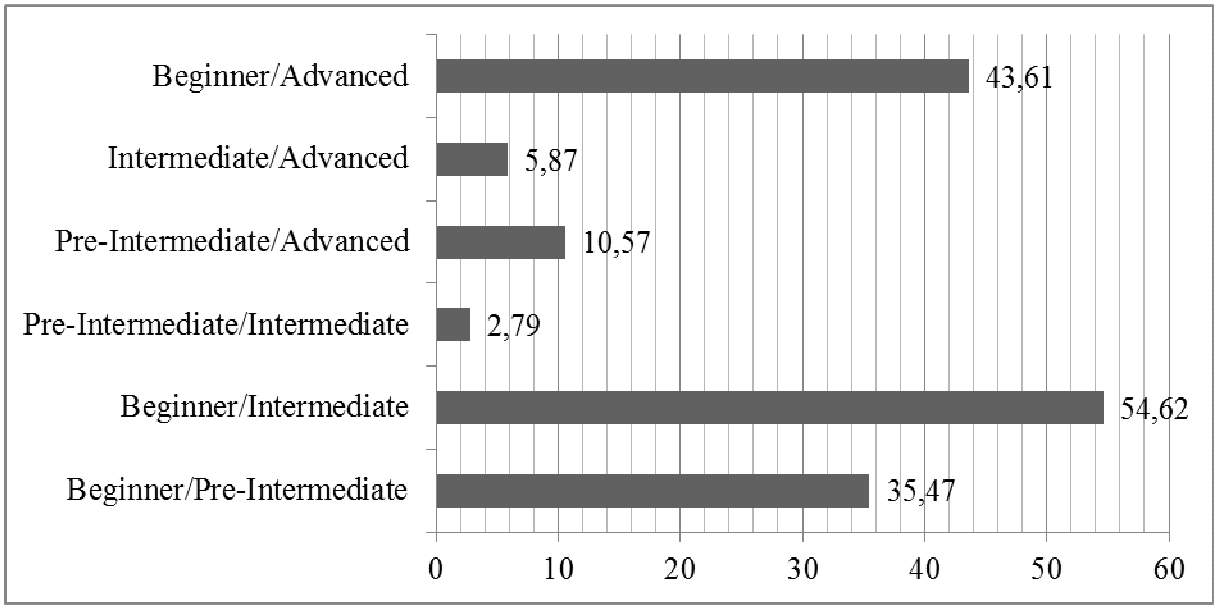
Assim como houve variação mínima na produção de itens metafóricos entre os níveis pré-intermediário e intermediário, o contraste da relação *type/token* desses níveis também mostrou diferença mínima. Retomando os resultados encontrados, a diferença do número de metáforas de um corpus e de outro apresentou diferença de apenas uma ocorrência por 1.000 palavras. Já a variação da relação forma/item (TTR) foi de 0,02%. Essas comparações parecem indicar que existe alguma ligação entre a variedade lexical do corpus (TTR) e sua frequência de metáforas, já que ambas apresentam pouca variação e que o subcorpus com percentual mais alto na relação *type/token* apresentou a frequência mais alta de itens metafóricos.

Com o objetivo de verificar se as diferenças observadas são estatisticamente significativas, aplicou-se o teste LL. Foram calculados os valores estatísticos da comparação entre os níveis da seguinte forma:

- *Beginner/Pre-Intermediate*
- *Beginner/Intermediate*
- *Pre-Intermediate/Intermediate*
- *Pre-Intermediate/Avanced*
- *Intermediate/Advanced*
- *Beginner/Advanced*

O valor 6,6 do LL é a linha divisória para verificar se as diferenças entre níveis acontecem de forma aleatória (menor que 6,6) ou se são estatisticamente significativas e apresentam alguma motivação linguística (maior que 6,6). Os números obtidos encontram-se no gráfico 2.

Gráfico 2: Comparação estatística entre as frequências de metáforas no contraste entre os níveis de proficiência



Os resultados obtidos no teste foram superiores a 6,6 em quatro dos contrastes entre níveis, o que indica 99% de chance de as variações não serem aleatórias, mas serem linguisticamente motivadas e acontecerem por alguma razão específica. Esses resultados (acima de 6,6) podem ser divididos em dois grupos: um grupo no qual houve diferença significativa de um nível para outro imediatamente após; e outro no qual as diferenças se mostraram significativas quando as comparações foram realizadas entre níveis não imediatamente posteriores um ao outro, conforme a tabela 16.

Tabela 16: Contrastes entre níveis com resultados estatísticos significativos

Contraste	**Resultados estatísticos significativos entre níveis imediatamente posteriores um ao outro**	**Resultados estatísticos significativos entre níveis não imediatamente posteriores um ao outro**
Beginner/ Pre-Intermediate	X	
Beginner/ Intermediate		X
Pre-Intermediate/ Advanced		X
Beginner/ Advanced		X

Conforme a tabela 16, observa-se que dentre as comparações realizadas, as que obtiveram valor acima de 6,6 são entre níveis não imediatamente posteriores um ao outro. Com relação a níveis imediatamente posteriores um ao outro, o único contraste com valor estatístico acima de 6,6 foi entre os níveis *Beginner* e *Pre-Intermediate.*

A verificação estatística obtida foi considerada aleatória em duas das comparações realizadas. Considerar um resultado aleatoriamente estatístico significa dizer que as variações estão atreladas a fatores incertos e que não acontecem por uma razão específica. O resultado do LL foi abaixo de 6,6 em comparações realizadas entre níveis imediatamente posteriores um ao outro, conforme a tabela 17.

Tabela 17: Contrastes entre níveis com resultados estatísticos aleatórios

Contraste	**Resultados estatísticos aleatórios em níveis imediatamente posteriores um ao outro**	**Resultados estatísticos aleatórios em níveis não imediatamente posteriores um ao outro**
Pre-Intermediate/ Intermediate	X	
Intermediate/ Advanced	X	

Acredita-se que os valores obtidos no LL possam estar relacionados à maneira como os informantes do BELC foram classificados. Como explicado em um dos capítulos anteriores, Pacheco (2010), ao compilar o BELC utilizou o *Placement Test* da *Oxford University Learning Center* (OULC). Originalmente, o teste é dividido em três níveis: (i) inglês muito baixo *(Too Low)*; (ii) inglês para propósitos sociais ou acadêmicos *(English for social or Academic Purposes)*; e (iii) avançado *(Advanced)*. O teste utilizado na compilação do corpus foi o OULC, mas a classificação de proficiência de acordo com a pontuação obtida no teste foi modificada. Os três níveis do OULC foram transformados em quatro níveis *(Beginner, Pre-Intermediate, Intermediate* e *Advanced)* pela autora do BELC. Segundo Pacheco (2010), o teste escolhido foi o OULC por ter sido considerado "neutro", no sentido de não fazer parte do material didático utilizado nas aulas dos aprendizes e por ser considerado "modelo" diante dos propósitos de muitos dos informantes (estudar no exterior). Além disso, a autora entende que outros testes poderiam ser complexos demais para boa parte dos aprendizes. Sobre a mudança na classificação dos alunos segundo suas proficiências linguísticas, Pacheco apresenta duas razões: (i) o grande número de informantes participantes da coleta do corpus; e (ii) a grande diferença de proficiência linguística neles observada. Ambas as classificações estão dispostas na tabela 18.

Tabela 18: Classificação de proficiência do OULC e do BELC

Pontuação (OULC)	**Classificação (OULC)**	**Pontuação (BELC)**	**Classificação (BELC)**
0 – 30	*Too Low*	0 – 20	*Beginner*
31 – 40	*English for Social or Academic Purposes*	21 – 30	*Pre-Intermediate*
41 – 50	*Advanced*	31 – 40	*Intermediate*
--------	----------	41 – 50	*Advanced*

A principal diferença entre uma classificação e outra é que o nível *Too Low* do OULC corresponde a dois níveis no BELC, *Beginner* e *Pre-Intermediate*. O contraste entre as variações desses níveis foi o único contraste entre níveis imediatamente posteriores um ao outro que apresentou resultado estatisticamente significativo. Esse fato despertou a curiosidade de tentar entendê-lo.

Ao observar que na classificação do BELC os níveis *Beginner* e *Pre-Intermediate* correspondem a um único nível no OULC, optei por unir os números de *tokens* e os números brutos de metáforas desses subcorpora, transformando-os em um único subcorpus. Dessa forma, o BELC seria composto por três níveis de proficiência: (i) um correspondente aos níveis *Beginner* + *Pre-Intermediate*[43]; (ii) um correspondente ao nível *Intermediate*; e (iii) outro ao nível *Advanced*, conforme a tabela 19.

Tabela 19: Números dos subcorpora de níveis de proficiência organizados conforme a classificação do OULC

Subcorpus	***Tokens***	**Valor bruto de <m>**	**Valor normalizado de <m> (frequência por 1.000 palavras)**
Beginner + Pre-Intermediate	59.036	716	12,128193
Intermediate	39.504	617	15,6186715
Advanced	5.053	103	20,3839303

Trabalhando com os números dispostos na tabela 19, foram encontrados resultados que talvez expliquem algumas das verificações realizadas anteriormente. O número de metáforas dos níveis Pré-Intermediário e Intermediário, por exemplo, eram quase iguais, apresentavam variação de cerca de uma ocorrência metafórica. Agora, no entanto, do nível *Beginner + Pre-Intermediate* para o nível *Intermediate*, há variação crescente de pouco mais de três ocorrências. Esse dado revela que, quando observados dessa forma, os corpora apresentam variação crescente de um nível para outro, sendo que essa variação ocorre de forma mais uniforme do que na análise anterior. Acredita-se que a baixa variação entre os níveis pré-intermediário e intermediário encontrada na análise anterior se deva à classificação

[43] Daqui em diante, me refiro à união dos níveis *Beginner* e *Pre-Intermediate* da seguinte forma: *Beginner + Pre-Intermediate.*

dos informantes do BELC conforme seus níveis de proficiência, visto que, originalmente, os níveis iniciais *Beginner* e *Pre-Intermediate* correspondem a um único nível.

Em resumo, a interpretação das frequências sugere que as probabilidades de uso de linguagem metafórica apresentam variação durante o processo evolutivo de aprendizes de inglês como LE, falantes de PB como L1. A variação ocorre de forma crescente. Ou seja, quanto mais alto o nível de proficiência linguística do aprendiz, maior o número de ocorrências de <m>. Considerando a forma como os informantes do BELC foram classificados de acordo com seus níveis de proficiência em inglês, esperava-se que a diferença na produção de metáforas de um nível para outro fosse uniformemente crescente. Entretanto, esse aumento não se dá de forma uniforme e proporcional entre um nível e outro, mas apresenta baixa variação entre dois dos quatro níveis. A baixa variação encontrada entre os níveis pré-intermediário e intermediário reflete fatores relacionados ao modo de classificação dos alunos segundo suas proficiências. Quando considerados os níveis originais do OULC, a variação entre os níveis *Beginner* + *Pre-Intermediate* e *Intermediate* aumenta, revelando assim variação crescente entre todos os níveis. Essa constatação mostra a relevância de medir a capacidade linguística de aprendizes através de testes de proficiência de forma a determinar o grau de competência e o domínio da LE, com base no desempenho. Além disso, esse tipo de teste proporciona informações sobre deficiências linguísticas dos aprendizes, difíceis de serem obtidas de outra forma. Esses aspectos são relevantes na formação de grupos nivelados e no monitoramento dos resultados alcançados em cada nível. Os resultados desta análise também mostraram a importância de utilizar uma classificação confiável, feita com base em rigorosos estudos teóricos e metodológicos. Para os fins da pesquisa de Pacheco (2010) pode ser que o fator classificação de proficiência não tenha gerado interferências nos resultados. No entanto, a análise da variação de ocorrências metafóricas em cada nível desta pesquisa revelou que, dependendo do que se está buscando verificar, o modo de classificação dos aprendizes segundo suas proficiências linguísticas exerce influência direta nos resultados.

Além das questões concernentes aos níveis de proficiência, observou-se que os números da relação *type/token* parecem ser um indicativo da variação da frequência de metáforas em cada nível.

6.3 SUBCORPORA DE TIPOS DE TAREFA

Esta seção aborda os três tipos de tarefa que compõem o BELC. Antes de iniciar a análise da frequência metafórica nos textos, considero importante apresentar os números de *types*, *tokens* e da relação *type/token* (TTR) dos subcorpora (tabela abaixo), assim como a temática de cada tarefa. Antes de apresentar as frequências de metáforas linguísticas extraídas dos corpora, vale dispensar atenção às informações dispostas na tabela 20.

Tabela 20: Descrição dos subcorpora de tipos de tarefa em números e descrição da temática das tarefas

Tipo de tarefa	***Types***	***Tokens***	***Type/token ratio***	**Temática**
Tarefa 1	3.714	39.026	9,4	Texto descritivo com informações pessoais em 1ª pessoa
Tarefa 2	3.079	27.280	11,42	Texto descritivo com informações pessoais em 3ª pessoa
Tarefa 3	3.775	37.288	10,20	Texto narrativo sobre uma viagem

Conforme a tabela 20, ao valor TTR da tarefa 1 é o mais baixa dos três tipos de tarefa. Nesse subcorpus, 9,4% do léxico ocorre apenas uma vez. Ou seja, 90,6% das palavras se repetem nos textos dos aprendizes. Em seguida, aparece a tarefa 3, em que 10,20% do número de *tokens* do corpus não se repetem. A tarefa com maior diversidade lexical é a 2, em que 11,42% das palavras ocorrem apenas uma vez dentre todos os 27.280 itens do corpus. Esses números parecem estar associados aos tipos de tarefa e suas temáticas, assim como a aspectos gramaticais e lexicais de cada um. O tipo de tarefa 1, por exemplo, é a descrição de informações pessoais em 1ª pessoa. O tema desse tipo de tarefa não exige uma variedade lexical muito grande, visto que o tempo verbal se repete e que as informações são veiculadas através de palavras e frases como *My name is..., I am ... years old, I live in....* Ou seja, independente do nível em que a tarefa 1 foi produzida, as informações veiculadas são as mesmas, as quais são normalmente transmitidas através de frases como as mencionadas. Essa parece ser uma das razões para seu valor TTR ser mais baixo que os dos tipos de tarefa 2 (informações pessoais em 3ª pessoa) e 3 (informações sobre uma viagem). As tarefas 2 e 3

apresentam valor TTR de 11,42 e 10,20 respectivamente. Esses números parecem indicar a existência de diferenças linguísticas e de características particulares de cada tarefa, sendo a frequência de metáforas uma delas.

Ainda com relação aos valores da relação *type/token* de cada tipo de tarefa, na observação da frequência de metáforas nos subcorpora de níveis de proficiência (análise anterior) e da relação *type/token* (TTR) de cada nível, a diversidade lexical dos corpora pareceu ser um indicativo da frequência de ocorrências metafóricas encontrada. A verificação mostrou que o nível com maior valor TTR, foi o nível em que foi observado o maior número de metáforas. Se o valor indicativo da diversidade lexical (TTR) dos subcorpora de tipos de tarefa seguir a mesma lógica encontrada na análise anterior, espera-se que a tarefa 2 apresente maior número de ocorrências de <m>.

Apresento agora o número de ocorrências de <m> nos três subcorpora (tarefas 1, 2 e 3). Discuto os resultados numa relação contrastiva entre os tipos de tarefa produzidos durante a coleta do BELC. Pretende-se através dos contrastes de frequência apresentados nesta seção, responder a terceira questão norteadora desta pesquisa: Há variação na produção de metáforas no corpus de estudo com relação ao tipo de tarefa?

A frequência de metáforas em cada tipo de tarefa do BELC foi extraída com o *Concord* do *WordSmith Tools.* Na tabela 21, estão dispostas as frequências de metáforas encontradas em cada tarefa. Na primeira coluna, estão as tarefas 1, 2 e 3. Na segunda coluna, estão os números brutos de ocorrências de <m> em cada subcorpus e na terceira, seus valores normalizados. As comparações realizadas entre os números encontrados têm o objetivo de verificar se a presença de metáforas é mais característica de um tipo de tarefa do que de outro e se pode ser vista como uma característica mais específica de uma das tarefas. Conforme a tabela 21, na tarefa 1 são produzidas cerca de 13 metáforas por 1.000 palavras, na tarefa 2 o número apresenta um aumento de quatro ocorrências, passando para 17 itens metafóricos em 1.000 palavras. Em oposição, a comparação entre as tarefas 2 e 3 revela uma diminuição de cerca de 6 metáforas de uma tarefa para outra, caindo de 17 para 11 ocorrências de <m> por 1.000 palavras.

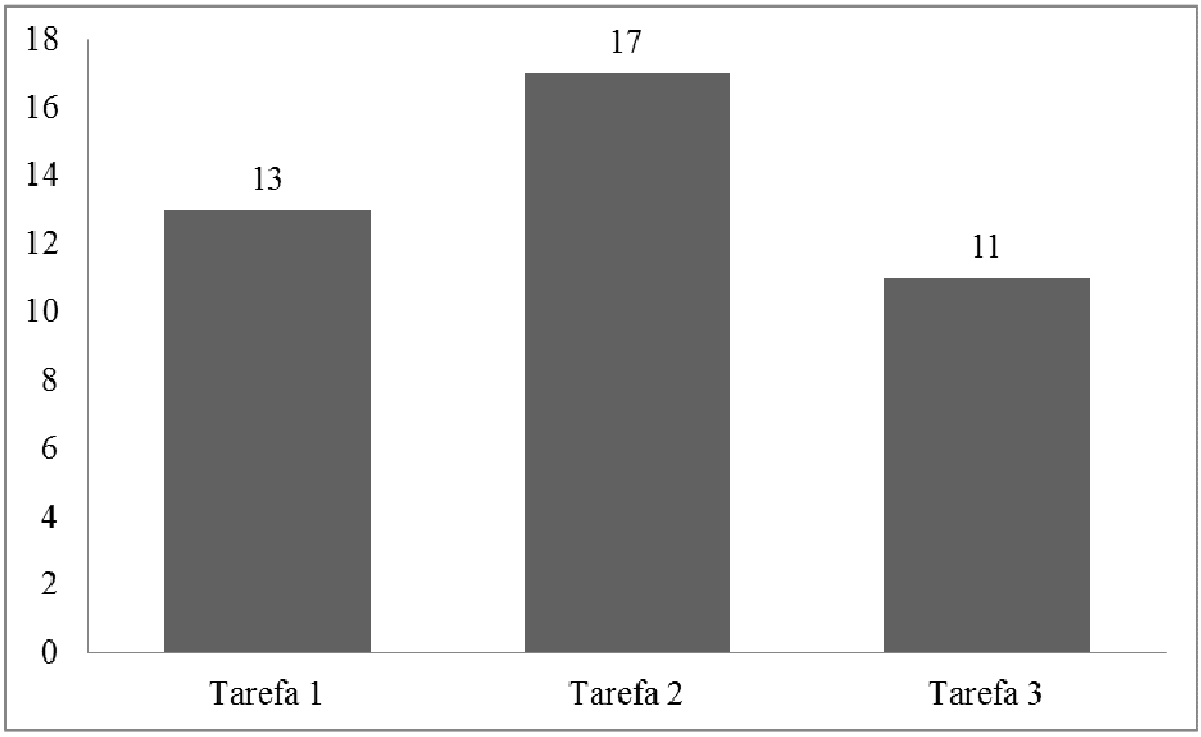
Tabela 21: Frequência de metáforas nos subcorpora de tipos de tarefa

Tipo de tarefa	**Valor bruto de <m>**	**Valor normalizado de <m> (frequência por 1.000 palavras)**
Texto 1	532	13,63
Texto 2	473	17,33
Texto 3	431	11,55

Os números encontrados revelam que as 1.436 ocorrências metafóricas no corpus não são uniformemente distribuídas entre os três tipos de tarefa. O número de ocorrências de <m>, como se pode ver, varia de forma crescente na seguinte ordem: tarefa 3 → tarefa 1 → tarefa 2. O contraste entre os resultados indica que a presença de metáforas é mais peculiar à segunda t tarefa. Portanto, quando se tratando da produção de metáforas por aprendizes de inglês como LE, falantes de PB como L1, há variação na frequência de <m> e a presença de metáforas caracteriza os três textos produzidos na coleta do corpus, mas aparece mais frequentemente em textos descritivos com informações pessoais em 3ª pessoa (tarefa 2). Os resultados encontrados parecem indicar que as diferenças são motivadas por características de cada texto.

A frequência disposta na tabela 21 pode ser melhor visualizada no gráfico 3.

Gráfico 3: Frequência de metáforas por 1.000 palavras nos tipos de tarefa



Com base no número de ocorrências metafóricas e seus valores normalizados, pode-se calcular a cada quantas palavras uma metáfora é produzida, conforme a tabela 22.

Tabela 22: Razão de produção metafórica nos tipos de tarefa: produção de uma metáfora/palavras

Tarefa	**Produção de uma metáfora/palavras**
Tarefa 1	73
Tarefa 2	57
Tarefa 3	86

Conforme a tabela 22, em ordem crescente de produção, na tarefa 3, a cada 86 palavras produzidas, uma ocorrência metafórica é encontrada. Na tarefa 1, o número de palavras diminui de 86 para 73. Nesse caso, a cada 73 palavras da tarefa, uma é metáfora. Se da tarefa 3 para a tarefa 1 houve uma diferença de 13 palavras, da tarefa 1 para a tarefa 2, a diferença é ainda maior. Na tarefa com maior produção metafórica (tarefa 2), uma metáfora é

produzida a cada 57 palavras na tarefa. Esse dado mostra que a tarefa mais lexicalmente rica (maior valor TTR) é também o que apresenta frequência metafórica mais elevada.

A tabela 23 indica a densidade de metáforas nos subcorpora. Como se pode observar, na tarefa 1, o número de metáforas corresponde a 1,36% dos *tokens* do corpus. Já na tarefa 2, das 27.280 palavras, 1,73% são itens metafóricos. Na tarefa 3, as ocorrências metafóricas representam 1,15% do subcorpus.

Tabela 23: Densidade de metáforas nos tipos de tarefa

Tarefa	**Metáforas**	**Palavras**	**Densidade**
Tarefa 1	532	39.026	1,36
Tarefa 2	473	27.280	1,73
Tarefa 3	431	37.288	1,15

As tabelas e o gráfico apresentados permitem visualizar os mesmos resultados de maneiras diferentes. Todas as formas de apresentação dos resultados, seja da frequência normalizada ou da densidade de itens metafóricos nos corpora, ilustram os mesmos dados, porém, de maneiras diferentes. O número de metáforas em cada um dos tipos tarefa analisados revela, em primeiro lugar, que as 1.436 ocorrências de <m> extraídas do BELC não são uniformemente distribuídas entre os três tipos de tarefa. As frequências variam de 11 a 17 metáforas por 1.000 palavras no corpus e essa variação parece ser motivada por necessidades comunicativas e linguísticas de cada tarefa. Os tipos de tarefa foram superficialmente analisados e identificou-se que a tarefa 2, tarefa com maior número de ocorrências metafóricas, é escrito na 3ª pessoa do presente simples. Os informantes descrevem pessoas queridas em suas vidas (pai, mãe, irmãos, amigos, namorados) e, para isso, se utilizam da metáfora, mesmo que de forma inconsciente. Além de escreverem sobre a rotina, o trabalho e os finais de semana dessa 3ª pessoa, os aprendizes apresentam suas descrições físicas e emocionais. No caso das descrições emocionais e da demonstração do que essas pessoas representam na vida dos informantes é que aparece maior frequência metafórica, conforme os exemplos do quadro 7.

Quadro 7: Exemplos de metáforas do tipo de tarefa 2

aves more like my sister. She is a very **strong <m>** woman. After several tries she go wn hair and brown eyes. She's adorable, **sweet <m>** and funny, sometimes stressed but clean my roon every necesary. She has a **hot <m>** ’s personality. When I (era) children s. He don't like some jokes and he've a **black <m>** humor hehehe. He don't like farm's

Com relação ao tipo de tarefa 1, observou-se o uso da 1ª pessoa do presente simples em todos os textos. O uso de frases como *My name is..., I live in..., I study at..., I like…, I have a (brother, sister, boyfriend)* aparece em todos os textos, desde os níveis mais iniciais até os mais avançados. No entanto, nos níveis mais avançados, apesar de as estruturas básicas se manterem as mesmas, as construções são mais complexas. Ainda com relação às narrativas pessoais em 1ª pessoa, Berber Sardinha (comunicação verbal)[44] coloca que elas tendem a apresentar menor frequência metafórica em relação a outros tipos textuais, característica essa que pode ser observada também nos textos do BELC.

No tipo de tarefa 3, os informantes relatam, em 1ª pessoa, uma viagem ou um passeio realizado. Além de se tratar de um texto em 1ª pessoa, como o texto 1, acredita-se que as razões da baixa frequência estejam relacionadas também às repetições das mesmas formas lexicais e dos mesmos tipos de construções frasais. No quadro 8 são apresentadas ocorrências de uma construção frequente no corpus.

[44] Palestra do professor Tony Berber Sardinha, no IV Congresso Internacional sobre Metáfora na Linguagem e no Pensamento, ocorrido nos dias 26, 27 e 28 de outubro de 2011, na Universidade Federal do Rio Grande do Sul.

Quadro 8: Exemplos do tipo de tarefa 3

to Floripa again. N5G1S03PPT3 **In 2001 I went to** Bombinhas with my classmates . I Cup Championship. N8G1S15PIT3 **In 2003 I went to** an Exchange Program on Canad d in next summer. N5G1S05PPT3 **In 2003 I went to** Rio de Janeiro. It was the most r work and study. N4G3S07PPT3 **In 2004 I went to** Ceará with my parents and my tw der South Africa. N7G3S18PAT3 **In 2004 I went to** Houston, Texas, in order to make there next year. N3G2S11PPT3 **In 2004 i went to** Porto Seguro with my classmates. again this year. N3G4S09PIT3 **In 2005, I went to** Taiwan, near China, there is whe IT3 When I was 14 years old, **in 2005, I went to** Disneyland with my friends. We v comeback to USA. N8G2S11PPT3 **In 2006 I went to** a trip with two of my best frien g in morning all. N2G3S08PIT3 **In 2006 I went to** África do Sul. I invited a frien them on "mangue". N7G1S03PIT3 **In 2007 I went to** Peru with my mother for 2 weeks

A análise das variações de frequência e das características das tarefas 1, 2 e 3 sugere que o nível de proficiência linguística não é fator determinante na frequência metafórica dos tipos de tarefa, visto que as três tarefas foram produzidas em todos os níveis de proficiência do BELC.

As três tarefas foram analisados seguindo os mesmos critérios metodológicos de identificação e extração de frequência e demonstraram variação no número de itens metafóricos. Com o objetivo de verificar se essas diferenças são estatisticamente significativas, aplicou-se o teste estatístico LL. Foram calculados os valores estatísticos da comparação entre as tarefas da seguinte forma:

- Tarefa 1/Tarefa 2
- Tarefa 1/Tarefa 3
- Tarefa 2/Tarefa 3

Os números obtidos encontram-se na tabela 24. Dois dos resultados estatísticos foram maiores que 6,6 e revelam que as variações apresentam alguma motivação linguística.

Tabela 24: Comparação estatística entre os tipos de tarefa

Contraste	LL
Tarefa 1/Tarefa 2	14,38
Tarefa 1/Tarefa 3	6,51
Tarefa 2/Tarefa 3	37,06

Os contrastes entre as seguintes tarefas apresentaram valor maior que 6,6, o que indica 99% de chance de as diferenças não serem aleatórias, mas estatisticamente significativas e acontecerem por alguma razão específica:

- Tarefa 1/Tarefa 2
- Tarefa 2/Tarefa 3

A comparação que apresentou diferença estatisticamente aleatória foi a seguinte:

- Tarefa 1/Tarefa 3

Do ponto de vista da linguagem como sistema probabilístico, as variações de frequência entre as tarefas 1 e 2; e 2 e 3, não ocorrem de forma aleatória. Os valores estatísticos sugerem que as variações entre essas tarefas acontecem em razão dos traços linguísticos e dos contextos situacionais de uso da linguagem em cada tarefa. Pode ser que esses valores estejam associados ao fato de que ambas os contrastes (tanto entre 1 e 2 quanto entre 2 e 3) são entre textos escritos um em 1ª e o outro em 3ª pessoa. Já na comparação entre os textos 1 e 3, a qual apresentou valor menor que 6,6 no LL, os textos são ambos escritos em 1ª pessoa, apesar de um ser descritivo e outro narrativo.

Por se tratarem de tipos de tarefa diferentes, com temáticas também diferentes, imaginava-se desde o princípio que o contraste entre as frequências de <m> nas tarefas revelasse variação de acordo com o tipo de tarefa e que a presença de metáforas caracterizasse mais uma tarefa do que outra. Entretanto, apesar de prever alguma variação, não se tinha muita ideia sobre a tarefa que apresentaria o número mais alto de ocorrências de <m>, visto que todas as três tarefas foram produzidas nos quatro níveis de proficiência. A análise mostrou que a presença de metáforas ocorre nas três tarefas, mas é mais característica de textos descritivos com informações pessoais em 3ª pessoa (tarefa 2). Entretanto, acredita-se que para compreender melhor as razões das variações encontradas entre as tarefas do BELC,

seja necessário realizar uma investigação de cunho qualitativo, com foco nas características discursivas de cada texto.

6.4 SUBCORPORA INDIVIDUAIS

Esta parte da análise tem o objetivo de apresentar e discutir as frequências de <m> encontradas em cada subcorpus individual do BELC. Um subcorpus individual corresponde a uma tarefa específica produzida em um determinado nível. Serão observadas, portanto, as frequências de metáforas de todas as tarefas em todos os níveis. O nível intermediário, por exemplo, foi transformado em três subcorpora: um subcorpus correspondente à tarefa 1, um à tarefa 2 e outro à tarefa 3. Dessa forma divididos, serão observadas as frequências em 12 subcorpora. A identificação dos mesmos foi feita através de uma letra correspondente ao nível (B, P, I, A) e um número correspondente à tarefa (1, 2, 3). O código I3, por exemplo, corresponde à tarefa 3 produzida no nível *Intermediate.*

Antes de iniciar a análise propriamente dita, faço uma breve descrição dos corpora em questão. Na tabela 25, estão elencados os doze subcorpora, seus números de *types* (formas), *tokens* (itens) e suas relações *type/token* (forma/item).

Tabela 25: Descrição dos subcorpora individuais em números

	Types	***Tokens***	***Type/token ratio***
B1	1.500	8.465	17,98
B2	1.143	5.565	20,81
B3	1.375	7.826	17,66
P1	1.865	13.708	13,80
P2	1.648	10.201	16,34
P3	1.948	13.271	14,79
I1	2.018	14.889	13,72
I2	1.666	10.310	16,33
I3	2.152	14.305	15,16
A1	621	1.964	31,96
A2	426	1.204	35,71
A3	598	1.886	32,31

O subcorpus mais rico em léxico é o da tarefa 2 produzida no nível avançado (A2). Observa-se também que em ordem decrescente de diversidade lexical, após o subcorpus A2, aparecem o A3 e o A1, sendo que neles mais de 30% do total de palavras do corpus ocorre apenas uma vez no texto. Os textos mais lexicalmente diversificados, portanto, foram todos produzidos no nível avançado do processo de aprendizagem. Isso indica que a produção de discursos lexicalmente ricos está atrelada ao nível de proficiência na LE. Além disso, a relação forma/item (TTR) é um dado que pode ser utilizado como forma de monitorar a aquisição de léxico ao longo do processo de aprendizagem.

Ainda sobre o valor da relação forma/item (TTR), nas análises de níveis de proficiência e de tipos de tarefa, as frequências mais altas de <m> foram encontradas nos subcorpora com maior valor TTR. Com base no que foi mencionado e nos números dispostos na tabela 26, imagina-se que, neste caso, o A2 seja o subcorpus individual com maior produção metafórica.

Após algumas considerações iniciais, apresento as frequências de ocorrências metafóricas nos subcorpora. A extração do número de metáforas linguísticas foi realizada com a ferramenta *Concord,* o concordanciador do *WordSmith Tools.* Os resultados encontrados estão dispostos na tabela 26 e serão discutidos numa relação contrastiva entre os subcorpora individuais do BELC. Os contrastes aqui realizados buscam verificar mais a fundo a variação da produção metafórica entre níveis e entre tipos de tarefa. Além disso, permitem identificar características específicas dos subcorpora e caracterizá-los individualmente. Na primeira coluna da tabela 26, apresento o nível e a tarefa a que cada subcorpora se refere, na segunda, o total bruto de ocorrências de <m> e na terceira, seus valores normalizados por 1.000 palavras.

Tabela 26: Frequência de metáforas nos subcorpora individuais do BELC

Subcorpus	**Valor bruto de <m>**	**Valor normalizado de <m> (frequência por 1.000 palavras)**
B1	65	7,67
B2	61	10,96
B3	64	8,17
P1	181	13,20
P2	203	19,90
P3	142	10,70
I1	245	16,45
I2	179	17,36
I3	193	13,49
A1	41	20,87
A2	30	24,91
A3	32	16,96

A tabela 26 mostra que a frequência da produção metafórica varia de 7 a 24 metáforas por 1.000 palavras. Esses números indicam que existe variação entre os subcorpora individuais e que, dentre todos, no nível avançado, tarefa 2 (subcorpus A2), encontra-se o maior número de ocorrências metafóricas. Isso indica que a suposição realizada com base na comparação entre a relação *type/token* e o número de itens metafóricos foi corroborada. Da mesma forma que na análise de níveis e na análise de tipos de tarefa, o subcorpus com maior valor TTR apresentou a frequência mais alta de metáforas linguísticas, nesta análise o subcorpus individual com maior valor TTR é o subcorpus com maior frequência metafórica (A2). Entretanto, apesar dessa relação ter sido observada nesta análise e nas anteriores, a suposição não pode ser generalizada, visto que os números da relação forma/item (TTR) e da frequência metafórica só mostraram alguma relação nos primeiros colocados das listas de frequência de metáforas nos corpora. Ou seja, os resultados sugerem a existência de alguma relação entre esses números, a qual não pode ser generalizada, visto que foi observada apenas nos primeiros colocados das listas de frequências metafóricas de cada análise (níveis de proficiência, tipos de tarefa e subcorpora individuais).

Com base nos números de ocorrências de <m> e seus valores normalizados, pode-se calcular a razão da produção metafórica nos subcorpora individuais (produção de uma metáfora/palavras), conforme a tabela 27, que também ilustra a variação de frequência entre os subcorpora.

Tabela 27: Razão de produção metafórica nos subcorpora individuais: produção de uma metáfora/palavras

Subcorpus	Produção de uma metáfora/palavras
B1	130
B2	91
B3	122
P1	75
P2	50
P3	93
I1	60
I2	57
I3	74
A1	47
A2	40
A3	58

As tabelas dispostas nesta parte da análise ilustram de maneiras diferentes os resultados obtidos na extração de frequências dos subcorpora. Todas as formas de apresentação dos números apontam para a variação na produção metafórica entre os subcorpora individuais. No entanto, não fica claro se a variação está mais relacionada ao nível de proficiência ou ao tipo de tarefa.

Para que os subcorpora e seus números de ocorrências de <m> possam ser analisados mais a fundo, a fim de verificar a influência do nível de proficiência e do tipo de tarefa na variação de frequência metafórica, eles serão observados em quatro grupos, de acordo com os níveis de proficiência do BELC. Cada um dos grupos é formado por três subcorpora, os quais correspondem aos três tipos de tarefa produzidos em um mesmo nível. As três tarefas, portanto, serão observados em todos os níveis, o que facilita a análise dos números. Neste momento, a ordem crescente/decrescente de frequência de metáforas não será considerada.

O primeiro dos grupos corresponde aos três tipos de tarefa produzidos no nível inicial do processo de aprendizagem (B1, B2 e B3). A tarefa 1, produzida no nível B, apresenta 7 metáforas por 1.000 palavras, o que corresponde a uma metáfora a cada 130 palavras produzidas no corpus. Na tarefa 2 (subcorpus B2), em que os aprendizes escreveram informações pessoais em 3ª pessoa, o número de metáforas aumenta em relação à tarefa 1, passando para 10 metáforas por 1.000 palavras. Esse número corresponde à produção de um item metafórico a cada 91 palavras no texto. Já na terceira tarefa (subcorpus B3), há uma frequência de 8 metáforas por 1.000 palavras no texto. Isso significa a mesma coisa que dizer

que uma metáfora linguística ocorre a cada 122 *tokens* no corpus. Esses números estão dispostos na tabela 28.

Tabela 28: Frequência de metáforas nos subcorpora do nível *Beginner*

Subcorpus	**Frequência de <m> por 1.000 palavras**	**Produção de uma metáfora/palavras**
B1	7	130
B2	10	91
B3	8	122

Observa-se na tabela 28 que o número de metáforas no nível *Beginner* apresentou pouca variação entre as três tarefas, mantendo-se quase que constante. Esse número corrobora os resultados encontrados na análise dos subcorpora de níveis de proficiência, mostrando que a frequência metafórica nas produções escritas de aprendizes está atrelada ao estágio do processo de aprendizagem em que os mesmos se encontram. No caso dos subcorpora B1, B2 e B3, o número de ocorrências de <m> parece estar mais ligado ao nível de proficiência na LE do que ao tipo de tarefa, visto que a frequência se mantém quase que constante nos três tipos de tarefa produzidos no mesmo nível.

No segundo grupo, foram observadas as variações entre a frequência de itens metafóricos dos subcorpora correspondentes aos três tipos de tarefa produzidos no nível pré-intermediário de proficiência na LE, os subcorpora P1, P2 e P3. No primeiro tipo de tarefa produzido no nível pré-intermediário, a frequência encontrada foi de 13 metáforas por 1.000 palavras, o que corresponde a uma ocorrência metafórica a cada 75 palavras produzidas no texto. Da tarefa 1 para a tarefa 2, o número de metáforas aumenta para 19 metáforas por 1.000 palavras ou um item metafórico a cada 50 palavras no corpus. A menor frequência foi encontrada no tipo de tarefa 3, em que foram produzidas 10 metáforas por 1.000 palavras no corpus, o que equivale à produção de uma metáfora a cada 93 palavras, conforme a tabela 29.

Tabela 29: Frequência de metáforas nos subcorpora do nível *Pre-Intermediate*

Subcorpus	**Frequência de <m> por 1.000 palavras**	**Produção de uma metáfora/palavras**
P1	13	75
P2	19	50
P3	10	93

Nesse caso, a variação da frequência de itens metafóricos nos três subcorpora se manteve menos constante do que na análise anterior. As frequências dos subcorpora P1 e P3 apresentaram pouca variação, entretanto, no contraste desses subcorpora com o subcorpus P2 a variação foi maior. Diferentemente dos subcorpora analisados anteriormente (B1, B2, B3), em que o nível de proficiência pareceu exercer influência direta na frequência de itens metafóricos, os contrastes aqui realizados aponta para maior influência do tipo de tarefa na frequência. As frequências parecem continuar atreladas ao nível, visto que as frequências entre P1 e P3 não variaram tanto. No entanto, o fato de o número de ocorrências nos subcorpora ter se mantido menos constante, parece ter sido motivado pelas características linguísticas da tarefa 2. É relevante lembrar que na análise de tipos de tarefa, o subcorpus corresponde à tarefa 2 apresentou maior frequência de <m> que as tarefas 1 e 3.

O terceiro grupo corresponde aos três tipos de tarefa produzidos no nível intermediário de proficiência na LE (subcorpora I1, I2 e I3). De acordo com a tabela 30, na primeira tarefa do nível intermediário, foram produzidas 16 metáforas por 1.000 palavras, o que equivale à produção de um item metafórico a cada 60 palavras no corpus. Na comparação entre as tarefas 1 e 2 produzidas nesse nível, o número se mantém quase que constante, visto que são produzidas 17 metáforas por 1.000 palavras, o que corresponde à ocorrência de uma metáfora a cada 57 palavras no corpus. A variação aumenta no contraste das tarefas 1 e 2 com a tarefa 3. Se no subcorpus I1 são produzidas 16 e no subcorpus I2, 17 metáforas por 1.000 palavras no corpus, no subcorpus I3 o número cai para 13 ocorrências metafóricas. Esse número equivale à ocorrência de uma metáfora a cada 74 palavras no corpus.

Tabela 30: Frequência de metáforas nos subcorpora do nível *Intermediate*

Subcorpus	**Frequência de <m> por 1.000 palavras**	**Produção de uma metáfora/palavras**
I1	16	60
I2	17	57
I3	13	74

A interpretação desses números aponta novamente para a influência tanto do nível de proficiência quanto do tipo de tarefa. A baixa variação entre as frequências do I1 e do I2 parece ser consequência do nível de proficiência na língua. Entretanto, no caso do subcorpus I3, a variação aponta para razões concernentes ao tipo de tarefa, visto que na análise dos subcorpora de tipos de tarefa, o subcorpus da tarefa 3 apresentou a menor frequência.

Com relação ao nível avançado de proficiência linguística, no subcorpus A1 observou-se a ocorrência de 20 metáforas por 1.000 palavras, o que corresponde a um item metafórico produzido a cada 47 palavras no corpus. No subcorpus A2 a frequência aumenta para 24 metáforas por 1.000 palavras. Nesse caso, uma metáfora é produzida a cada 40 palavras no corpus. No subcorpus A3 a frequência diminui em relação ao A1 e ao A2. O número de ocorrências cai para 16 <m> por 1.000 palavras, o que equivale à produção de uma metáfora a cada 58 palavras no corpus. Esses números estão dispostos na tabela 31.

Tabela 31: Frequência de metáforas nos subcorpora do nível Advanced

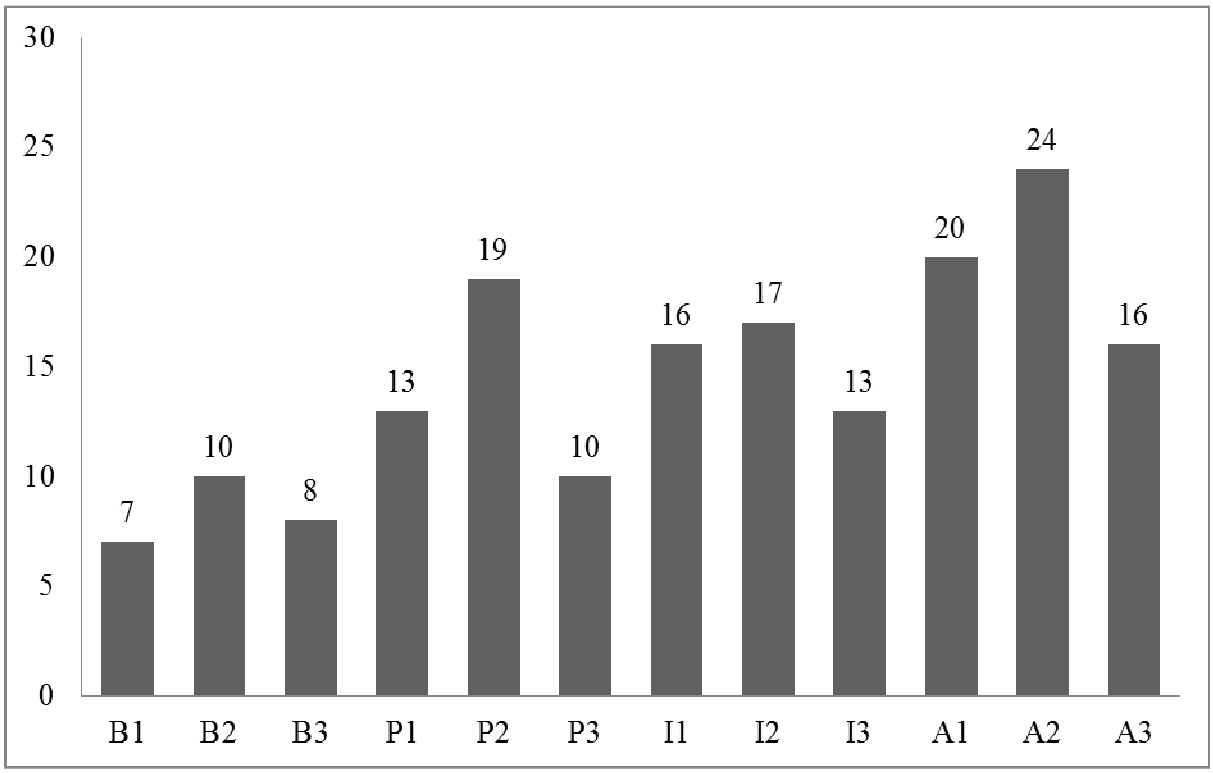
Subcorpus	**Frequência de <m> por 1.000 palavras**	**Produção de uma metáfora/palavras**
A1	20	47
A2	24	40
A3	16	58

Os números apontam novamente para a influência tanto do nível de proficiência quanto do tipo de tarefa na frequência de itens metafóricos e corroboram os resultados encontrados nas análises de níveis de proficiência e de tipos de tarefa. Ou seja, no nível avançado há maior frequência de metáforas em relação aos níveis anteriores. Com relação ao tipo de tarefa, os números aqui obtidos crescem na mesma ordem que os números encontrados

na análise de tipos de texto. A frequência aumenta na seguinte ordem tarefa 3 → tarefa 1 → tarefa 2.

No gráfico 4, estão dispostas as frequências de todos os doze subcorpora. O gráfico está organizado em ordem crescente de proficiência.

Gráfico 4: Frequência de <m> por 1.000 palavras nos subcorpora individuais



Percebe-se que nos três subcorpora do nível *Beginner* o número de metáforas produzidas foi inferior em relação aos outros níveis. Entre os textos 1, 2 e 3 desse nível, observa-se que há pouca variação. Já nos subcorpora do nível pré-intermediário (P1, P2 e P3), a frequência foi mais elevada em relação ao nível anterior, porém menos constante. No nível intermediário, a frequência é ainda mais elevada. A observação das frequências nos subcorpora I1, I2 e I3 mostra que no nível intermediário a produção metafórica nos textos é mais constante que no nível pré-intermediário. Com relação ao nível avançado, a frequência aumenta ainda mais e o número de ocorrências nos subcorpora (A1, A2 e A3) é mais variável do que nos subcorpora correspondentes aos níveis *Beginner* e *Intermediate.*

De forma geral, observa-se a influência tanto do nível de proficiência quanto do tipo de tarefa na variação de frequência de <m> entre os subcorpora individuais. Com relação aos níveis de proficiência, observa-se que os subcorpora correspondentes a um mesmo nível (A1, A2 e A3, por exemplo) apresentam variação de frequência, porém as variações se mantêm dentro de certos limites, os quais são particulares de cada nível. Quanto à influência do tipo de tarefa, em todos os quatro grupos acima analisados, a tarefa 2 apresentou frequência mais alta, o que além de corroborar os resultados da análise de subcorpora de tipos de tarefa, mostra que mesmo em uma análise mais profunda, o tipo de tarefa continua a exercer influência no número de ocorrências metafóricas.

6.5 ALGUMAS CONSIDERAÇÕES QUALITATIVAS

A metodologia de busca de metáforas no BELC não partiu de uma lista preestabelecida de expressões metafóricas. A anotação feita foi de cunho *bottom-up*, a fim de identificar todas as metáforas (que se encaixavam entre os limites estabelecidos nesta investigação) presentes no corpus. Como mencionado no capítulo que relata a metodologia desta pesquisa, a unidade de análise foi a metáfora linguística. O objetivo não foi identificar metáforas conceptuais subjacentes, por exemplo. Entretanto, as teorias abordadas no capítulo sobre estudos da metáfora fizeram com que os dados do corpus fossem lidos e analisados mais criteriosamente, durante a anotação. Assim sendo, durante a leitura do corpus, foi natural "enxergar" particularidades das ocorrências marcadas e observar as teorias que as sustentavam.

Não pretendo realizar nesta seção, uma análise qualitativa exaustiva das ocorrências metafóricas do corpus. O objetivo é esboçar algumas considerações de cunho qualitativo sobre pontos observados durante a anotação do corpus. Além disso, pretende-se realizar breves comentários sobre ocorrências consideradas relevantes e peculiares ao tipo de língua analisado.

6.5.1 Metáforas ontológicas: o verbo *‘to have’*

Metáforas conceptuais, em geral, são consideradas um fenômeno linguístico que vai além do nível das palavras. Segundo Lakoff e Johnson (1980), uma metáfora não é só uma palavra semântica ou pragmaticamente incongruente com o discurso ao seu redor, mas é um sentido codificado no nível do pensamento. Consideradas um processo cognitivo, metáforas conceptuais vão além da linguagem poética e perpassam o entendimento dos mais variados conceitos, como os relacionados aos sentimentos, por exemplo. Segundo Berber Sardinha (2007d), a metáfora oferece meios e se faz necessária para dar conta e sentido ao que passamos e sentimos durante a vida, como alegrias e tristezas, por exemplo.

Em *Recontando a vida em narrativas pessoais: um estudo de metáforas na perspectiva da Linguística de Corpus,* Berber Sardinha (2007d) realiza uma investigação sobre a metáfora em narrativas pessoais. O autor analisa uma coletânea de 32 narrativas em que pessoas com mais de 60 anos contam suas vidas. A metodologia utilizada foi de cunho *bottom-up,* com o objetivo de realizar um levantamento de todas as metáforas presentes no corpus, através do programa identificador de metáforas. A palavra com maior probabilidade metafórica indicada pelo programa foi *tenho.* Dentre as 499 ocorrências de *tenho* no corpus, 149 eram metafóricas, o que equivale a 29,9% de uso metafórico. As 149 ocorrências eram casos de metáforas ontológicas em que algo abstrato estava sendo conceptualizado como um objeto concreto, como “tenho contato”, “não tenho dúvida” e “tenho interesse”. Dentre as conclusões, o pesquisador relata que a metáfora ontológica é um recurso importante para falar de conceitos da vida e relatar experiências vividas ao longo dos anos.

Assim como as narrativas analisadas por Berber Sardinha *(Ibidem),* os três tipos de tarefa que compõem o BELC giram em torno do mesmo eixo: a vida dos informantes (tarefa 1: texto descritivo com informações pessoais em 1ª pessoa; tarefa 2: texto descritivo com informações pessoais em 3ª pessoa, tarefa 3: texto narrativo sobre um passeio/viagem). Dessa forma, o BELC também apresenta metáforas ontológicas com o verbo *to have.* As linhas de concordância dispostas no quadro 9 foram retiradas do BELC e ilustram a presença dessas metáforas no corpus.

Quadro 9: Exemplos de metáforas ontológicas no BELC

lunteer work, knows English and **has <m>** big goals and strong skills of leadership. A
le. He like velocity and my mother **have <m>** fear about this. We like a handgum but
Alemanha, Japão and others. I **have <m>** "saudades" and a have "vontade" of go back
y classes are almost ending. I **have <m>** no idea why I've chosen Journalism, because

Em todas as quatro linhas de concordância, conceitos abstratos *(goals, fear,* saudades, *idea)* são conceptualizados em termos de objetos concretos. Ou seja, assim como na L1, como mostrado por Berber Sardinha (2007d), os aprendizes do BELC falam de sentimentos e experiências como se fossem objetos concretos, os quais podemos 'ter', 'ganhar' ou 'perder', por exemplo. Com relação à frequência destas metáforas no corpus, há 333 ocorrências metafóricas de formas do verbo *to have,* conforme a tabela 32.

Tabela 32: Frequência de metáforas ontológicas com *to have* no BELC

Forma metafórica do verbo *to have*	**Número de ocorrências**
Have	208
Has	33
Had	80
Having	10
Haved	1
Havin'	1
TOTAL	333

Dentre as 333 ocorrências, 208 são da forma *have,* 33 da forma *has,* 80 da forma *had* e 10 da forma *having.* Há também duas exceções: *haved* e *havin'.* A ocorrência da forma *haved* sugere que houve a generalização do passado simples dos verbos regulares. Já em *havin'* ocorreu a supressão do *g* final do presente contínuo. Essas ocorrências representam 23,18% do número total de metáforas no BELC.

Os resultados encontrados deixam claro o quanto a metáfora conceptual, a metáfora ontológica em especial, é um fenômeno de pensamento embutido na maneira de entender, compreender e descrever o mundo. O recurso é inerente à forma de pensar do ser humano, o qual não é dissociado do sistema conceptual dos indivíduos nem durante a aprendizagem de uma LE. Outro fator que aponta para essa ideia é o fato de que as 333 ocorrências encontradas

estão distribuídas entre os níveis de proficiência. Ou seja, são produzidas desde os estágios iniciais do processo de aprendizagem. A presença de metáforas ontológicas, portanto, não é peculiar ao tipo de língua aqui analisado, mas parece ser significativa o suficiente para caracterizá-lo.

6.5.2 *Fight x Argue*

Outro caso curioso observado durante a anotação do BELC foi o uso da unidade lexical *fight.* Literalmente, a palavra significa lutar fisicamente. Nas 15 ocorrências de *fight* anotadas com o código <m>, observou-se que foram utilizadas no sentido de agressão verbal *(argument),* não física, conforme o quadro 10.

Quadro 10: Exemplos de metáforas linguísticas com *fight* no BELC

ble but it's good, because we hard ever **fight <m>** . He's my best friend. He likes rock
e in a lot of things that's why we **fight <m>** as frequently as we have funny times tog
good relationship, but sometimes we **fight <m>** . I love my parents, they are everythi

As ocorrências foram identificadas como metafóricas, pois de acordo com os critérios do MIP (PRAGGLEJAZ, 2007), os significados básico e contextual da unidade lexical *fight* são diferentes, mas o significado contextual pode ser entendido na comparação com o significado básico da palavra, conforme o esquema abaixo.

- Significado contextual: Nas ocorrências de *fight* anotadas, a palavra aparece com o significado de agressão verbal, discussão.
- Significado básico: O significado mais básico de *fight* é agressão física.
- Significado contextual x significado básico: O significado contextual é diferente do significado básico, mas pode ser entendido na comparação com ele.
- Unidade lexical utilizada metaforicamente? Sim.

A metáfora conceptual subjacente às expressões linguísticas mencionadas é DISCUSSÃO É GUERRA. Nesse caso, as frases *Ele brigou com ela* e *Ele discutiu com ela* são consideradas equivalentes.

O uso das formas metafóricas de *fight* parece ter sido motivado por lacunas no vocabulário dos aprendizes, problema comum na aprendizagem de uma L2. Dessa forma, no esforço de tentar superar a dificuldade comunicativa encontrada, o aprendiz busca estratégias que o auxiliem a veicular o significado desejado. A utilização da metáfora como recurso para suprir necessidades comunicativas mostra que, mesmo que de forma inconsciente, o fenômeno faz parte do discurso de aprendizes no processo de aprendizagem da LE. As ocorrências observadas sugerem que a falta do vocabulário específico tenha sido suprida através de duas formas principais: (i) transferência da L1 para a LE; e (ii) semelhanças semânticas entre as palavras *fight* e *argument.* No caso da transferência de uma língua para outra, quando o conhecimento da LE não é suficiente para elaborar seu discurso, o aprendiz busca apoio na sua L1. Ou seja, os conhecimentos e habilidades da L1 são utilizados na resolução de problemas comunicativos encontrados na LE. As ocorrências metafóricas de *fight* no corpus apontam também para a influência exercida pelo português, já que no PB as palavras lutar *(fight)* e discutir *(argue)* são muitas vezes utilizadas indistintamente. A frase *Tive uma briga séria com meu namorado*, por exemplo, é quase sempre usada com o sentido de agressão verbal e não de agressão física. Outro fator que parece influenciar nos casos observados é a semelhança semântica entre os substantivos *fight* e *argument* ou entre os verbos *to fight* e *to argue*, visto que ambos exprimem o sentido de agressão, sendo uma física e outra verbal.

6.5.3 *Water down x Waterfall*

Durante a anotação de metáforas no corpus de nível pré-intermediário, tarefa 3, identificou-se um caso curioso, porém não metafórico. Ao descrever uma viagem que havia feito a Gramado e Canela, o informante relata que visitou o Parque do Caracol, um lugar bonito com uma cascata (cachoeira, queda d'água) fantástica. A ocorrência encontra-se no quadro 11.

Quadro 11: Uso da expressão *water down* no BELC

In Canela we visited the Caracol Park's, a beautiful place, with an fantastic **'water down'**

O fato de o aprendiz não saber como dizer a palavra cascata em inglês, fez com que ele "criasse" uma expressão para veicular seu significado. A palavra cascata corresponde à palavra *waterfall* em inglês. Na falta desse vocabulário, o aprendiz criou meios de veicular o significado desejado, utilizando a expressão *water down.* A tradução literal dessa expressão é a junção das palavras água e para baixo (direção). A ocorrência não foi anotada com o código <m>, ou seja, não foi considerada metafórica segundo os critérios estabelecidos pelo MIP na identificação de metáforas linguísticas, visto que contextualmente o significado das palavras *water down* e *waterfall* é o mesmo.

A estratégia utilizada pelo aprendiz foi bem sucedida e, apesar de *water down* não veicular o significado exato de cascata, remete à imagem de água caindo. Conforme o quadro 11, observa-se também que a expressão utilizada é apresentada entre aspas no texto. Talvez essas aspas sejam um indicativo de que o aprendiz tinha conhecimento de que estava utilizando uma expressão não considerada padrão na língua.

Este caso aponta para uma lacuna no vocabulário do aprendiz, comum no processo de aprendizagem de uma língua que não a materna e mostra um dos recursos utilizados durante o processo de aprendizagem para suprir dificuldades comunicativas. Como forma de conseguir veicular o significado desejado, o aprendiz mostra que existem outros recursos que, apesar de não considerados padrão na língua, satisfazem suas necessidades, evitando problemas de comunicação.

7 CONSIDERAÇÕES FINAIS

Nesta seção, serão retomadas as questões norteadoras desta investigação e, com o objetivo de respondê-las, serão apresentadas as etapas metodológicas e os resultados obtidos. Posteriormente, serão sugeridos pontos para pesquisas futuras.

O objetivo geral desta pesquisa foi investigar quantitativamente o processo de produção metafórica entre falantes de uma LE em diferentes níveis de proficiência e tipos de tarefa, no BELC (PACHECO, 2010), através de uma abordagem baseada em corpus. A fim de alcançar esse objetivo, foram estabelecidas as seguintes questões de pesquisa:

1. Os aprendizes de inglês como LE, falantes de PB como L1, como evidenciado pelo BELC, produzem metáforas?
2. Há variação na frequência da produção metafórica no corpus de estudo com relação ao nível de proficiência linguística em LE?
3. Há variação na produção de metáforas no corpus de estudo com relação ao tipo de tarefa?

As hipóteses que nortearam esta investigação são:

1. Aprendizes brasileiros de inglês como LE, falantes de PB como L1, produzem metáforas.

2. Há variação na produção metafórica com relação aos níveis de proficiência linguística, sendo que quanto mais avançado o nível, maior o número de ocorrências metafóricas.

3. Há variação na produção metafórica com relação ao tipo de tarefa, sendo que probabilidades de uso da linguagem metafórica variam de acordo com tipos textuais específicos.

Com o intuito de responder as questões de pesquisa e verificar as hipóteses acima, as seguintes etapas metodológicas foram seguidas:

a) Identificação e anotação manual de metáforas linguísticas no BELC, com base nos procedimentos de Cameron (2003) e do Grupo Pragglejaz (2007)

Durante a escolha do método desta pesquisa, surgiram alguns problemas (relatados no capítulo 5). A resolução desses percalços culminou com a opção pela leitura e anotação manual do corpus na busca por metáforas linguísticas. A identificação de metáforas foi realizada com base nos procedimentos de Cameron (2003) e do Grupo Pragglejaz (2007). Sendo o BELC um corpus composto por pouco mais de 100.000 palavras, o equivalente a 170 páginas do *Word* corridas, a anotação foi um processo demorado e minucioso. A anotação do corpus mostrou que o fato de o MIP (PRAGGLEJAZ, 2007) não ser um método específico para a identificação de metáforas em um corpus de aprendiz impõe algumas limitações na anotação, visto que há ocorrências peculiares à linguagem do aprendiz não previstas na metodologia. Como mencionado no decorrer do trabalho, um dos aspectos dessa natureza evidenciado pelo BELC foi a dificuldade no uso de preposições, as quais parecem ser utilizadas como unidades desprovidas de conteúdo semântico. Também foram identificados desvios da língua padrão e transferências da L1 para a LE. Diante dos pontos mencionados e da inexistência de um método específico para a identificação de metáforas em corpora de aprendizes, foi necessário estabelecer critérios que auxiliassem a lidar com as peculiaridades da língua de aprendizes brasileiros de inglês como LE, falantes de PB como L1.

b) Extração da frequência de metáforas no BELC e em seus subcorpora (níveis de proficiência, tipos de tarefa e individuais), através do concordanciador do *WordSmith Tools.*

Subsequente à anotação do corpus, em que as ocorrências foram anotadas com a etiqueta <m>, foram extraídas as frequências de metáforas do BELC como um todo e de seus subcorpora de níveis de proficiência, de tipos de tarefa e individuais.

c) Contraste das frequências de metáforas linguísticas entre os subcorpora de níveis de proficiência.

Nesta etapa, foram contrastadas as frequências extraídas dos subcorpora dos níveis de proficiência *Beginner, Pre-Intermediate, Intermediate* e *Advanced.* A análise dos contrastes entre os quatro níveis mostrou variação na frequência de metáforas produzidas. A variação se dá de forma crescente. Quanto mais alto o nível, maior o número de ocorrências de <m>. Entretanto, considerando a forma como os informantes do BELC foram classificados de

acordo com suas proficiências linguísticas na LE, esperava-se que a elevação do número de metáforas ocorresse de forma uniforme. Na tabela 33, já apresentada na seção 6.2, observa-se que esse aumento não se dá de forma uniforme e proporcional entre um nível e outro.

Tabela 33: Frequência de metáforas nos níveis de proficiência

Nível de proficiência	**Valor bruto de <m>**	**Valor normalizado de <m> (frequência por 1.000 palavras)**
Beginner	190	8,69
Pre-Intermediate	526	14,14
Intermediate	617	15,61
Advanced	103	20,38

Acredita-se que a baixa variação encontrada entre os níveis pré-intermediário e intermediário seja reflexo de fatores relacionados ao modo de classificação dos alunos segundo suas proficiências linguísticas, visto que o teste utilizado era originalmente dividido em três níveis de proficiência, os quais foram transformados em quatro na compilação do BELC.

d) Contraste das frequências de metáforas linguísticas entre os subcorpora de tipos de tarefa.

O contraste entre as frequências extraídas dos três subcorpora de tipos de tarefa mostrou que as ocorrências metafóricas do BELC não são uniformemente distribuídas entre os três tipos de tarefa que compõem o corpus. Presumia-se que isso aconteceria, visto que são tarefas diferentes, com temáticas também diferentes. Apesar de prever variação, não imaginava-se qual das três tarefas apresentaria frequência metafórica mais alta, já que todos foram produzidas nos quatro níveis de proficiência. Observou-se que a variação de ocorrências de <m> se dá de forma crescente na seguinte ordem: tarefa 3 (informações sobre uma viagem) → tarefa 1 (informações pessoais em 1ª pessoa) → tarefa 2 (informações pessoais em 3ª pessoa), conforme a tabela 34, reproduzida da seção 6.3.

Tabela 34: Frequência de metáforas nos subcorpora de tipos de tarefa

Tipo de tarefa	Valor bruto de <m>	Valor normalizado de <m> (frequência por 1.000 palavras)
Tarefa 1	532	13,63
Tarefa 2	473	17,33
Tarefa 3	431	11,55

A extração e análise das frequências mostraram que a presença de metáforas caracteriza os três tipos de tarefa, mas é mais peculiar à tarefa 2 (texto descritivo em 3ª pessoa). Acredita-se que essas variações sejam motivadas em razão dos traços linguísticos e dos contextos situacionais de uso da linguagem em cada tarefa. Entretanto, entendo que para compreender melhor as razões das variações encontradas entre os tipos de tarefa do BELC, seja necessário realizar uma investigação de cunho qualitativo, com foco nas características discursivas de cada tarefa.

e) Contraste das frequências de metáforas linguísticas entre os subcorpora individuais.

Nesta etapa, foram verificados os números de ocorrências de <m> nos subcorpora individuais do BELC. O contraste das frequências mostrou que a produção metafórica varia de 7 a 24 metáforas por 1.000 palavras, sendo que dentre todos os subcorpora, o subcorpus A2 (tarefa 2 produzida no nível avançado) apresentou o maior número de ocorrências. As variações parecem ser influenciadas tanto pelos níveis de proficiência dos aprendizes quanto pelos tipos de tarefa. Com relação aos níveis de proficiência, observou-se que os subcorpora correspondentes a um mesmo nível (A1, A2 e A3, por exemplo) apresentaram variação de frequência, porém essas variações se mantiveram dentro de certos parâmetros, os quais são particulares de cada nível. Quanto à influência do tipo de tarefa, em todos os subcorpora correspondentes a um mesmo nível (B1, B2 e B3, por exemplo), a tarefa 2 apresentou frequência de <m> mais alta. Essa constatação corroborou os resultados da análise de subcorpora de tipos de tarefa e mostrou que mesmo em uma análise mais profunda (como a realizada na seção 6.4), a tarefa continua a exercer influência na variação da produção de metáforas na escrita dos aprendizes.

f) Discussão de algumas ocorrências consideradas peculiares ao tipo de língua sob investigação.

A busca de metáforas no BELC não partiu de uma lista preestabelecida. A anotação feita foi de cunho *bottom-up* e teve como objetivo principal identificar com o código <m> todas as metáforas linguísticas encontradas no corpus que estivessem dentro dos limites estabelecidos na seção 5.6. Dessa forma, durante a leitura do corpus, foram também identificadas ocorrências peculiares à linguagem dos aprendizes durante o processo de aprendizagem da LE. Optou-se então por realizar alguns comentários de cunho qualitativo sobre aspectos e ocorrências consideradas particulares ao tipo de língua aqui investigado. Entre os pontos observados estão: (i) o alto número de ocorrências de metáforas ontológicas composta por formas do verbo *to have*, dando origem a expressões como *have fear* (ter medo); (ii) o uso de formas lexicais semanticamente similares, na tentativa de tentar suprir lacunas no vocabulário, como o uso de *fight* no lugar de *argue*, em frases como *We have a good relationship, but sometimes we fight;* (iii) estratégias utilizadas com a intenção de suprir lacunas de vocabulário e necessidades comunicativas na LE, como no uso da expressão *water down* no lugar de *waterfall* (cachoeira).

Além das constatações acima, observou-se a existência de uma relação entre o valor TTR dos corpora e seus números de ocorrências de <m>. Nos três contrastes realizados (entre subcorpora de níveis de proficiência, de tipos de tarefa e individuais do BELC), observou-se que o subcorpus com valor TTR mais alto foi sempre o mesmo com o maior número de ocorrência metafóricas. Isso parece indicar uma relação entre os dois valores, a qual não pode ser generalizada, visto que foi observada apenas nos primeiros colocados das listas de frequência. Nos níveis de proficiência, por exemplo, o valor TTR decresce na seguinte ordem: *Advanced → Beginner → Intermediate → Pre-Intermediate.* Em relação ao número de ocorrências metafóricas, em ordem decrescente de frequência, os níveis são: *Advanced → Intermediate → Pre-Intermediate → Beginner*. Ou seja, o único nível que ocupa a mesma posição em ambas as listas, tanto do valor TTR quanto do número de ocorrências de <m>, é o *Advanced.* Observou-se o mesmo nos subcorpora de tipos de tarefa.

Em resumo, as hipóteses foram corroboradas. Concluiu-se que há variação na frequência da produção metafórica por aprendizes de inglês como LE, falantes de PB como

L1, entre níveis de proficiência e tipos de tarefa, mostrando também que o discurso do aprendiz, assim como outros tipos de língua, é permeado pela presença de metáforas.

Diante da carência de estudos que descrevam a linguagem de aprendizes de uma LE no tocante à produção de metáforas linguísticas, acredito que esta pesquisa contribui para preencher a lacuna existente tanto no campo da LdC e da aquisição de línguas, quanto nos estudos da metáfora. Com relação à LdC, a mesma se mostrou ser uma ferramenta extremamente útil no estudo da variação de frequência, visto que apresenta o número de ocorrências anotadas no corpus com rapidez e exatidão. Destaco também a importância do uso de corpora de aprendizes em pesquisas linguísticas, devido às novas percepções que eles podem auxiliar a revelar na descrição de aspectos do processo de aprendizagem de uma LE. Em oposição, considero a metodologia de busca e anotação de metáforas uma das limitações desta pesquisa. Como relatado no capítulo 5, inicialmente, a ideia era utilizar um método que evitasse a leitura e anotação manual do corpus, em função de diversos fatores, dentre os quais destacam-se a subjetividade do processo e o trabalho manual envolvido na análise. Diante da impossibilidade de uso de outros métodos, o método utilizado foi a leitura e anotação manual do corpus, com base em procedimentos de identificação de metáforas linguísticas considerados criteriosos, a fim de conferir maior confiabilidade à anotação. Posteriormente, realizou-se a validação da anotação, também como forma de garantir um processo de identificação de metáforas linguísticas mais confiável.

Por fim, com base em ocorrências metafóricas encontradas durante a anotação (como as mencionadas no item f) desta seção), sugiro pesquisas futuras que se proponham a investigar a influência da língua materna na produção de metáforas na língua alvo e que auxiliem a desvendar questões concernentes tanto à produção metafórica em LE quanto à influência exercida pela L1 no processo de aquisição de LEs.

REFERÊNCIAS

ARISTÓTELES. (séc IV a.C.) **A poética clássica.** 7. ed. São Paulo: Cultrix, 1997.

BERBER SARDINHA, Tony. MCI, um identificador de candidatos à metáfora em corpora. In: SHEPHERD, Tania et al. (Org.) **Caminhos da Linguística de Corpus.** São Paulo: Mercado de Letras, 2012. p. 87-105.

_____. Metaphor and Corpus Linguistics. **Revista Brasileira de Linguística Aplicada.** v. 11, n.2, p. 329-360, 2011a.

_____. Metáforas e Linguística de Corpus: Metodologia de Análise Aplicada a um Gênero de Negócios. **D.E.L.T.A.**, São Paulo, v. 27, n.1, p.1-20, 2011b.

_____. Como usar a Linguística de Corpus no Ensino de Língua Estrangeira – por uma Linguística de Corpus Educacional brasileira. In: VIANA, Vander et al. (Org.). **Corpora no Ensino de Línguas Estrangeiras.** São Paulo: HUB Editorial, 2010. p. 293-348.

_____. Metáforas de teleconferências de negócios. **Caderno Est. Ling.**, Campinas, v. 50, n. 2, p. 171-188, jul./dez 2008.

_____. Análise de metáfora em corpora. **Ilha do Desterro**: A Journal of English Language, Literatures in English and Cultural Studies, Florianópolis, n. 52, p. 67-199, jan./jun, 2007a.

_____. **Metáfora.** São Paulo: Parábola Editorial, 2007b.

_____. Metaphor in corpora: a corpus-driven analysis of Applied Linguistics dissertations. **Revista Brasileira de Linguística Aplicada**, v.7, n. 1, p. 11-35. 2007c.

_____. Recontando a vida em narrativas pessoais: um estudo de metáforas na perspectiva da Linguística de Corpus. **Metáfora em Perspectiva** – Organon – Revista do Instituto de Letras da Universidade Federal do Rio Grande do Sul, Porto Alegre, v. 21, n. 43, p. 143-159, jul/dez 2007d.

_____. **Pesquisa em Linguística de Corpus com WordSmith Tools**. [S.I.]: [s.n.], 2006.

_____. A influência do tamanho do corpus de referência na obtenção de palavras-chave usando o programa computacional *WordSmith Tools*. **The ESPecialist**, São Paulo, v. 26, n. 2, p. 83-204, 2005.

_____. **Linguística de Corpus.** São Paulo: Manole, 2004.

_____. Linguística de Corpus: histórico e problemática. **D.E.L.T.A.**, São Paulo, v. 16, n.2, p. 323-367, 2000.

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi. **Corpus Linguistics:** Investigating Language Structure and Use. New York: Cambridge University Press, 1998.

BIBER, Douglas. Representativeness in Corpus Design. **Literary and Linguistic Computing**, Oxford, v.8, n.4, p. 243-257, 1993.

_____. Methodological Issues Regarding Corpus-Based Analysis of Linguistic Variation. **Literary and Linguistic Computing**, Oxford, v.5, n.4, p.257-269. 1990.

_____. **Variation across speech and writing**. Cambridge: Cambridge University Press, 1988.

CAMERON, Lynne; DEIGNAN, Alice. The Emergence of Metaphor in Discourse. **Applied Linguistics**, Oxford, v. 27, n. 4, p. 671-690, 2006.

CAMERON, Lynne. **Metaphor in Educational Discourse**. London: Continuum, 2003.

CAMERON, Lynne; LOW, Graham. **Researching and Applying Metaphor.** Cambridge: Cambridge University Press, 1999.

DEIGNAN, Alice. Corpus Linguistics and Metaphor. In: GIBBS, Raymond (Ed.). **The Cambridge Handbook of Metaphor and Thought**. New York: Cambridge University Press, 2008. p. 280-294.

_____. **Metaphor and Corpus Linguistics:** convergence evidence in language and communication research. Amsterdam: John Benjamins, 2005.

GASS, Susan; SELINKER, Larry. **Second Language Acquisition**: An introductory Course. 3. ed. New York: Routledge, 2008.

GIL, Maitê. **Metáfora no ensino de língua materna:** em busca de um novo caminho. Dissertação (mestrado em Linguística Aplicada). Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

GRANGER, Sylviane. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In: AIJMER, Karin. (Ed.). **Corpora and Language Teaching.** Amsterdam: John Benjamins, 2009. p. 13-32.

_____. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. **TESOL Quarterly**, v. 37, n. 3, p. 538-546, autumn, 2003.

_____. A Bird's eye view of learner corpus research. In: _____ et al. (Ed.). **Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.** Amsterdam: John Benjamins, 2002. p. 3-33.

_____. The computer learner corpus: a versatile new source of data for SLA research. In: _____(Ed.). **Learner English on Computer**. New York: Longman, 1998. p. 3-18.

HUNSTON, Susan. **Corpora in Applied Linguistics.** London: Cambridge University Press, 2002.

KAUFFMANN, Carlos H. **O corpus do jornal**: Variação linguística, gêneros e dimensões da imprensa diária escrita. Dissertação (mestrado em Linguística Aplicada). Pontifícia Universidade Católica de São Paulo, São Paulo, 2005.

LAKOFF, George; JOHNSON, Mark. **Metaphors we live by.** Chicago: The University of Chicago Press, 1980.

MCCARTEN, Jeanne. **Teaching Vocabulary**: Lessons from the corpus, lessons for the classroom. New York: Cambridge University Press, 2007.

MCENERY, Tony; XIAO, Richard; TONO, Yukio. (2006) **Corpus-Based Language Studies** – An advanced resource book. Oxon: Routledge, 2007.

MCENERY, Tony; WILSON, Andrew. (1996) **Corpus Linguistics:** An introduction. Edinburgh: Edinburgh University Press, 2004.

O'KEEFFE, Anne; McCARTHY, Michael; CARTER, Ronald. **From Corpus to Classroom:** Language Use and Language Teaching. Cambridge: Cambridge University Press, 2007.

OLIVEIRA, Lucia Pacheco. Linguística de Corpus: teoria, interfaces e aplicações. **Matraga**, Rio de Janeiro, v.16, n.24, jan./jun., 2009.

PACHECO, Aline. **A aquisição de morfemas em inglês como L2:** Uma análise dos padrões evolutivos através do BELC *(Brazilian English Learner Corpus).* Tese (doutorado em Teoria e Análise Linguística). Universidade Federal do Rio Grande do Sul, Porto Alegre, 2010.

PRAGGLEJAZ Group. MIP: A method for identifying metaphorically used words in discourse. **Metaphor and Symbol.** v. 22, n., p. 1–39, 2007.

RAYSON, Paul. **Matrix:** A statistic method and software tool for linguistic analysis through corpus comparison. Tese (doutorado em Ciência da Computação). Universidade de Lancaster, Lancaster, 2002.

REPPEN, Randi. Building a corpus: What are the key considerations? In: O'KEEFFE, Anne et al. (Ed.). **The Routledge Handbook of Corpus Linguistics.** New York: Routledge, 2010. p. 31-37.

SARMENTO, Simone. Linguística de Corpus e o Desenvolvimento de Material Didático para Inglês com Propósitos Específicos. In: _____ et al. (Ed.). **O Ensino do Inglês como Língua Estrangeira**: Estudos e Reflexões II. Porto Alegre: EDIPUCRS, 2009. p. 259-290.

_____. **O uso dos verbos modais em manuais de aviação em inglês:** um estudo baseado em corpus. Tese (doutorado em Teorias do Texto e do Discurso). Universidade Federal do Rio Grande do Sul, Porto Alegre, 2008.

SCOTT, Mike. **WordSmith Tools**. (1996). Oxford: Oxford University Press. Versão 6, 2012.

SHEPHERD, Tania M. G. O estatuto da Linguística de Corpus: metodologia ou área da Linguística? **Matraga**, Rio de Janeiro, v.16, n.24, p. 150-172, jan./jun.,2009.

STEEN, Gerard et al. **A Method for Linguistic Metaphor Identification**: From MIP to MIPVU. Amsterdam: John Benjamins, 2010.

TAGNIN, Stella; FROMM, Guilherme. CoMaprend – a experiência da construção de um corpus de aprendizes para estudo. **Domínios de Lingu@gem**, Uberlândia, v.2, n.2, 2008. Não paginado.

TOGNINI-BONELLI, Elena. **Corpus Linguistics at Work.** Amsterdam: John Benjamins, 2001.

VEREZA, Solange. O *lócus* da metáfora: linguagem, pensamento e discurso. **Cadernos de Letras da UFF –** Dossiê: Letras e cognição. Rio de Janeiro, n. 41, p. 199-212, 2010.

_____. Metáfora e Argumentação: Uma abordagem cognitiva-discursiva. **Revista Linguagem em (Dis)curso**, v. 7, n. 3, set./dez., 2007. Não paginado.

ANEXOS

ANEXO 1

Estrutura geral do BELC

Nível	Tarefa	Sujeitos	Palavras	Total de textos	Total de palavras
Beginner	Tarefa 1	90	8.465	252	21.856
	Tarefa 2	82	5.565		
	Tarefa 3	80	7.826		
Pre-Intermediate	Tarefa 1	113	13.708	314	37.180
	Tarefa 2	107	10.201		
	Tarefa 3	94	13.271		
Intermediate	Tarefa 1	107	14.889	271	39.504
	Tarefa 2	92	10.310		
	Tarefa 3	72	14.305		
Advanced	Tarefa 1	13	1.964	33	5.053
	Tarefa 2	11	1.204		
	Tarefa 3	9	1.886		
TOTAL				870	103.593