Universidade do Estado do Rio de Janeiro

Centro de Educação e Humanidades

Instituto de Letras

Gustavo Estef Lino da Silveira

Análise de quadrigramas na escrita em inglês como língua estrangeira: um estudo baseado em *corpus*

Rio de Janeiro

2014

Gustavo Estef Lino da Silveira

Análise de quadrigramas na escrita em inglês como língua estrangeira: um estudo baseado em *corpus*

Dissertação apresentada como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Letras, da Universidade do Estado do Rio de Janeiro. Área de concentração: Linguística.

Orientadora: Profª. Dra. Tania M. G. Shepherd

Rio de Janeiro

2014

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/CEHB

S587 Silveira, Gustavo Estef Lino da.
Análise de quadrigramas na escrita em inglês como língua estrangeira: um estudo baseado em corpus / Gustavo Estef Lino da Silveira. – 2014.
98 f.: il.

Orientadora: Tania Maria Granja Shepherd.
Dissertação (mestrado) – Universidade do Estado do Rio de Janeiro, Instituto de Letras.

1. Língua inglesa – Estudo e ensino – Falantes de português – Teses. 2. Linguística aplicada – Teses. 3. Linguística de corpus – Teses. 4. Línguas – Estudo e ensino – Rio de Janeiro (RJ) – Teses. 5. Aquisição da segunda linguagem - Teses. 6. Ensino auxiliado por computador – Teses. I. Shepherd, Tania M. G. (Tania Maria Granja). II. Universidade do Estado do Rio de Janeiro. Instituto de Letras. III. Título.

CDU 801

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou parcial desta dissertação desde que citada a fonte

_____ _____

Assinatura Data

Gustavo Estef Lino da Silveira

Análise de quadrigramas na escrita em inglês como língua estrangeira: um estudo baseado em *corpus*

Dissertação apresentada como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Letras, da Universidade do Estado do Rio de Janeiro. Área de concentração: Linguística.

Aprovada em 19 de março de 2014.

Banca examinadora:

_____

Prof. Dra. Tania M. G. Shepherd (Orientadora)

Instituto de Letras - UERJ

_____

Prof. Dra. Patricia Bértoli-Dutra

Instituto de Letras - UERJ

_____

Prof. Dra. Deise Prina Dutra

Universidade Federal de Minas Gerais

Rio de Janeiro

2014

DEDICATÓRIA

Dedico este trabalho aos meus pais, Sebastião (*in memoriam*) e Zoraide, por sempre acreditarem e investirem na minha educação.

AGRADECIMENTOS

À Deus, pois antes de iniciar todas as coisas, já escrevera o fim.

Ao meu pai, Sebastião, que sempre me incentivou o hábito de ler comprando livros, revistas em quadrinhos e jogos; tornando-se o melhor pai que pôde ser.

À minha mãe, Zoraide, pelo apoio incondicional aos meus estudos, por ser presente em momentos difíceis da minha vida e ter sido a maior incentivadora de eu ter estudado para este curso de mestrado.

À minha professora e orientadora Tania, pelas horas gastas de orientação e por ter acreditado e confiado em mim.

À minha tia Solange e meu padrinho Junior pelo apoio e suporte familiar tanto nos tempos bons quanto difíceis.

EPÍGRAFE

Seja passado o passado. Tome-se outra vereda e pronto.

Miguel de Cervantes

RESUMO

SILVEIRA, Gustavo Estef Lino da. *Análise de quadrigramas na escrita em inglês como língua estrangeira: um estudo baseado em corpus*. 2014. 98 f. Dissertação (Mestrado em Linguística) – Instituto de Letras, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2014.

O presente estudo tem como objetivo geral traçar um perfil das escolhas léxico-gramaticais da escrita em inglês de um grupo de aprendizes brasileiros na cidade do Rio de Janeiro, ao longo dos anos de 2009 a 2012, através da análise de sua produção de quadrigramas (ou blocos de quatro itens lexicais usados com frequência por vários aprendizes) em composições escritas como parte da avaliação final de curso. Como objetivo específico, a pesquisa pretendeu analisar se os quadrigramas produzidos estavam dentre aqueles que haviam sido previamente ensinados para a execução da redação ou se pertenceriam a alguma outra categoria, isto é, quadrigramas já incorporados ao uso da língua ou quadrigramas errôneos usados com abrangência pela população investigada. Para tal, foram coletadas composições escritas por aprendizes de mesmo nível de proficiência de várias filiais de um mesmo curso livre de inglês na cidade do Rio de Janeiro. Em seguida, essas composições foram digitadas e anotadas para constituírem um *corpus* digital facilmente identificável em termos do tipo e gênero textual, perfil do aprendiz, filial e área de origem do Rio de Janeiro. O estudo faz uso de preceitos e métodos da Linguística de *Corpus*, área da Linguística que compila grandes quantidades de textos e deles extrai dados com o auxílio de um programa de computador para mapear uso, frequência, distribuição e abrangência de determinados fenômenos linguístico ou discursivo. O resultado demonstra que os aprendizes investigados usaram poucos quadrigramas ensinados e, coletivamente, preferiram usar outros que não haviam sido ensinados nas aulas específicas para o nível cursado. O estudo também demonstrou que quando o gênero textual faz parte de seu mundo pessoal, os aprendizes parecem utilizar mais quadrigramas previamente ensinados. Isto pode querer dizer que o gênero pode influenciar nas escolhas léxico-gramaticais corretas. O estudo abre portas para se compreender a importância de blocos léxico-gramaticais em escrita em L2 como forma de assegurar fluência e acuracidade no idioma e sugere que é preciso proporcionar maiores oportunidades de prática e conscientização dos aprendizes quanto ao uso de tais blocos.

Palavras-chave: Linguística aplicada. Escrita em inglês. Linguística de *corpus*. Quadrigramas.

ABSTRACT

SILVEIRA, Gustavo Estef Lino da. *Analysis of quadrigrams in EFL writing*: a corpus-based study. 2014. 98 f. Dissertação (Mestrado em Linguística) – Instituto de Letras, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2014.

This study seeks to trace the profile of lexico-grammatical choices of a group of apprentice writers in the city of Rio de Janeiro, between 2009 and 2012. To this end it analyses the apprentices´ production of 4-grams (or rather blocks of four lexical items used with relative frequency by a number of apprentices) in written compositions, as part of their final assessment. Specifically, the research aimed to analyse whether the 4-grams produced by the apprentices had been taught previously as part of their composition lessons or whether they belonged to some other category. In other words, namely 4-grams already internalized as part of their language use of erroneous 4-grams used frequently and extensively by the subjects investigated. Thus, compositions written by apprentices at the same proficiency level were collected at various branches of a private English school in the city of Rio de Janeiro. Subsequently, these compositions were typed and tagged in order to compile a digital *corpus* easily identified in terms of type and textual genre, apprentice profile, branch and area of the city of Rio de Janeiro. The study makes use of precepts and methods of *Corpus* Linguistics, an area of Linguistics that collects large quantities of texts and from them extracts data with the help of a computer programme in order to map use, frequency, distribution and range of a certain linguistic or discursive phenomena. The results demonstrate that the apprentices studied made little use of 4-grams that had been taught them and, collectively, they preferred to use other n-grams that had not been taught in the specific lessons of the level. The study has also shown that when the textual genre is part of one's personal life, the apprentices seem to make use of more previously taught 4-grams. This may lead to believe that the genre may influence the choice of correct lexico-grammatical items. The study creates a research space for the understanding of the importance of lexico-grammatical chunks in L2 writing as a means of ensuring fluency and accuracy in the target language. In addition, it also suggests that more opportunities of practice should be offered to learners so that they become aware of the use of such chunks.

Keywords: Applied linguistics. Writing in English. *Corpus* linguistics. 4-grams.

SUMÁRIO

	INTRODUÇÃO	10
1	**LÍNGUA E LINGUAGEM**	17
1.1	**O que são língua e linguagem?**	17
1.2	**De onde vem a linguagem?**	18
2	**APRENDIZAGEM DE UMA SEGUNDA LÍNGUA: BREVE INTRODUÇÃO**	23
2.1	**Aprendizagem ou aquisição de uma segunda língua?**	23
2.2	**Interlíngua**	30
3	**AVALIANDO A PRODUÇÃO DO APRENDIZ: O ERRO**	33
3.1	**Breve perspectiva histórica do erro ...**	34
3.1.1	O que é o erro?	35
3.1.2	Tipos de erros	37
3.2	**Escrita de aprendiz e análise de erros baseada por computador**	40
3.3	**Colocação e coligação**	42
3.4	**Blocos léxico-gramaticais**	43
3.5	**O princípio idiomático e o princípio da livre escolha**	45
4	**MATERIAIS E MÉTODOS**	47
4.1	**Introdução: por que usar o computador no tratamento de dados textuais empíricos?**	47
4.2	***Corpus: corpus* de aprendiz**	49
4.3	**O *corpus* de estudo**	51
4.4	**A identificação de arquivos em um *corpus***	55
5	**ANÁLISE DOS DADOS**	57

5.1 **Redações e suas distribuições por filiais, gêneros textuais e tipos de erro** ... 57

5.2 **Quadrigramas em redações de curso livre** 60

5.3 **Quadrigramas errôneos** 64

5.4 **Para uma categorização de quadrigramas: a proposta de Hyland** 73

CONCLUSÃO 80

REFERÊNCIAS 89

ANEXOS 93

INTRODUÇÃO

Esta dissertação encontra-se na confluência de duas subáreas da Linguística - a Linguística Aplicada (LA) e a Linguística de *Corpus* (LC). Isso se dá porque, em primeiro lugar, a dissertação analisa blocos léxico-gramaticais de quatro palavras (doravante chamados quadrigramas) na escrita de aprendizes de inglês como língua estrangeira como ponto de partida para reflexões sobre a aprendizagem e aquisição de aspectos de uma segunda língua (doravante L2). Em segundo lugar, são utilizados princípios da Linguística de *Corpus* enquanto área e metodologia de pesquisa, ao compilar, extrair e analisar grandes quantidades de dados oriundos das composições escritas por esses aprendizes, com o auxílio de um programa de computador.

O presente estudo entende a amplitude da Linguística e como esta se encontra multi-ramificada, abrangendo diversas áreas do conhecimento científico. Com isso, é possível dizer que esta pesquisa se alia aos estudos de Dutra (2010) e Berber-Sardinha (2004) que aproximam ambas a Linguísticas Aplicada e de *Corpus* na tentativa de entender o fenômeno da escrita em L2 do aprendiz brasileiro.

Em inglês, o foco em blocos léxico-gramaticais tem uma história que remete a um passado recente. Esses blocos em inglês receberam uma gama de rótulos, que à primeira vista parecem sinônimos. Entretanto, uma rápida busca no site de pesquisa Google n-gram viewer nos permite constatar que os rótulos estão sempre atrelados a pesquisadores e a concepções distintas.

"Lexical chunks" e "lexical bundles", por exemplo, surgiram há bem pouco tempo nos livros acadêmicos em língua inglesa, mas de formas distintas. O primeiro deles, o item "lexical chunks", aparece pela primeira vez no sub*corpus* de inglês geral do *corpus*[1] online acessado pelo n-gram viewer, em 1970 e mesmo assim com uma frequência inexpressiva, tendo sido apenas encontrado em quantidade considerável no ano de 2003, atingindo um ápice com cinco vezes mais ocorrências que no fim dos anos 1980, conforme a Figura 1 a seguir:

Figura 1- Evolução do uso do item "lexical chunks" no site Google n-gram viewer.
Fonte: https://books.google.com/ngrams.

[1] Definirei *corpus* e *corpus* de aprendiz na seção 4.2 desta dissertação.

A busca por “lexical bundles” mostra que a expressão aparece discretamente na década de 1960, depois desaparece, retornando em frequência ascendente nos anos 1990, mas com grande pico no ano de 2008 aumentando em cinco vezes mais sua frequência e provavelmente nos demais anos a partir de então, já que o *corpus* termina naquele ano, de acordo com a Figura 2 abaixo:

Figura 2- Evolução do uso do item “lexical bundles” no site Google n-gram viewer.
Fonte: https://books.google.com/ngrams.

O aumento de ambas as ocorrências “lexical chunks” e “lexical bundles” no final dos anos 1990 conforme demonstram as Figuras 1 e 2 parecem coincidir com a publicação de Lewis (2003), na qual grande ênfase é dada ao ensino de blocos lexicais e à visão da linguagem como aglomerados de blocos de palavras. Pode-se ver que no início do aumento da frequência demonstrada pelas figuras grande atenção estava sendo dada ao ensino de blocos lexicais. Já quando as pesquisas se voltam para os estudos computacionais de feixes lexicais com as publicações de Biber (2006) e Hyland (2008) pode-se ver na Figura 2 o ano de 2008 culminando com o ápice de achados da combinação “feixes lexicais” pois o foco dos estudos está na extração e descrição de dados em *corpora*.

Esta busca informal por termos em *corpora* de livros acadêmicos pode sugerir que estudos envolvendo blocos e feixes lexicais são recentes e podem ser considerados de crescente importância na área de Linguística Aplicada, pois ainda há espaço para novas abordagens e investigações acerca da aquisição de blocos léxico-gramaticais por aprendizes de uma segunda língua.

Desta forma, este estudo tem como foco a análise de composições escritas em língua inglesa como L2, como base da avaliação da aquisição de blocos de quatro itens léxico-gramaticais, aqui chamados de quadrigramas. Para entendermos o que é um quadrigrama, precisamos saber sobre a composição do termo n-grama. Esse é um termo que advém da área

de linguística computacional que significa uma sequência contígua de um número ‘n’ de itens, encontrados em um texto ou textos. Portanto, um quadrigrama é uma sequência de quatro itens. A presente dissertação se alinha mais ao uso descrito pela Figura 2 pois utiliza *corpora* digitais e extração de dados com programas de computador. Alinha-se também aos trabalhos desenvolvidos no Brasil acerca do assunto, destacando-se os de Shepherd (2009), Dutra (2010) e Dutra e Berber-Sardinha (2010).

Independentemente do nome que recebem, o que se sabe sobre n-gramas é que eles podem constituir o que entendemos como combinações pré-construídas de multi-palavras denominadas ‘princípio do idioma ou idiomático’ – a alternativa que se conjuga com o ‘princípio da livre escolha’ normalmente utilizados por usuários de uma língua na construção de textos escritos ou falados (Sinclair, 1991). A presença dessas combinações pré-construídas confere idiomaticidade a esses textos e por isso, deveriam ser incorporadas às práticas de aprendizes de inglês como L2 como forma de lhes garantir fluência, como veremos na seção 4.3 do capítulo 4 quando discuto a diferença entre os princípios idiomático e de livre escolha.

O objetivo deste estudo é fazer um levantamento dos quadrigramas utilizados por uma população de um curso livre de inglês em redações produzidas durante provas. A finalidade do trabalho é tentar entender a natureza desses blocos como elementos ensinados ou criados. A escolha de provas como *corpus* de estudo justifica-se pelo fato de que assim, consegue-se otimizar o número de aprendizes analisados, já que nas tarefas de casa, nem todos os alunos são obrigados a escrever as redações. Sendo assim, durante as provas, até mesmo quem não praticou a habilidade escrita tem de produzir um texto como parte da avaliação do semestre.

Outro motivo para a escolha de provas como *corpus* de estudo está no fato de que os alunos que já escreveram textos semelhantes ao longo do semestre, já tiveram a chance de praticar a escrita das redações, receberam o *feedback* do professor com as correções necessárias e, assim, podem escrever um texto melhor. Um último motivo seria assegurar que os aprendizes não fizeram consulta a nenhum material de referência durante o processo de escrita, tais como anotações ou dicionários. Em outras palavras, quaisquer blocos lexicais usados pelos aprendizes seria resultado de internalização e não de cópia.

Começo este capítulo descrevendo a motivação desta pesquisa através da minha prática docente como professor de inglês. Em seguida, justifico a pertinência de se estudarem quadrigramas na aprendizagem de uma segunda língua. Mais adiante, teço as perguntas que nortearão esta pesquisa e que tentarei responder no capítulo de conclusão. Por fim, reitero a importância de se revisitarem alguns preceitos teóricos que fundamentam a Linguística para,

depois, seguir rumo a áreas mais específicas da mesma, conforme será detalhado no capítulo de fundamentação teórica.

Uma motivação inicial

Há cerca de uma década venho exercendo o magistério no ensino da língua inglesa em cursos livres de idiomas e parte desse tempo, venho trabalhando com as abordagens Comunicativa e Lexical. Em minha prática docente tenho observado que o léxico e a gramática andam lado a lado. Se para alguns educadores e alunos, essas áreas ainda parecem ser distintas, pode-se ressaltar que o professor de idiomas da geração pós-*corpus* está começando a quebrar esse paradigma. Mesmo muitas vezes inconscientemente, esse novo professor parece ter mais entendimento que a língua em uso consiste de um grande conjunto de blocos léxico-gramaticais. Esse novo professor entende que não se deve apenas ensinar esses blocos, mas também incentivar que seus alunos os utilizem através da prática constante em atividades em sala de aula, a fim de que tais blocos possam ser internalizados e utilizados de maneira espontânea tanto na fala quanto na produção escrita em inglês.

A ideia inicial de meu estudo surgiu a partir de um grupo de aprendizes que estava, à época, estudando comigo há cerca de um ano e meio. Ao longo das aulas e no nível em que os alunos se encontravam (intermediário-avançado), grande ênfase era dada ao ensino de blocos léxico-gramaticais e à prática dos mesmos para que os aprendizes “soassem como um falante nativo”, ou seja, adquirissem mais idiomaticidade em seu discurso e não lançassem mão de estruturas calcadas na língua materna. Tal direcionamento era dado pelo programa curricular da instituição que já fornecia o esboço das aulas e o conteúdo a ser trabalhado em sala de aula.

O que pôde ser informalmente observado por mim é que o professor ensinava blocos léxico-gramaticais, os alunos aparentemente entendiam o significado através do contexto ou explicação do próprio professor ou de algum colega e praticavam os novos blocos em um contexto restrito, muitas vezes oral. Embora houvesse reforço da matéria através do dever de casa, a maioria dos alunos na faixa etária estudada (entre 15 e 18 anos) parecia não dar tanta importância ao dever de casa do curso de inglês, já que estavam envolvidos com o pré-vestibular ou iniciando a vida universitária. O que poderia ter sido uma oportunidade extra de aprendizagem e revisão do conteúdo ensinado era relegado a um plano secundário.

Na aula seguinte, um novo conteúdo era apresentado e a mesma dinâmica de aula ocorria sem haver muita flexibilidade no calendário ou oportunidades no guia do professor

para exercícios de revisão, cabendo a esse, quando quisesse e achasse necessário, criar suas próprias atividades de revisão, fossem elas orais ou escritas.

Mas mesmo havendo alunos que se destacassem tanto em fluência quanto em acuracidade, havia outros que cometiam erros muito básicos ao usarem a língua. Tais erros, em alguns casos, eram os mesmos produzidos por aprendizes iniciantes: flexão da terceira pessoa do singular, plurais irregulares, entre outros. Alguns alunos tinham também muita dificuldade em incorporar os novos blocos léxico-gramaticais, estruturas ditas avançadas, em seu discurso quando participavam de alguma conversação livre e espontânea, demonstrando um nível linguístico equiparável a de níveis anteriores. É importante mencionar que os alunos pareciam aprender os novos blocos léxico-gramaticais na aula em que os mesmos eram ensinados, pois os praticavam em uma situação controlada pelo professor, acertavam em boa parte o dever de casa, quando o faziam, mas pareciam esquecê-los ao longo do caminho conforme novos blocos eram introduzidos nas aulas seguintes.

O “X” da questão parecia se concentrar no fato de que grande parte do conteúdo somente seria revisto pelo aluno na aula anterior à prova - a aula de revisão, na qual os aprendizes fariam exercícios escritos de fixação, preenchendo colunas ou fazendo múltipla escolha.

Outras formas de se avaliar a aprendizagem de blocos léxico-gramaticais além da avaliação formal da prova escrita seriam a redação como tarefa de casa e a prova oral. Entretanto, a prática de escrever um texto como tarefa de casa após uma aula de composição escrita, fazia com que muitos alunos produzissem um texto simples sem evidência dos blocos léxico-gramaticais ensinados e contendo usos errôneos linguísticos aparentemente já sedimentados. É importante que se diga que não fazia parte da prática pedagógica da instituição pedir que os alunos reescrevessem as redações refazendo as correções sinalizadas pelo professor.

Portanto, ao final do semestre, a sensação que nos restava enquanto professor desse nível era que nada tinha sido retido, internalizado ou aprendido e a minha sensação era de que o que eu ensinara não servira de nada.

A justificativa do estudo e perguntas de pesquisa

Muitos estudos têm demonstrado a necessidade de incluir blocos léxico-gramaticais no ensino da língua estrangeira. Outros têm ressaltado a importância de analisarmos esses blocos

como forma de avaliarmos idiomaticidade e por conseguinte o nível de proficiência em um idioma (HYLAND, 2008; LEWIS, 2009). Entretanto, estudos dessa natureza envolvendo populações de aprendizes brasileiros de língua inglesa ainda são escassos. Este estudo vem, portanto, acrescentar aos estudos existentes, não só por ter como sujeitos de pesquisa aprendizes adolescentes de inglês, mas também por ter um grupo com características bem homogêneas, isto é todos os adolescentes estudavam em um mesmo nível de uma mesma instituição particular de ensino de língua inglesa, de uma única cidade (Rio de Janeiro), e todos começaram a aprender a língua inglesa ainda crianças.

A presente dissertação analisa coletivamente os dados de um *corpus* de textos redigidos por adolescentes, sob condições de testagem. Os aprendizes, oriundos de várias regiões do Rio de Janeiro, começaram a estudar inglês entre 8 e 10 anos de idade e estão concluindo um curso livre de inglês que durou cerca de 7 anos. Os gêneros das redações já foram estudados em classe e blocos léxico-gramaticais específicos ensinados a esses alunos. Partindo dessas condições de coleta, e tendo como foco o *corpus* de redações desses alunos, as perguntas que norteiam a pesquisa são:

a) Os textos dos alunos quando submetidos a um programa de análise revelam se os aprendizes fazem uso de quadrigramas conforme o ‘princípio idiomático’ ou do ‘princípio da livre escolha’? Dos quadrigramas que porventura são extraídos quais foram ensinados formalmente, quais não foram?
b) Os quadrigramas ensinados formalmente são usados exatamente como foram ensinados, ou sofrem algum tipo de elaboração/transformação? Em caso positivo, essa transformação estaria baseada em erros?
c) Usando-se o sistema para categorização de blocos léxico-gramaticais desenvolvidos por Hyland (2008), como podemos identificar que quadrigramas são usados pelos aprendizes?
d) Considerando que os quadrigramas foram contextualizados a partir de gêneros, e que esses foram também ensinados, há alguma diferença numérica no uso dos quadrigramas nos diferentes gêneros? Em caso positivo, que gênero propicia condições maiores de utilização dos quadrigramas ensinados?
e) A partir dos resultados finais, é possível se formular alguma conclusão sobre a pertinência de ensinar quadrigramas para posterior testagem?

O trabalho investiga se os alunos utilizam em seus textos quadrigramas ensinados no programa curricular. Além disso, é levantada a questão sobre a influência do gênero na escolha de blocos léxico-gramaticais. Também é escopo deste trabalho uma análise das possíveis transformações feitas quando os aprendizes produzem quadrigramas diferentes daqueles ensinados.

Essas são as perguntas que tentarei responder ao longo desta pesquisa. No entanto, sendo este estudo uma interseção de áreas da Linguística como a Linguística Aplicada e Linguística de *Corpus*, convêm que se resgatem alguns princípios teóricos acerca desses três campos de estudo. O primeiro deles de caráter mais amplo e genérico será explicitado na seção a seguir, onde discutirei a dicotomia língua e linguagem. As demais, por serem mais pertinentes ao estudo, serão esmiuçadas nos capítulos que seguem.

Para responder a essas perguntas a dissertação está estruturada em 5 capítulos, além de introdução e conclusão. Na introdução apresento o objetivo do estudo e as perguntas de pesquisa. No primeiro discuto a relevância dos conceitos de língua como sistema e língua/gem como uso e a pertinência de incluí-los em uma dissertação que alia uma área da LA, que é a da aquisição de determinados fenômenos de uma língua estrangeira e a coleta de dados empíricos de forma não experimental. Se o capítulo anterior trata de conceitos como língua e linguagem de uma forma ampla, no segundo, faço um apanhado histórico sobre a diferença conceitual entre aprendizagem e aquisição de uma segunda língua (L2). O capítulo dois apresenta as preocupações da LA sobre a produção de aprendizes na área pré-*corpus* e dessa forma discute a diferença entre aprendizagem e aquisição, e a possível existência de uma interlíngua. No terceiro, apresento perspectivas analíticas sobre o conceito de ´erro´. Em seguida, discuto a importância que a LC deu à análise de erros trazendo vigor a essa área outrora esquecida e negligenciada. Menciono como estudos envolvendo erros tomaram um novo rumo através da compilação de grandes quantidades de textos digitais e digitalizados e extração de erros com o auxílio de programas de computador. No capítulo quatro, descrevo o porquê de se utilizar a LC como metodologia de pesquisa. Faço um detalhamento da compilação dos *corpora* de estudo e de referência e detalho procedimentos da pesquisa. No capítulo cinco, analiso os blocos extraídos do *corpus* e a análise dos mesmos. Por fim, na conclusão, respondo às perguntas de pesquisa e teço as conclusões finais deste trabalho, trazendo os questionamentos surgidos com a pesquisa e os possíveis direcionamentos futuros.

1 **LÍNGUA E LINGUAGEM**

Neste capítulo discorro sobre itens que, à primeira vista, não parecem ter muito a ver com ensino e aprendizagem de uma segunda língua ou língua estrangeira, mas que se fazem necessários para refletir sobre a gênese dos pensamentos mais recentes sobre aquilo que é língua como sistema e linguagem como uso e sobre métodos de ensino de uma segunda língua.

Começo a seção traçando um panorama da dicotomia língua e linguagem. Mais adiante, revisito as duas principais correntes linguísticas que tentam explicar como ocorre a aquisição de uma língua, correntes essas, que são contrastantes, mas de suma importância para se compreender a visão a que o presente estudo se atrela.

1.1 **O que são língua e linguagem?**

Os seres humanos são os únicos animais dotados da faculdade da fala. Embora primatas e golfinhos possuam um considerável quociente de aprendizagem e características genéticas semelhantes a dos seres humanos, os mesmos não possuem à sua disposição um órgão fonador responsável por desenvolver a fala em um contexto criativo e independente de uma resposta a estímulos (GURGEL, 1999).

Todavia, inúmeros estudos dão conta de uma suposta linguagem animal. Os estudos sobre o assunto sugerem que os animais estabelecem padrões de comunicação entre si a fim de executar tarefas e transmitir mensagens em seus grupos. Ainda assim, a linguagem animal ainda é de fato abstrata se comparada à capacidade de comunicação humana. Os seres humanos fazem uso da linguagem de uma forma criativa, plural, emitindo sons que não são simples repetições ou respostas a estímulos emocionais internos ou externos. Isso é um dos inúmeros fatores que diferencia a nossa linguagem da dos animais (GURGEL, 1999).

A dicotomia ‘língua’ e ‘linguagem’ foi levantada nos estudos de Saussure. Para o linguista, a língua (“langue”) seria exterior ao indivíduo, um dado sistemático, social, coletiva, abstrata e regulada por princípios. Seria um sistema de signos, um conjunto de unidades que se relacionam organizadamente dentro de um todo (PETTER, 2011). Já a linguagem (“parole”), ou fala seria individual, particular. Cada falante possuiria a sua própria

fala, ou seja, a sua maneira específica de fazer uso da “langue”. Logo, a “parole” seria assistemática, individual e caótica e por isso não observável.

Para o linguista genebrino, portanto, o objetivo dos estudos da Linguística deveria ser a “langue” porque era observável e não a “parole”, porque era individual e aparentemente caótica.

Anos mais tarde, como discutirei quando tratar especificamente da Linguística de *Corpus*, ficaria evidenciado que a “parole”, ainda que individual, não é caótica, é padronizada. Daí podermos embarcar em um estudo diagnóstico como o da presente dissertação, envolvendo algumas centenas de aprendizes.

1.2 **De onde vem a linguagem?**

Duas grandes correntes linguísticas que tentam explicar o processo de aquisição da linguagem e tem visões contrastantes entre si são o gerativismo e o funcionalismo. Na primeira, temos Noam Chomsky como seu maior expoente com a visão racionalista de ver a linguagem e na segunda, temos os estudos de Halliday e a Gramática Sistêmico-Funcional atrelada à visão empirista da língua. Na abordagem racionalista “há uma maior preocupação com a forma pela qual a mente processa a linguagem, enquanto que na abordagem empírica, a preocupação está em observar a ocorrência de dados naturais” (ALMEIDA, 2010, p. 25).

Segundo Chomsky todos os falantes são dotados de uma gramática universal inata que ajuda a explicar as similaridades no desenvolvimento da aquisição de uma língua materna. (THORNBURY, 2008). Para Chomsky, essa gramática é formada por princípios invariáveis e permeáveis a todas as línguas. Isso explica, em parte, o fato de crianças de diferentes línguas passarem pelos mesmos estágios durante a aquisição da linguagem.

Os gerativistas acreditam que “as infinitas sentenças de uma língua são formadas a partir da aplicação de um finito sistema de regras (a gramática)” (KENEDY, 2010, p. 131). Sendo assim, somos capazes de produzir um sem número de frases em uma língua devido à capacidade de criação do falante e à um número infinito de possibilidades que a língua oferece. É importante ressaltar que, segundo Wilkins (1976), a mesma língua opera de acordo com regras finitas e pré-estabelecidas que restringem as opções do falante. Ou seja, embora os meios pelas quais as frases são construídas sejam limitados o número de sentenças distintas em uma língua é infinito.

Segundo Kenedy (2010), Chomsky aprofundou esse conceito ao abordar a criatividade do falante e sua habilidade de construir frases através da substituição de sintagmas tanto no eixo paradigmático quanto no sintagmático da frase. Dessa forma, o linguista chama a atenção para o fato de a língua ser dotada de possibilidades infinitas de combinações e o falante ter a criatividade de fazer quaisquer combinações que lhe aprouver.

Kenedy (2010) afirma ainda que um falante consegue distinguir a agramaticalidade de uma frase inconscientemente através do conhecimento interno das leis que regem a língua e governam a formação das frases, a este fenômeno é chamado de “competência linguística”.

Em via oposta, há o “desempenho linguístico”, que é a performance linguística do falante em inúmeras situações e “envolve diversos tipos de habilidades que não são linguísticas, como atenção, memória, emoção, nível de estresse, conhecimento de mundo, etc.” (KENEDY, 2010, p. 134).

Pode-se dizer que é interesse exclusivo do gerativismo apenas a competência linguística e não o desempenho. Por isso, os estudos gerativistas não retiram dados linguísticos do uso concreto da língua na vida cotidiana. Toda e qualquer variável extralinguística que possa influenciar no uso da linguagem não é domínio do gerativismo.

Por isso, Chomsky sugeria que “um *corpus* não poderia ser uma ferramenta útil para um linguista uma vez que o *corpus* deveria descrever a língua através da competência e não da performance. Além disso, um *corpus* não poderia representar uma língua em sua totalidade; então, não poderia ser representativo dessa língua” (ALMEIDA, 2010, p. 26).

Sabemos hoje em dia que tal afirmativa não é verdadeira, com a LC, que será discutida no capítulo 4, podemos representar a linguagem de aprendizes, a norma padrão de uma língua, um dialeto específico, entre outros mediante a coleta adequada de um *corpus*.

Por outro lado, o funcionalismo vê a língua como um sistema de probabilidades, no qual frequência e uso são os fatores que irão ditar os rumos da análise linguística. Embora os gerativistas trabalhem com um sistema de possibilidades, para os funcionalistas a língua deve ser analisada segundo suas funções, senão não há sentido em utilizá-la.

A visão da linguagem como sistema probabilístico pressupõe que, “embora muitos traços linguísticos sejam possíveis teoricamente, tais traços não ocorrem com a mesma frequência.” (BERBER-SARDINHA, 2004, p. 30-31). Tal característica é extremamente importante quando se trata de um *corpus* de aprendizes. Aqueles aprendizes que alcançam proficiência em um determinado idioma dominam os traços linguísticos que são possíveis teoricamente em termos de frequência e distribuição. Aqueles que não adquiriram a língua

podem usar traços que sejam possíveis, mas não os usam nem com a frequência e nem com a distribuição desejadas, como será visto no capítulo de análise do *corpus*.

Temos aqui o paradigma entre linguagem possível e linguagem provável. Para Chomsky e os racionalistas a língua é formada de um número infinito de sentenças denominada linguagem possível. Este estudo tenta discorrer sobre a probabilidade de certas construções ocorrerem na língua, mesmo sendo consideradas gramaticalmente corretas.

Se a língua é formada de possibilidades infinitas, a língua possível de ser falada pode, não necessariamente, ser provável de ser produzida por um falante de uma determinada comunidade discursiva. Em outras palavras, nem sempre "linguagem possível" pode vir a significar "linguagem provável". Isso quer dizer que, o que é gramaticalmente correto pode vir a não estar mais em voga ou não ser usado em determinado contexto. Com o auxílio de métodos computacionais a LC nos permite compilar grandes quantidades de textos e analisar se a língua produzida pelos aprendizes de L2 estudados seria uma linguagem correta mas que não estaria mais em uso, ou seja, que não seria possível de ser produzida por falantes nativos de inglês. Para este estudo, a visão gerativista de possibilidade e linguagem correta e possível mas não provável de ocorrer não irá nos servir de parâmetro de análise.

Em oposição à visão gerativista da língua, o funcionalismo compreende a linguagem sob outra perspectiva, a da probabilidade de determinada sentença ser produzida em um determinado contexto ou não. "Os funcionalistas concebem a linguagem como um instrumento de interação social, alinhando-se, assim, à tendência que analisa a relação entre linguagem e sociedade." (CUNHA, 2010, p. 157). Logo, os funcionalistas irão analisar a língua segundo contextos que envolvam os interlocutores, os propósitos comunicativos e o contexto discursivo. A análise vai além da sentença gramatical e os estudos focam no uso e na possibilidade de uso e não na estrutura gramatical em si.

Os funcionalistas entendem a linguagem como um sistema de probabilidades. Ou seja, investigam a probabilidade de um falante pronunciar determinada frase dentro de certo contexto. Mesmo sendo considerada incorreta, há uma grande probabilidade de os falantes de determinada região ou em um contexto específico fazerem uso de uma sentença vista pelos gerativistas como agramatical. É campo de investigação do funcionalismo o estudo desses fenômenos extralinguísticos que influenciam a produção linguística.

Dentro desse paradigma sobre a visão da linguagem temos de um lado Halliday que vê a linguagem como um sistema de probabilidades e do outro Chomsky, que a enxerga como possibilidade.

As divergências entre funcionalistas e gerativistas vão além porque adentram o campo da aquisição de linguagem. Para estes, o processo de aquisição de linguagem é inato, logo, todos os seres humanos já nascem com uma pré-disposição biológica a desenvolver a faculdade da linguagem. Já os funcionalistas defendem que a língua é adquirida a partir da exposição da criança em situações de interação com outros membros de sua comunidade de fala e que a criança constrói a gramática da sua língua de acordo com sua necessidade e habilidade comunicativa (CUNHA, 2010).

O funcionalismo opõe-se ao gerativismo em relação ao processo de aquisição da linguagem pela forma como ambas as correntes compreendem-na. Por um lado, os gerativistas explicam a aquisição da linguagem em termos de uma capacidade humana específica para a aprendizagem da língua (CUNHA, 2010). Por outro lado, para os funcionalistas, é com base nos dados linguísticos a que a criança é exposta em situações de interação com os membros de sua comunidade de fala, que esta constrói a gramática de sua língua. Em outras palavras, o que Cunha (2010) quer dizer é que para o funcionalismo a aquisição da linguagem é um processo social, construído a partir das relações com outros falantes, enquanto que para os gerativistas ela é um processo biológico e interno ao homem.

Embora o funcionalismo defenda que a língua é adquirida a partir da exposição e interação social em contextos comunicativos, a criança é dotada de uma capacidade cognitiva que torna possível a aprendizagem da língua, assim como o aprendizado de outras habilidades, sendo a língua apenas uma vertente entre tantas outras habilidades que a criança irá desenvolver.

Parece com isto, que todo o esforço de linguistas alinhados à visão de língua de Sausurre, que estava preocupado em analisar o sistema ou a “langue” pode vir a ser considerado um discurso esvaziado de sentido e importância nos dias atuais. Se a língua é mutável e no caso da aprendizagem de uma língua estrangeira, os aprendizes adquirem apenas um fragmento daquela, seria interessante estudar a linguagem em uso e toda idiossincrasia que cada aprendiz carrega com sua variedade linguística.

Muito esforço parece ter sido empregado pelos gerativistas ao longo dos anos em se estudar a língua externa ao aprendiz, ao sistema de regras e à competência linguística, negligenciando habilidades como o desempenho do falante em certas situações e a real função de um enunciado e a probabilidade de uso e ocorrência de um enunciado dentro de uma língua. Com isso, é de suma importância que se revisitem os arcabouços teóricos da Linguística para que se reafirme a visão de linguagem a que este estudo se atrela, já que muitos teóricos da área de aquisição de uma segunda língua costumam estar anelados à visão

gerativista ao definirem seus conceitos de língua e aquisição de uma segunda língua, conceitos estes como interlíngua que irei destrinchar no capítulo a seguir.

Em suma, neste capítulo revisitei alguns conceitos como a dicotomia língua e linguagem, além das principais ideias acerca da aquisição da linguagem para funcionalistas e gerativistas porque existe um fio condutor que se inicia com a discussão dos estudos Saussureanos. Ao falar da “parole”, esta nos fornece meios para se iniciar a discussão acerca da língua em uso como é o caso do presente estudo. Passo, no capítulo a seguir, a focar em conceitos e estudos pertinentes à área de Linguística Aplicada, mais especificamente, aquisição de uma segunda língua e a diferença entre aprender e adquirir uma L2 e conceitos como interlíngua e erros porque esses conceitos são importantes para se entender a mentalidade do linguista aplicado da era pré-*corpus*.

2 **APRENDIZAGEM DE UMA SEGUNDA LÍNGUA: BREVE INTRODUÇÃO**

Neste capítulo, serão apresentados, em primeiro lugar, alguns conceitos-chave sobre aprendizagem de uma segunda língua. Em seguida, vem uma seção explanatória sobre aquisição de segunda língua e o que significa adquiri-la. Como ficará claro no decorrer da discussão, é necessária a inserção desses dois tópicos como preâmbulo a nosso capítulo analítico. A dissertação explora uma situação de 'ensino' formal de expressões compostas de quatro palavras associadas a determinados gêneros escritos para verificar se e como as referidas expressões são produzidas em situação de testagem. Entender o que é aprender e adquirir é importante para lidar com os resultados analíticos do capítulo 5.

2.1 **Aprendizagem ou aquisição de uma segunda língua?**

Em 1989, Spolsky (1989, p.28) elaborou um quadro sinótico sobre os elementos envolvidos no processo de aprendizagem de uma L2. De acordo com esse quadro, replicado na Figura 3 a seguir, para que haja aprendizagem em uma L2 uma gama de fatores seria necessária. Em primeiro lugar, Spolsky cita a questão do contexto social como sendo de vital importância para iniciar esse processo. Para o autor o contexto social propiciaria oportunidades de aprendizagem tanto formais quanto informais. Spolsky também afirma que o contexto social leva a diversos tipos de atitudes que podem aparecer no aprendiz sob a forma de motivação. A motivação, que viria a ser estudada por Ur (1996) pode ser intrínseca ou extrínseca, ou seja, inerente ao aprendiz ou advinda de fatores externos, respectivamente. No entanto, em sua obra de 1989, Spolsky não menciona os tipos de motivação. Por conseguinte, o autor sugere que a motivação se une a outras quatro características pessoais que seriam influenciadoras no processo de aprendizagem de uma L2, a saber: idade, personalidade, capacidades e conhecimento prévio. Todas essas características explicariam o uso que o aprendiz faz das oportunidades de aprendizagem disponíveis (sejam elas formais ou informais). Por último, o autor sugere que parece ser necessário que haja uma ação do aprendiz sobre a situação em que se encontra, podendo, então, o aprendiz obter resultados linguísticos e não linguísticos.

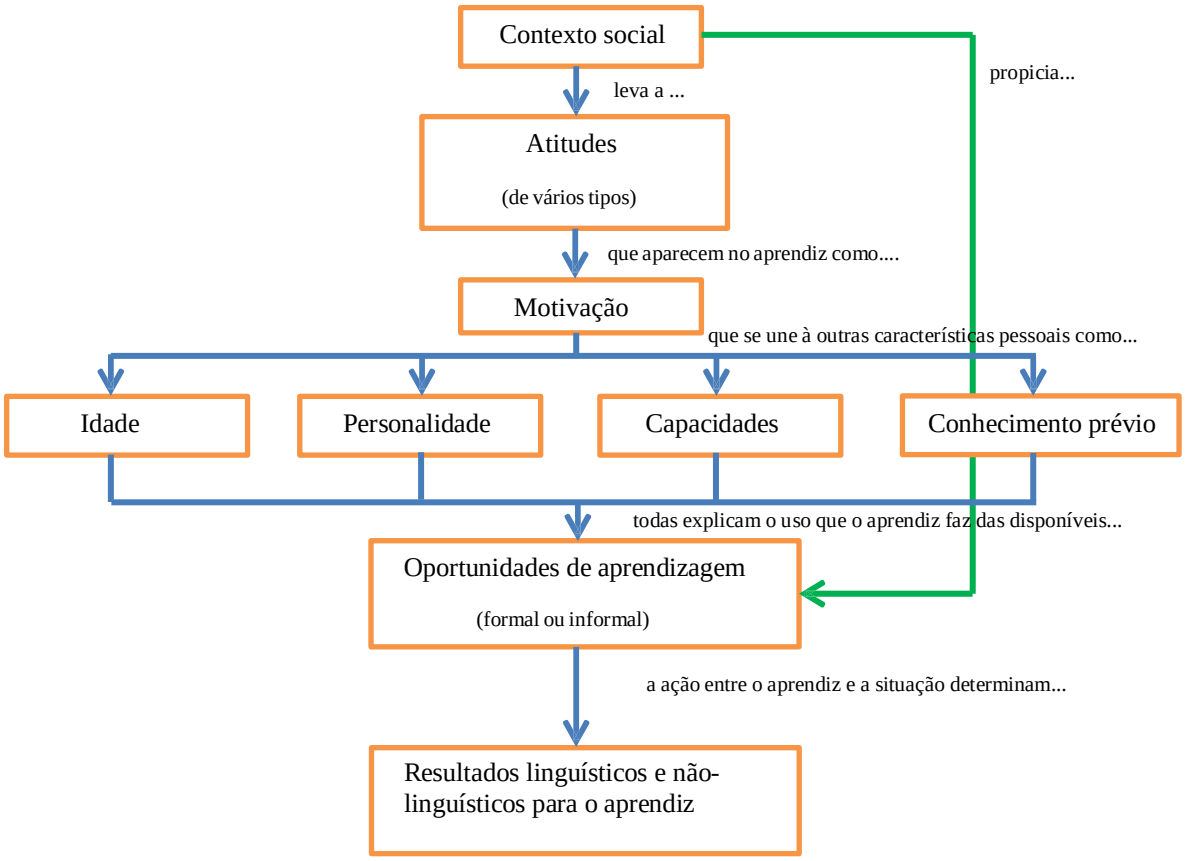

Figura 3- Modelo geral de Spolsky de aprendizagem de uma segunda língua. Fonte: SPOLSKY, 1989, p. 28.

O que Spolsky parece sugerir é que a aprendizagem de uma L2 não seria como uma fórmula mágica que teria um resultado concreto e satisfatório a ser obtido. Spolsky demonstra em sua Figura um caminho mais curto e outro mais longo a ser percorrido pelo aprendiz: ambos partindo do contexto social e desembocando em oportunidades de aprendizagem. Entretanto, em nenhum momento o autor garante que a aprendizagem irá ocorrer pois o mesmo faz uso da expressão "oportunidades de aprendizagem". O que Spolsky indica contraria a promessa de muitas instituições de ensino que tentam vender a todo custo: um método mágico para se aprender uma L2.

Segundo Spolsky (1989) seria preciso ainda abordar uma gama de tópicos, todos ilustrados na Figura 3, a fim de se discorrer sobre o processo de aquisição de uma L2. Por questões de relevância, irei ater-me aos detalhes que concernem à diferença entre aprender e adquirir uma segunda língua.

Acredita-se que aprender a L2 seria uma habilidade com suas particularidades. Embora hoje em dia muito se use o termo "aquisição de segunda língua", estudos da década de 1970 usavam a nomenclatura "aprendizagem de segunda língua" (GARDNER e

LAMBERT, 1972 e COOK, LONG e McDONOUGH, 1979) sendo que as duas conceituações parecem conter visões opostas que tentam definir o que significa poder falar, escrever e entender uma L2.

Aprender uma L2 não significa simplesmente repetir enunciados mas fazer escolhas linguísticas em situações específicas. Já em 1976, Wilkins afirmava que o aprendiz precisa ser posto em uma situação que o obrigue a fazer escolhas. Sendo assim, o uso controlado da língua em sala de aula, em resposta a situações como *repetition drills* não indicariam que o aluno, de fato, teria aprendido a língua alvo, mas essa técnica seria apenas um ensaio ou uma prática controlada em que o professor poderia monitorar padrões de pronúncia, entonação, ritmo e acuracidade gramatical.

Com essa afirmativa há quase quarenta anos, Wilkins sugeria que aprender não significa apenas memorizar expressões, mas utilizar a língua em situações de livre escolha motivadas por fatores intrínsecos, por aquilo que o aluno deseja aprender. Na verdade, o que Wilkins parecia estar dizendo é que o conteúdo a ser ensinado precisa ser significativo para que o aprendiz se sinta motivado a aprender. Apesar de antigo, esse parece ser um ponto de partida fundamental para esta dissertação, que se propõe a verificar o quanto foi aprendido em relação à quantidade de itens ensinados. Veremos claramente na seção de análise de dados que não basta ensinar coletivamente para que um item seja aprendido coletivamente. Ou seja, o aprendizado não se da em nível coletivo, mas em nível individual e esse precisa de motivação.

A aprendizagem é um processo que estaria estreitamente ligado a fatores motivacionais. Se o aprendiz sente necessidade de comunicar algo em um contexto real, ele poderá encontrar a motivação necessária para aprender o que precisa.

Krashen, linguista que aperfeiçoou as teorias de Chomsky acerca da gramática universal, foi um dos autores que estabeleceu as diferenças entre aprender e adquirir uma segunda língua. Segundo Krashen (1982, p. 10), aprendizagem seria o conhecimento consciente de uma segunda língua, o que incluiria saber regras, estar ciente delas e ser capaz de falar sobre elas. Entretanto, o autor reitera que é preciso ter em mente que grande parte das atividades conduzidas em sala de aula são atividades que, de certa forma, imitam situações do mundo real. Até mesmo situações como *role play*, discussões ou debates, são forjadas, no sentido que não partem do desejo do aluno de aprender um determinado item lexical para se comunicar em uma situação específica. Por isso, abordagens de ensino como o *communicative approach* sugerem como um dos últimos estágios da aula de língua estrangeira um passo chamado “personalização”.

A personalização seria o momento da aula em que o aluno une todo o conhecimento linguístico praticado de forma controlada e o utiliza de forma livre para falar de situações pessoais (HARMER, 2007, p. 53). Com isso, o professor criaria um ambiente em que o aluno possa ver sentido no uso de um determinado item da língua em um contexto real e pode vir a considerar significativo aprendê-lo.

Wilkins (1976, p. 39) já defendia essa posição ao dizer que um dos fatores que promove a aprendizagem de uma língua seria quão significativa é a língua a ser aprendida. Mais uma vez, pode-se sugerir que se determinado item não tiver importância para o aprendiz, existe uma possibilidade de que o mesmo não irá aprendê-lo.

O autor também afirmaria, em 1976, que é comum muitos professores acreditarem que algumas pessoas possuem uma aptidão linguística maior que outras. Isto significaria que alguns aprendizes teriam um potencial de aprendizagem e produção de línguas maior que outros. Aliado a isso, Wilkins postulava que a aprendizagem é cíclica e o que foi aprendido precisa ser constantemente praticado. Se o aprendiz não faz uso da língua ou não utiliza determinado item léxico-gramatical, existe uma grande possibilidade de que o que fora aprendido venha a cair em desuso e então seja esquecido (WILKINS, 1976, p. 41-42).

Em contrapartida, para Krashen, todos adquirimos uma língua da mesma maneira, através da exposição a um *input* compreensível, ou seja, a aquisição não se dá sem que o aprendiz seja exposto ao idioma, o que pode se dar de forma autêntica ou restrita. A exposição autêntica acontece quando o aprendiz é exposto à língua de forma natural; já a forma restrita se dá quando a língua é simplificada e selecionada para se chamar a atenção do aprendiz a um item léxico-gramatical que se deseja transmitir. Através desse mecanismo externo ao aprendiz, o mesmo irá ativar seu mecanismo de processamento cognitivo a fim de perceber padrões na língua e com isso inferir significados e regras.

Krashen (s/data) cita o caso de uma aluna japonesa de quatro anos a quem teria ensinado a falar inglês. O início da produção linguística da menina se assemelhava em muito à aquisição em língua materna, à medida que a menina produzia uma palavra ou outra isoladamente, depois duas palavras juntas. Após alguns meses, as frases foram ficando maiores.

Outra característica mencionada pelo autor foi a de que a língua surgira de forma muito rápida, ao ponto que após cinco meses convivendo com outras crianças falante-nativas de inglês ela conseguiu voltar ao Japão falando quase como uma delas. Para o linguista, isso poderia ser explicado pelo fato de que ao longo daqueles cinco meses a menina fora exposta ao *input* compreensível exaustivamente, mesmo não falando.

Logo, para Krashen, falar não é o inicio do processo de aquisição em uma segunda língua, mas sim ouvir. A exposição a uma segunda língua seria o fator-chave para o desenvolvimento da habilidade linguística.

Com isso, o que Krashen parece sugerir é que habilidades receptivas como compreensão auditiva e leitura servem de *input* para aquisição de uma língua e que o aprendiz deve receber *input* dessas habilidades antes que se espere que produza a L2 imediatamente. Segundo o autor, um falante nativo pode saber fazer uso de estruturas de um tempo verbal como o *Present Perfect* muito bem, conseguindo fazer uso do tempo verbal com precisão e nas situações corretas, por ter adquirido a língua. Entretanto, não saberia, se indagado, explicar a estrutura usada para formá-lo ou as regras de como usá-lo. Por outro lado, um falante não-nativo, que aprendeu a L2 pode saber muito de um tempo verbal e sua estrutura mas não usá-la adequadamente em contextos específicos. A esse último falta a percepção da noção de uso em contextos apropriados. Poder-se-ia dizer, com isso, que esse falante teria aprendido a L2 mas não a teria adquirido.

Krashen (1982, p. 11) atenta para o fato que a distinção aprendizagem-aquisição não se dá somente no contexto de aquisição da L2, pois certamente “aprendemos” pequenas partes de nossa língua materna na escola. Para o autor, embora a aquisição da primeira língua possa ter ocorrido em casa, informalmente, mais tarde iremos “revisitar” a língua materna, agora no contexto escolar e estudá-la a fim de aprendermos, refletirmos e analisarmos regras em um contexto formal.

Krashen também menciona fatores externos como sendo vitais no processo de aquisição de uma segunda língua. Para o linguista, a motivação é uma palavra-chave nesse cenário, já que alunos motivados aprendem melhor. Autoestima alta também parece afetar a aquisição de uma segunda língua, já que os alunos irão produzir com mais confiança. Por fim, o teórico menciona a questão da ansiedade: quanto mais baixa a ansiedade, melhor. Para a aquisição ocorrer a ansiedade deve ser zero, como em uma conversa entre amigos.

Ellis parece concordar com as ideias de Krashen ao afirmar que as crianças ao aprenderem a língua materna recebem uma quantidade muito grande de *input* durante o chamado “período de silêncio”. Nesse estágio da aquisição da linguagem as crianças aprendem muito apenas ouvindo, pois essa fase serve como uma espécie de preparação para a próxima, que seria a produção linguística (ELLIS, 2000, p. 20).

Ellis afirma que os aprendizes também trazem para o processo de aprendizagem de uma segunda língua uma grande quantidade de conhecimento prévio de mundo que os auxiliaria na tarefa de aprendizagem de uma L2. Além disso, os mesmos aprendizes também

fazem uso da língua materna como um recurso adicional. Na verdade, os aprendizes podem vir a fazer uso da língua materna como uma espécie de muleta para dar os primeiros passos rumo à língua alvo.

Ellis (2000) menciona as estratégias de comunicação que todo falante de uma língua possui. Para o autor, o falante é dotado de criatividade para explicar expressões e léxico que não conhece, parafraseando definições. Por fim, o autor define o conceito de aptidão. Aptidão parece explicar porque alguns aprendizes parecem ter mais facilidade em aprender uma língua. Isso pode vir a ser explicado devido a uma aptidão inata de cada falante em assimilar certas habilidades da língua (sendo as quatro habilidades de uma língua: escrita, fala, compreensão oral, compreensão leitora) que outras. Isto talvez explique o porquê de alguns aprendizes expostos ao mesmo conteúdo e dentro de um mesmo ambiente educacional adquirirem uma maior habilidade na língua que outros. Por conseguinte, também é possível explicar o fato de alguns aprendizes serem melhores em apenas algumas das quatro habilidades da língua e não todas. Como por exemplo, alguns aprendizes têm mais facilidade com a leitura que a fala, outros com a escrita e não com a compreensão auditiva.

Corroborando a definição do que seja adquirir uma L2, Ellis (2000) atenta para o fato de que a memorização de certos blocos lexicais usados por aprendizes iniciantes para executar determinadas funções pode dar a falsa impressão de que aquele aprendiz já é fluente na língua e tem sua competência linguística estabelecida. Para que se ateste tal fato, é preciso que se avalie o falante em um espectro mais amplo além de funções que exijam do falante a produção de blocos léxico-gramaticais, pois o próximo passo no processo de aquisição da linguagem é a omissão de determinadas palavras de uma frase. Os falantes em estágios iniciais de aquisição não fazem frases completas, mas têm a tendência de omitir algumas palavras, assim como as crianças que aprendem a língua materna.

Em suma, a aprendizagem estaria ligada a um contexto formal de educação, seria um processo consciente, no qual o aluno estuda uma determinada estrutura para utilizá-la, envolveria memorização, técnica, conhecer as regras de uma língua e como fazer uso das mesmas. Entretanto, parece que aprender não significa necessariamente adquirir pois esse é um processo inconsciente, que não precisa ocorrer em um ambiente educacional. Adquirir significaria fazer uso da língua, internalizá-la, sem obrigatoriamente ter de conhecer a terminologia e as regras associadas a uma linguagem em si, ou seja, a metalinguagem.

Portanto, neste estudo, faz-se a tentativa de investigar se, quando expostos a um ambiente educacional que proporcione a aprendizagem de blocos léxico-gramaticais comumente utilizados por falantes nativos, os aprendizes, de fato, assimilaram a língua e a

utilizam naturalmente ou não. É claro que a situação de avaliação não é uma situação ideal para medir se algo foi ou não foi assimilado. A aquisição de L2 é materializada através de sinais que Ellis caracterizou em quatro pontos nodais para se mensurar se, de fato, um aprendiz assimilou uma língua ou não (ELLIS, 2000).

O autor menciona o fato de o aprendiz ser capaz de produzir padrões de uso da língua mais ou menos de acordo com os produzidos por um falante nativo. No entanto, essa questão deve ser analisada com sensibilidade, pois muitas vezes o falante de uma L2 sabe muito mais do que realmente produz.

Outra questão a ser levantada seria o uso de blocos lexicais na fala do aprendiz. Hyland (2008) afirma em seus estudos que o uso de blocos lexicais por aprendizes de uma segunda língua pode dizer muito sobre o nível de proficiência do mesmo e que sua ausência pode vir a indicar baixa proficiência na língua. Contudo, e talvez porque o estudo de Ellis date do ano 2000, o autor parece não priorizar o uso holístico de expressões. Ao atentar para o fato de que se um aprendiz memoriza um bloco como "*Can I...?*" para pedir permissão, talvez o aprendiz não tenha de fato internalizado os diferentes usos do verbo modal.

Um terceiro ponto a ser discutido seria o sobreuso ou uso excessivo. Muitos aprendizes ao assimilarem certas estruturas tendem a fazer uso dos mesmos blocos léxico-gramaticais em inúmeras situações, por muitas vezes errôneas e com isto generalizando a estrutura linguística e a usando em contextos sociais inapropriados.

Além desses três pontos, Ellis atenta para o fato de que os aprendizes devem estar engajados no aprendizado do item e do sistema. O autor entende por item a internalização de blocos da língua e por sistema, as regras de uso em contexto e função. Ou seja, não basta ao aprendiz apenas internalizar os blocos lexicais de uma língua, mas saber utilizá-los em um contexto apropriado e de acordo com a função que tal bloco léxico-gramatical possui, como por exemplo, fazer um pedido, pedir permissão, dar uma ordem, entre outros.

Parece, então, que a aquisição de L2 é perpassada pelos princípios mencionados acima. O fato é que não existe uma fórmula mágica para a aquisição ocorrer e o próprio Ellis (2000) ressalta que adversidades externas podem vir a afetar a aquisição, como por exemplo, pressões de tempo, motivação, idade, entre outros.

Em se tratando da língua materna os estudos mencionados nesta dissertação indicam que primeiro adquirimos uma língua e depois a aprendemos, no sentido escolar. Já o mesmo não pode ser dito sobre a L2. Krashen nos anos 1970 afirmava que o sucesso em uma segunda língua ocorre devido à aquisição e não à aprendizagem (THORNBURY, 2008, p. 19). Muitos aprendizes de L2 estudam a língua em um contexto educacional formal de sala de aula e

aprendem, em sua maioria, primeiro as regras. São muitas vezes negligenciadas noções como o uso, contexto e registro. Ou seja, aprende-se a segunda língua, mas não há aquisição. Por fim, o autor afirma que conhecimento aprendido não poderá se tornar conhecimento adquirido.

Os estudos mais recentes fazem uso do termo aquisição de L2, não fazendo distinção se a língua que o aprendiz está adquirindo é de fato, segunda, terceira ou quarta língua, ou se a mesma está sendo aprendida em um contexto de língua estrangeira ou de uma segunda língua nativa.

2.2 **Interlíngua**

Para se compreender a aquisição de uma segunda língua é importante mencionar um ponto crucial nesse processo: a interlíngua do aprendiz. Interlíngua é um conceito proposto por Selinker (1972) sobre o processo de aquisição de L2. Muitos estudos sobre o assunto surgiram na área de aquisição de segunda língua na década de 1970. Segundo Schutz (2005), a interlíngua é a linguagem produzida a partir do início do aprendizado, caracterizada pela interferência da língua materna, até o aluno ter alcançado seu potencial máximo de aprendizado na língua estrangeira. Nas palavras de Ellis (2000), dentro do processo de aquisição de L2, existe um sistema linguístico abstrato com suas próprias características, erros e estilos que cada aprendiz ou grupo de aprendizes carrega, ou seja, a interlíngua.

Para Selinker, a interlíngua pode ser definida como um sistema linguístico que difere tanto da língua materna quanto da língua estrangeira. Esse sistema linguístico seria dotado de regras linguísticas abstratas transitórias. Isto quer dizer, que o falante é dotado de uma "gramática mental" e pode modificá-la sempre que desejar adicionando ou apagando regras e reestruturando todo o sistema (ELLIS, 2000; SPRAT, PULVERNESS e WILLIAMS, 2005, p.44).

Sendo a interlíngua um processo contínuo em desenvolvimento, poder-se-ia pensar que o ponto de partida do aprendiz seria a sua língua materna. Entretanto, Ellis (2000, p. 54) diz que o ponto de partida não é a L1 do aprendiz e que o mesmo não avança substituindo regras da L1 por regras da L2 mas construindo suas próprias regras internas. Contudo, isso não significa que o aprendiz de uma L2 não fará uso de sua L1 ao longo do processo de aquisição da L2, mas apenas quando acreditar que a L1 o auxiliará na aquisição da L2.

Selinker afirma que somente cerca de 5% dos aprendizes irão construir uma gramática mental semelhante à de um falante nativo. O falante nativo possui uma criatividade linguística que foi adquirida junto ao processo de aquisição da língua materna. Isso significa que ele consegue produzir e utilizar frases em situações que o aprendiz ainda não domina. Ou seja, o aprendiz pode vir a cometer erros situacionais, de inadequação de contexto ou de registro. O objetivo principal do aprendiz de uma língua estrangeira é adquirir essa criatividade de uso da língua em contextos apropriados e o domínio do sistema linguístico do falante nativo (ELLIS, 2000, p. 3).

Pode-se dizer, segundo Ellis, que os erros são característicos do sistema de uma interlíngua e esses algumas vezes poderiam ser decorrentes da influência da língua materna. No entanto, os erros provenientes da L1 do aprendiz são normalmente “temporários”, e deveriam cessar assim que o aprendiz alcançasse o nível de proficiência desejado na língua alvo.

Essa interlíngua, por ainda estar em desenvolvimento, carrega erros em si e conforme mencionado anteriormente, erros fazem parte do processo de aprendizagem de uma segunda língua. Consequentemente, é preciso que se investigue, conforme já questionava Wilkins (1976, p. 37) o porquê de os erros serem evitados a qualquer custo.

Spratt, Pulverness e Williams (2005, p. 44) indicam que os erros geralmente mostram que o falante está aprendendo, que seu processo mental interno está trabalhando e que o aprendiz está experimentando com a língua. Os autores comparam a aprendizagem de uma língua a um quebra-cabeça a ser montado, no qual o aprendiz vai fazendo tentativas com a língua, muitas vezes errôneas, mas que o mesmo está pensando e construindo suas regras mentais até conseguir encaixar as peças da língua.

À época de muitos estudos sobre a interlíngua do aprendiz, na década de 1970, os erros eram analisados de forma isolada e a língua materna era vista como interferência negativa e algo a ser evitado. A interlíngua era vista como uma interferência da língua materna e, consequentemente, a L1 deveria ser evitada e abolida do processo de ensino-aprendizagem da L2.

É inegável que a L1 do aprendiz exerça papel fundamental nos estudos de aquisição da L2, pois a L1 é um recurso de que todo aprendiz de L2 lança mão, consciente ou inconscientemente. O que se questiona, na verdade, é o fato de que praticantes da Análise Contrastiva entendam a influência da L1 de forma negativa. O uso da L1 pode ser concebido como uma contribuição positiva para a aquisição de uma L2 na medida em que a proficiência do aprendiz na língua alvo aumenta. Espera-se que haja menos interferência da L1 no

processo de aprender uma L2. Esse não é sempre o caso como veremos mais adiante no capítulo de análise.

De acordo com Granger (2002), conhecer a interlíngua de um aluno pode vir a auxiliar o professor a compreender melhor o processo de aquisição da L2, que é diacrônico, mutável e flexível e com isso ajuda-lo a repensar suas práticas docentes a fim de otimizar a aprendizagem do idioma.

3 **AVALIANDO A PRODUÇÃO DO APRENDIZ: O ERRO**

Esta dissertação discute, através de dados empíricos, aspectos léxico-gramaticais da produção escrita de aprendizes de inglês como L2. Ensinou-se a esses aprendizes a produzir um conjunto de quadrigramas associados a determinados gêneros escritos. A dissertação verifica se os alunos produziram, em condições de exame, os conjuntos ensinados, e mais especificamente quantos dos quadrigramas foram produzidos e em que tipos de texto.

O trabalho usa como metodologia a extração de quadrigramas a partir de um *corpus* digitalizado de redações de aprendizes de inglês como L2, com o auxílio de um programa de computador. Assume-se que poderão vir a ser encontrados quadrigramas errôneos, isto é, que um número de alunos possa vir a produzir com alguma frequência um número de quadrigramas que não possam ser considerados 'bem formados'. Assume-se também que a distribuição dos quadrigramas poderá ser diferente tanto de aprendiz para aprendiz, de gênero para gênero, como entre filiais do mesmo curso livre de inglês. É de suma importância, portanto, que se revisite a conceituação de erro para incorporar uma nova forma de entendimento do erro, que é aquela auxiliada por métodos computacionais, a qual lida com a frequência e abrangência.

Vimos no capítulo anterior que muitos estudos sobre erros baseavam-se em uma possível busca da influência da língua materna no processo de aquisição de uma segunda língua; outros estudos veem o erro pelo viés da interlíngua e a interferência da língua materna no desenvolvimento da linguagem do aprendiz.

Nesta seção será enfocada a tradição pré-análise de *corpus* no que tange aos estudos de erros. À época pré-análise digital de *corpora*, os erros eram vistos de forma negativa e de maneira isolada, devendo ser evitados a qualquer custo. Faço esta análise pré-*corpus* pois minha intenção é contrastar o que se fazia antes com o que se faz agora com o auxílio de programas de computador e preceitos da LC, conforme será discutido no capítulo seguinte.

Traçarei um panorama do conceito de erro. Em seguida, discorro sobre algumas definições de erro, pois, de acordo com a literatura da área, esse parece ser um conceito de difícil definição. Mais adiante, discuto os tipos de erros existentes na produção do aprendiz de L2.

3.1 **Breve perspectiva histórica do erro**

Antes do surgimento de grande parte dos estudos sobre aquisição de segunda língua nos anos 1970, a teoria behaviorista, dominante no ensino de língua estrangeira, pregava que os erros deveriam ser evitados a qualquer custo durante o processo de aprendizagem. Seus defensores (SKINNER, 1974) afirmavam que os aprendizes deveriam receber um estímulo para então produzir uma resposta que deveria ser reforçada. Acreditava-se à época que os erros sofriam interferência da língua materna.

Para os behavioristas, a língua era aprendida através da aquisição de hábitos; os maus hábitos, ou seja, erros deveriam ser evitados a qualquer custo. Talvez, então, a única forma de saná-los seria através da repetição da forma correta exaustivamente. No entanto, isto poderia vir a gerar uma super correção do erro, ou seja, quando o professor corrige muito o aprendiz tentando evitar ao máximo que o mesmo cometa erros, ocasionando assim desmotivação no aprendiz (FERRIS, 2002). Como assumia-se que os erros originavam-se na língua materna do aprendiz, surgiram muitos estudos em um campo da Linguística Aplicada, que receberam o nome de Análise Contrastiva, que visava comparar a língua do aprendiz com a língua alvo (ERDOGAN, 2005, p. 262). Acreditava-se que a comparação entre a L1 do aprendiz e a L2 auxiliaria a entender os erros mais comuns que determinados grupos de falantes poderiam cometer e com isso evitá-los para que os mesmos não interferissem na aprendizagem da língua.

No início dos anos 1970 outras teorias de aquisição de L2 começaram a surgir e viu-se, então, que o behaviorismo não dava conta de explicar os fenômenos de aquisição de uma segunda língua baseado em princípios como estímulo-reposta. Mais tarde, com o desenvolvimento da Linguística Aplicada houve uma proliferação de estudos no ramo da Análise de Erros. De acordo com Erdogan (2005, p. 263) essa área lida com a ‘performance’ dos aprendizes em termos dos processos cognitivos usados ao reconhecer ou codificar o *input* recebido na língua alvo.

Se antes a língua materna era vista como interferência, o que pode ter uma denotação negativa, nos estudos atuais ela é tratada como influência e pode vir a ter um caráter mais neutro ou até mesmo positivo. A transferência não sendo negativa é vista como um processo cognitivo de que o aprendiz faz uso para assimilar o conteúdo linguístico da L2.

3.1.1 O que é o erro?

É lugar-comum dizer que todos os aprendizes de uma língua cometem erros. Crianças cometem ‘erros’ quando aprendem sua língua materna, como também os aprendizes de L2 em termos de aprendizagem de escrita da L2, Ferris (2002, p.3) relata que as aulas de composição escrita em L2 tradicionalmente sofreram a influência da psicologia comportamental e da linguística estruturalista que consistiam basicamente em atividades controladas ou guiadas. Devido a essa influência behaviorista, grande ênfase era dada à acuracidade dos textos dos aprendizes de L2 fazendo com que os professores corrigissem todos e quaisquer erros tentando evitar que um mau hábito se formasse.

Mais tarde, no entanto, ilustra o fato de que nos anos 1970, ao invés de se focar nas formas corretas das composições, parágrafos e frases, alunos e professores se concentravam no desenvolvimento de ideias, rascunhos, revisão, trabalho colaborativo e compartilhamento de sucessos (ELLIS, 2000, p. 3, 4). Essa técnica era chamada de abordagem por processo (*process-approach*) e acreditava que se os alunos escolhessem os tópicos e recebessem direcionamento certo para tomarem as decisões a fim de moldar e polir seus próprios textos, seu produto final iria melhorar como uma consequência natural de um processo de escrita mais consciente.

Junto ao advento do *process-approach* nos anos 1970 e 1980, surgiram estudiosos como Truscott (*apud* FERRIS, 2002) afirmando que não havia evidência científica de que os aprendizes se beneficiavam com a correção de erros. Toda essa tendência levou a uma diminuição na correção dos erros de aprendizes.

No entanto, sabe-se que a maioria dos professores de língua estrangeira e também alunos acredita no potencial da correção de erros e nos efeitos positivos que tal correção trará ao desempenho geral do aprendiz (FERRIS, 2002).

Hoje em dia, parece haver maior ênfase na correção de erros pelos professores, mas esses tendem muitas vezes a marcar as formas errôneas e até mesmo corrigir os erros de seus alunos, reescrevendo-os.

Embora o erro possa ser identificado pelos professores, sua definição ainda parece depender da origem do erro – essas são diversas. Erros podem ser causados pela influência da língua materna, nesse caso usa-se a terminologia ‘interferência da L1’ para caracterizá-los. Mas acredita-se que a língua materna também pode vir a auxiliar o processo de aquisição da L2 (ELLIS, 2000, p.51).

Ferris (2002) afirma ainda que a aquisição de uma segunda língua leva tempo e ocorre em estágios. Alguns aprendizes podem nunca atingir o nível de controle de um falante nativo no aprendizado da L2, especialmente se a exposição à L2 for tardia. É também irreal esperar que a produção de aprendizes de escrita de L2 seja livre de erros. Tal fato pode ser explicado porque há evidencias de que somente cerca de 5% dos aprendizes irão construir uma gramática mental semelhante à de um falante nativo, conforme afirmava Selinker (1972).

Se para os defensores do behaviorismo os erros deveriam ser evitados, hoje em dia com métodos mais recentes e comunicativos, os mesmos são entendidos como parte natural da aquisição da língua. Erros são inevitáveis e parte necessária do processo de aprendizagem de uma língua (WILKINS, 1976).

Erros podem vir a indicar uma falha no processo de aprendizagem a partir do momento em que o aprendiz não consegue identificar a forma correta. Vemos então que o conceito de erro no ensino de uma língua estrangeira parece um pouco difícil de ser definido. Primeiro porque alguns autores classificam os erros de acordo com determinados princípios, diferenciando-os de acordo com o nível de consciência do aprendiz ao produzir a forma errônea.

Há autores como Sprat, Pulverness e Williams (2005, p. 44) que subdividem os erros em: falhas (*errors*) e deslizes (*slips*). Segundo essa visão, as falhas ocorreriam quando os aprendizes tentam dizer algo que está além do seu atual nível de processamento linguístico. Já os deslizes são resultado de fatores como cansaço, preocupação ou outras emoções ou circunstâncias temporárias.

Outra distinção feita por esses autores seria na habilidade que o aprendiz tem em conseguir corrigir ou não seu erro. A falha não pode ser corrigida porque o aprendiz não entende o que está errado, enquanto que o deslize pode ser corrigido pelo aprendiz pois este consegue compreender onde errou.

Erdogan (2005) é um outro autor que sub-classifica os erros em: falhas (*errors*) e equívocos (*mistakes*). Para o autor um erro seria o uso de um item linguístico de uma maneira tal que um falante nativo ou fluente da língua alvo o consideraria fruto de uma aprendizagem errada ou incompleta. Em outras palavras, o autor conclui dizendo que os erros acontecem porque o aprendiz não sabe o que está correto e, portanto não pode corrigí-lo.

Para Erdogan, equívoco seria o erro causado por falta de atenção, cansaço, descuido ou algum outro aspecto da ‘performance’ do aprendiz. Equívocos podem ser autocorrigidos quando chama-se a atenção do aprendiz para os mesmos. Se o aprendiz alterna entre formas corretas e errôneas, Ellis (2000) afirma que se trata de um equívoco, mas se o falante sempre

utiliza a forma errônea então trata-se de uma falha. Ferris (2002, p. 25) sugere que erro (ou falha) é a falta de acuracidade na língua-alvo, tanto por interferência da L1 ou pela não aquisição completa da L2.

Pode-se concluir desses estudos que temos dois tipos de erros: as falhas e os equívocos ou deslizes. No entanto, eu adicionaria um terceiro tipo que seria o erro de tentativa. Esse erro originar-se-ia devido ao fato de o aprendiz ainda estar manipulando a língua muito precocemente. Dessa forma, ainda não teria recebido *input* suficiente da língua, cometendo erros de colocações, contextuais, de registro, até mesmo erros léxico-gramaticais, mas todos originados pelas tentativas que se dão durante a aquisição da língua.

3.1.2 Tipos de erros

As tendências mais recentes de métodos comunicativos focam principalmente em desenvolver a habilidade oral do aprendiz, com os erros tendo menor importância, contanto que os mesmos não impeçam a comunicação e sejam produzidos em atividades cuja ênfase seja a fluência, os erros podem vir a ser negligenciados (UR, 1996). Se por um lado, temos no passado, uma abordagem considerando o erro um demônio a ser exorcizado, por outro, no presente, o mesmo já é visto como um eterno e inseparável companheiro (LÓPEZ, 2009). A verdade é que um estudo de erros oferece grandes vantagens para melhorar as práticas pedagógicas da língua. Isso significa que os resultados de uma possível analise baseada em erros evidenciam aquelas áreas da língua que os professores precisam dar ênfase, tais como gramática, vocabulário, discurso, etc. Ou seja, analisando os erros importantes sugestões para o desenvolvimento de métodos de ensino podem vir a ser criados, e isso envolve todas as áreas do planejamento pedagógico, desde o programa curricular até os materiais de ensino.

No entanto, a avaliação da competência de um aprendiz vai além da análise de erros analisados isoladamente, mas também em outros aspectos tais como o não uso de estruturas difíceis da língua, o que poderia ser um indicador de pouco progresso na aprendizagem de uma língua estrangeira (YIP, 1995).

A atitude em relação a erro sempre variou em um contínuo de total rejeição ao erro a uma certa leniência. No Quadro abaixo (1), identificamos quatro momentos dentro da LA com autores e suas visões de erros.

Autor	Ano	Tipos de erro
Corder	1973	Omissão, adição, seleção e ordem.
Wilkins	1976	Generalização da língua.
Ferris	2002	Global, local, tratável e não-tratável.
Thornbury	2008	Lexical, gramatical e discursivo. Transferência positiva e negativa.

Quadro 1- classificação dos tipos de erros segundo alguns autores.

Em 1973 Corder (*apud* ERDOGAN, 2005, p. 264) classificava os erros em quatro tipos: omissão, adição, seleção e ordem. Os erros de omissão seriam quando o aprendiz omitia uma palavra ou um morfema. Os de adição quando o aprendiz adicionava um vocábulo desnecessário à frase ocasionando um erro. Os erros de seleção seriam quando há uma escolha verbal errônea, um pronome errado ou verbo, ou seja, o aprendiz teria feito uma escolha errada. E por último, o erro de ordem, que poderia ser um erro de ortografia, ordem errada no eixo sintagmático da frase ou ordem de adjetivo e substantivo, por exemplo. Em outras palavras, Corder via o erro em L2 de uma forma isolada ignorando o sentido e o discurso.

De forma semelhante, Wilkins (1976) via o erro em L2 dissociado do sentido: assim que os aprendizes tivessem formado regras mentais sobre um fato linguístico, estenderiam essas regras mentais a outros fatos, generalizando e por vezes errando. A terminação do passado simples em inglês estendida ao passado irregular é um típico exemplo de generalização.

Ferris (2002) faz a distinção entre erros globais e locais e também entre erros tratáveis e não-tratáveis. Os erros globais seriam aqueles que afetam o entendimento de um texto, impedindo a comunicação e a compreensão da mensagem e erros locais seriam pequenos erros que não afetam a compreensão do texto, a mensagem pode ser perfeitamente entendida.

Erros tratáveis são aqueles relacionados a uma estrutura linguística que está relacionada à quebra de uma regra. São chamados de tratáveis porque o aprendiz pode ser direcionado a um livro de gramática e por meio de um conjunto de regras pode resolver o problema. Como exemplo poderíamos citar os: tempos e formas verbais, o uso de artigo, erros de ortografia e pontuação, etc.

Um erro não-tratável pode ser idiossincrático e o aluno precisará utilizar conhecimento adquirido da língua para fazer a autocorreção. Como exemplo poderíamos citar: a escolha errada de palavras, uma estrutura frasal inexistente na língua, seja por adição ou subtração de palavras.

Ferris faz sua contribuição aos estudos sobre o erro, mas a grande questão de seu trabalho é quando a autora fala dos erros tratáveis e não-tratáveis. Entretanto, esta dissertação não esta interessada nos erros individuais, pois estes não são tratáveis por serem idiossincráticos. O presente estudo não lida com erros individuais, mas sim com os coletivos.

Thornbury (2008, p. 114) diz que existem duas razões principais pelas quais os erros podem ocorrer. A primeira delas seria pela influência da língua materna do aprendiz, também chamada de interferência ou transferência e dessa forma o autor fala de dois tipos de transferência: positiva e negativa. Para o autor, ambas sofrem influência da língua materna do aprendiz. No entanto, a transferência positiva não resulta em um erro em si.

O autor afirma que há três tipos de erros que o aprendiz de uma segunda língua comete: erros lexicais, gramaticais e discursivos. Para o autor, a primeira categoria envolve os erros de escolhas de palavras errôneas pelo aprendiz para o significado que se queria transmitir. Como por exemplo, a escolha da forma errada da palavra (o uso do verbo *think* ao invés do substantivo *thing*) pela semelhança fonológica e gráfica e erros na ordem natural das palavras, como em português temos primeiro o substantivo e depois o adjetivo e em inglês o inverso (ex.: girl beautiful).

O segundo tipo de erro seria o erro gramatical. Esses estão atrelados à forma e tempo do verbo, na estrutura da frase (*the doorbell rangs; we speaked*) e também em casos onde o sujeito da frase é omitido (*it*).

O último tipo de erro listado por Thornbury seria o erro de discurso. Esses seriam erros macro, de construção e coesão textual onde palavras usadas erroneamente como um conectivo *eventually*, pode gerar no leitor a falsa impressão de que o texto está perto de sua conclusão.

Na visão do autor o erro pode ser micro ou macro. O macro seria um erro discursivo e o erro micro o erro da estrutura linguística em si.

O fato de eu não ter encontrado uma literatura específica que classifique erros nas décadas de 1980 e 1990 parece convergir com o estudo de Truscott (*apud* FERRIS, 2002) conforme já mencionado anteriormente nesta seção, onde o autor afirma que houve um declínio na correção, nos estudos e até mesmo na importância dada à análise de erros como um reflexo do surgimento do *process-approach* naquelas décadas. No entanto, o Quadro (1) sugere que parece ter havido um divisor de águas entre a era pré e pós *corpus*. Estudos pré-*corpus* envolviam erros nas décadas de 1970, mais à frente parece haver um intervalo onde os erros foram esquecidos e mais adiante no período pós-*corpus* os mesmos voltaram a estar em

voga com novas classificações e estudos nos anos 2000 já agora com o auxílio de ferramentas de programas de computador e princípios da LC.

Em 2000, Ellis (2000, p.52) notou que havia um tipo de erro que escapava a classificação clássica de erros em gramaticais ou lexicais, imperante até aquela época: erros também podem ter sua origem no uso excessivo de determinadas formas. Ellis notou que alguns aprendizes chineses usavam em excesso expressões de arrependimento ao se desculpar em inglês, e isso espelhava as regras pragmáticas de sua língua materna.

Mais recentemente com o auxílio de métodos computacionais aliados à Linguística Aplicada tais fatores tendem a incluir erros que envolvem a frequência de certas estruturas no texto, que vai desde o não uso da estrutura, até o uso excessivo, conforme proposto por Granger (2003) e que será mais bem detalhado no capítulo a seguir.

3.2 **Escrita de aprendiz e análise auxiliada por computador**

Esta dissertação discute através de dados empíricos a produção escrita de aprendizes de inglês como L2, os quais foram ensinados a produzir um conjunto de blocos léxico-gramaticais associados a determinados gêneros A dissertação verifica se os alunos produziram em condições de exame, os conjuntos ensinados, especificamente blocos léxico-gramaticais de quatro palavras (quadrigramas), quantos dos quadrigramas foram produzidos e em que tipos de texto.

O trabalho usa como metodologia a extração de quadrigramas a partir de um *corpus* digitalizado de redações de aprendizes de inglês como L2, com o auxílio de um programa de computador. Assume-se que poderão vir a ser encontrados quadrigramas errôneos, isto é, que um número de alunos possa vir a produzir com alguma frequência um número de quadrigramas que não possam ser considerados ‘bem formados’. Por isso, estendemos a discussão sobre erro para incorporar uma nova forma de entendimento do erro, que é a auxiliada por métodos computacionais. Essa nova concepção do erro em L2 lida com a frequência com a qual um bloco é ou deveria ser usado em determinado tipo de texto.

Vimos na seção anterior que muitos estudos sobre erros baseavam-se em uma possível busca da influência da língua materna no processo de aquisição de uma segunda língua, outros estudos veem o erro pelo viés da interlíngua e a interferência da língua materna no desenvolvimento da linguagem do aprendiz.

O fato é que, mais atualmente, tais conceitos convivem confortavelmente com outras visões do que vem a ser um erro. Hoje, com o advento do computador no tratamento de textos de aprendizes, a visão que se tem de erro (uso errôneo) foi somada à visão de sobreuso e subuso de determinadas expressões. Os conceitos de sobreuso, subuso e uso errôneo podem ser encontrados nas publicações de Granger (1998 et seq.), que tem focado suas pesquisas nos estudos de *corpora* de aprendiz.

> O estudo de sobreuso e sub uso marcam uma amplitude do escopo da análise de erros tradicional já que estes fenômenos, que são difíceis de ser identificados sem o uso de métodos computacionais, frequentemente não constituem erros. Na verdade, eles refletem áreas da língua do aprendiz que diferem da língua do falante nativo em termos de frequência e distribuição ao invés de correção. (GRANGER, 1998, p. 18)

Para Granger, os estudos a partir de *corpora* de aprendiz passaram da simples descrição do erro para o levantamento daquilo que é ou não nativo, em termos do que é usado ou não; do que é usado em excesso e do que é subusado. Granger (2004) afirma que a Linguística de *Corpus* abre uma perspectiva descritiva para o estudo e a análise de erros linguísticos de aprendizes de inglês como língua estrangeira já que a criação de *corpora* de aprendizes é um conceito relativamente novo, se comparado com os *corpora* de textos oriundos de falantes nativos.

Conforme Berber-Sardinha (2004, p. 266) afirma, “os conceitos de sobreuso e subuso são indicadores de desvio de um padrão de referência que, em geral, é a linguagem do falante nativo”. Para o autor (2004, p. 266), “o que muda é o foco, o método e o próprio entendimento da natureza do erro”. Se antes o erro era julgado e avaliado, hoje ele é descrito e analisado. Se antes ele deveria ser evitado, hoje, sabe-se que ele faz parte do processo de aprendizagem da L2.

Para analisar-se o erro, é vital analisá-lo dentro do seu contexto. A análise de erros hoje é bem diferente da feita nos anos 1970. Se à época, a análise era descontextualizada levando em consideração o erro de forma isolada, hoje, os erros são investigados de forma contextualizada.

O uso de ferramentas computacionais para extração e análise de erros permite que o linguista passe do erro micro, isolado e descontextualizado para uma visão macro e contextualizada e acima de tudo coletiva. Se antes, o analista tinha a visão apenas das folhas e flores de uma árvore, com o auxílio de programas de computador o mesmo consegue ver a árvore como um todo e muitas vezes parte da floresta, como é o propósito desta dissertação.

Essa proposta de contextualização do erro já era advogada por Granger (1998 et seq.), em muitos de seus estudos. A autora afirma que não se pode extrair o erro linguístico de seu texto original. Para a autora, os *corpora* de aprendizes são feitos de extensões contínuas do discurso e não de palavras ou frases isoladas. Tais extensões contêm tanto os exemplos de usos errôneos da língua quanto de usos corretos, isto é, os *corpora* de aprendizes contêm a linguagem em uso do aprendiz, seus erros e acertos, ou tentativas de acertos.

"Nem todo mundo usa a "linguagem correta" e, dessa forma, aquilo que não se encaixa no que é correto, simplesmente é considerado como errado." (ALMEIDA, 2010, p. 36). Tal conceituação tende a ver a linguagem de forma simplista. Pois é preciso que, antes de julgarmos a linguagem como correta ou errônea, levemos em conta fatores extralinguísticos atrelados à produção linguística.

3.3 **Colocação e coligação**

Dois conceitos que estão sob o viés da Linguística de *Corpus* são os de 'colocação' e 'coligação' e que podem estar estreitamente ligados à maneira de se analisar erros. É praticamente inviável nos estudos que envolvem linguagem de aprendiz e *corpora* que não se encontrem tais tipos de erros, haja vista que grande parte dos erros produzidos por aprendizes de L2 pode se originar no modo como as palavras se combinam com outras em seu entorno.

O termo 'colocação foi primeiramente usado por Firth (1968) e a ele atribui-se a definição de colocação como combinações frequentes de unidades lexicais fixadas na norma, que se destacam pela sua alta frequência de uso, como por exemplo, *redondamente enganado*, o que motiva que este tipo de construção seja às vezes identificado como sendo unidade semi-fixa ou semi-idiomática. Na verdade, a colocação seria a preferência de umas palavras por outras formando um elo inseparável. No entanto, é importante que se faça a correta distinção entre uma colocação e um n-grama, pois uma colocação pode fazer parte de um n-grama, mas um n-grama não contem necessariamente uma colocação, pois um n-grama é produto de uma extração por meios computacionais de sequências lexicais repetidas em um *corpus*.

Já por 'coligação' entende-se como a propriedade de as palavras preferirem determinadas posições sintáticas, na frase ou mesmo no texto (BERBER-SARDINHA e SHEPHERD, 2008). A coligação também envolve um bloco de palavras, mas a natureza de sua junção é gramatical (FIRTH, 1968, p.181).

A ideia básica de coligação é que assim como um item lexical pode vir a aparecer junto a outro item lexical, o mesmo também pode vir a ocorrer com uma função gramatical particular (HOEY, 2005, p.43). É como se houvesse um “gap” a ser preenchido na frase com uma função sintática em que não há muitas opções de escolhas.

Logo, o que pode se depreender dos estudos mencionados anteriormente é que a colocação está para o léxico assim como a coligação está para a gramática. Os colocados parecem, com isso, trazer sentido ao texto enquanto que os coligados teriam a função de unir ideias, de estruturar o texto.

Um aprendiz pode, portanto, juntar itens lexicais que normalmente não se justapõem ou não estão consolidados como ‘colocados’ pelas regras do uso, como por exemplo “pé-de-galo” ao invés de “pé-de-galinha”, ou pode colocar itens lexicais em posições gramaticais inusitadas, nem sempre intencionalmente, com vistas a efeitos metafóricos, como *algumas lágrimas atrás,* onde ‘lágrimas’ funciona como parte de uma circunstância de tempo.

3.4 **Blocos léxico-gramaticais**

Embora tenha tratado previamente da definição de colocados e coligados como sendo blocos léxico-gramaticais de origens distintas, ao abordar a questão da aprendizagem de uma segunda língua, é preciso que se analise não apenas a gramática por um lado e o léxico por outro. Se estudos e métodos de ensino antigos priorizavam o ensino do léxico separado da gramática, hoje em dia, percebe-se que ambos caminham juntos formando o que se pode chamar de léxico-gramática. Nesta seção, passo a explicar brevemente o que vem a ser tal conceito e a definir blocos léxico-gramaticais.

Grande importância tem o léxico nos estudos de uma língua. “Sem a gramática pouco pode ser dito, mas sem o léxico nada pode ser dito” (WILKINS, 1972, pp. 111-112). Logo, vemos nascer uma grande inquietação em relação a como a língua estava sendo ensinada, pois os métodos tradicionais focavam tanto no ensino da gramática como forma de se adquirir uma segunda língua, mas também reforçavam uma pretensa oposição entre léxico e gramática.

Com tais conceitos sendo amplamente divulgados e postos em prática vimos nascer novas abordagens no ensino de línguas como a abordagem lexical criada por Michael Lewis (1993). Este afirma que a língua é composta por blocos pré-fabricados de palavras e que a língua consiste de léxico gramaticalizado e não de gramática lexicalizada.

Logo, a língua é formada por blocos léxico-gramaticais, ou seja, palavras que andam juntas umas com as outras como se fossem, tijolos pré-fabricados, com os quais, o aprendiz vai moldando sua fala unindo tijolo a tijolo e construindo seu discurso. Esses tijolos se constituem de vários tipos e tamanhos de unidades compostas ou 'pré-padronizadas' tais como idiomas, frases fixas e colocações e por isso difíceis de delimitar e classificar (SHEPHERD, 2012).

Embora Lewis (1993) tenha chamado esses blocos de *lexical chunks* (*blocos lexicais) os blocos léxico-gramaticais já foram estudados desde a década de 1990 por diversos autores não necessariamente no ambiente de L2, também no ambiente do português como língua materna (ver SHEPHERD, ZYNGIER e VIANA, 2009). Outros estudiosos cunharam seus próprios termos para blocos semelhantes como 'padrões pré-fabricados' (GRANGER, 1998), 'lexemas frasais' (MOON, 1998), fórmulas (WRAY, 2002), entre outros. Podemos adicionar à lista o rótulo dado por Margareta Lewis (2009) em seu estudo sobre aquisição de uma segunda língua com aprendizes suecos, no qual a autora denomina os blocos de 'sequências formuláicas'.

Um estudo seminal sobre blocos léxico-gramaticais é o estudo de Sinclair (2004) no qual são analisados os "n-gramas". Isto é, blocos léxico-gramaticais extraídos por computador divididos em um número pré-determinado de itens lexicais, "n" sendo esse o número de palavras contidas no bloco. Sendo assim, um bigrama seria um bloco lexical formado por duas palavras, um trigrama por três e assim sucessivamente.

Outros estudos importantes cujo objetivo é a categorização dos blocos lexicais extraídos por vias computacionais a partir de *corpora* são os de Biber (2006) e o de Hyland (2008). Esse último, que será enfocado na parte analítica, constatou como a presença frequente de n-gramas tais como *as a result of, it should be noted that,* e *as can be seen* nos ajudam a identificar um certo texto como pertencente a um registro acadêmico, enquanto que *with regard to, in pursuance of* e *in accordance with* são frequentemente encontrados em outros registros, os textos legais.

O que Biber e Hyland estão sugerindo é o que outros autores como Granger também já estudaram, ou seja, que cada registro possui sua fraseologia característica. Em outras palavras, através da análise dos tipos de colocados e agrupamentos lexicais encontrados em um determinado tipo de texto pode-se dizer a que gênero o mesmo pertence. O gênero influencia as escolhas léxico-gramaticais (BAKHTIN, 2003, p. 312). No universo de textos produzidos por aprendizes de uma L2, espera-se que haja evidência de domínio não apenas do idioma em si, mas também da prática discursiva na qual ele está engajado. O aprendiz deve saber fazer

uso dos blocos que permeiam certos gêneros textuais, pois a ausência de determinados colocados em textos de aprendizes de língua estrangeira pode vir a sugerir uma falta de domínio do idioma ou daquela prática social (HYLAND, 2008).

3.5 **O princípio idiomático e o princípio da livre escolha**

A produção linguística proficiente não é somente calcada em blocos ou n-gramas. Essa produção se alterna entre duas escolhas que regem o uso da língua: o 'princípio da livre escolha' e o 'principio idiomático' (SINCLAIR, 1991). O primeiro está atrelado ao conceito de descrição tradicional da língua como contendo espaços a serem completados por um número grande de possibilidades de escolhas de palavras. O principio idiomático, por outro lado, vê a língua como uma série de blocos léxico-gramaticais que são unidos uns aos outros. Sinclair usa como exemplo o bloco *set eyes on* que costuma vir junto a outros blocos léxico-gramaticais como *the moment, the first time* e o verbo auxiliar *have/has* (SINCLAIR, 1991, p. 111). No entanto, há alguns conceitos que regem a natureza da língua que parecem ir de encontro ao princípio da livre escolha. O primeiro deles é que o gênero atrai um léxico específico para si (BAKHTIN, 2003) e o outro é que as palavras não aparecem juntas ou são escolhidas por acaso, mas elas se agrupam umas com as outras formando um bloco de palavras, como mencionado neste estudo, um bloco léxico-gramatical.

Sinclair sugere haver a predominância do "princípio idiomático" entre os falantes nativos. O princípio idiomático facilita tanto a produção de texto como o seu processamento. Por facilitar ambos, acredita-se que dominar o princípio idiomático confere idiomaticidade e fluência ao aprendiz. Entretanto, isso não pode ser generalizável, como mostra o estudo importante de Lewis (2009).

Lewis (2009) analisou um *corpus* de inglês oriundo de aprendizes adultos suecos que possuem considerável exposição à L2 fora da sala de aula. O objetivo do estudo era contrastar um *corpus* de falantes nativos com esse *corpus* de aprendizes de inglês a fim de avaliar a produção de sequências formuláicas, ou blocos léxico-gramaticais dos aprendizes, como forma de mensurar proficiência na língua alvo, idiomaticidade e fluência.

Lewis partiu da hipótese de que o aprendiz adulto sueco usa o princípio idiomático e não o da livre escolha na produção de textos. Para ela, o aumento da idiomaticidade e fluência

na língua são esperados conforme o aprendiz adquire sequências formulaicas e se aproxima de um certo nível de proficiência da língua alvo.

Apesar disso, a autora assumiu também que a proporção de sequências formuláicas usadas pelos aprendizes seria menor que no *corpus* de referência: os aprendizes além de usarem os blocos léxico-gramaticais aceitos e esperados, os usariam de uma forma diferente daquela dos falantes nativos. Com isso, erros de tentativa poderiam vir a ser encontrados, mas tais casos tenderiam a diminuir conforme o aluno adquirisse proficiência na língua (LEWIS, 2009, p. 6,7).

De fato, a autora constatou que alguns aprendizes cometeram erros ao tentar reproduzir certas sequências formulaicas como por exemplo *do a mistake (make), *look at TV (watch), have advantage of (take) e *accident by a car (car accident). Em relação a outros erros, o que pôde ser constatado no estudo foi que algumas sequências foram usadas em um contexto errôneo, assim como também houve erros na escolha de léxico, também foram muito comuns erros no uso de artigos e preposições.

Além disso, o estudo de Lewis mostrou, como a autora previa, que os aprendizes fizeram uso do principio idiomático. Entretanto, foi notado também que os aprendizes fizeram um maior número de repetições das mesmas sequências formuláicas que os falantes nativos, ou seja, reciclaram os blocos, o que pode ter afetado a percepção de uso desse princípio.

A conclusão do estudo de Lewis sugere que a presença pura e simples de linguagem convencionalizada, como sugeria Yorio (1989), não é necessariamente um indicativo de idiomaticidade e, portanto não pode ser uma forma de mensurar proficiência na língua. O uso repetido dos mesmos blocos, ainda que algo não errôneo e dentro do princípio idiomático, não torna o discurso mais proficiente – porque essa não é a constatação que se faz sobre a escrita de não aprendizes.

Nesse capítulo foi feito um levantamento, ainda que breve, sobre como vem sendo visto o erro na produção do aprendiz e como essa visão modificou-se e complicou-se diante das possibilidades combinatórias fornecidas pela LC. No próximo capítulo, sobre materiais e métodos da pesquisa em tela, é feita uma discussão sobre a Linguística de *Corpus* como sendo uma metodologia de pesquisa. Os processos de coleta e tratamento dos materiais de estudo são explicados bem como o método de análise empregado.

4 **MATERIAIS E MÉTODOS**

Para analisar quadrigramas em um número elevado de composições escritas por aprendizes, este estudo faz uso de análise auxiliada por computador e dos preceitos da Linguística de *Corpus.* Tradicionalmente, um capítulo de metodologia inicia-se com o passo-a-passo da pesquisa, com o detalhamento dos procedimentos de análise do estudo. Porém, neste trabalho, irei tecer alguns comentários de natureza teórica acerca da Linguística de *Corpus* como metodologia de pesquisa. Depois, justifico seu uso como método de análise de meu *corpus*. Em seguida, categorizo o que vem a ser um *corpus* de aprendizes, para mais adiante detalhar os procedimentos de compilação do *corpus* de estudo e caracterizar os sujeitos e o contexto da pesquisa. Menciono o que vem a ser um *corpus* anotado e sua utilidade. Por fim, detalho as características do *corpus* escolhido para este estudo.

4.1 **Introdução: por que usar o computador no tratamento de dados textuais empíricos?**

Este trabalho faz uso de ferramentas digitais para compilar, formatar, classificar e auxiliar na análise de dados oriundos de um grande número de textos escritos por aprendizes de inglês como língua estrangeira. Dentro da Linguística Aplicada, a disciplina que se ocupa de compilação e tratamento de dados digitalizados é a LC. Pode-se vir a questionar se a LC, é uma área da Linguística ou apenas um conjunto de métodos em si. A LC trabalha com dados empíricos, oriundos da linguagem em uso e com esses dados é possível fazer deduções sobre a língua.

Alguns linguistas consideram a LC uma metodologia e outros a veem como teoria. Em um dos primeiros volumes dedicados à LC, McEnery e Wilson (1996, p. 2) perguntam e respondem textualmente “A Linguística de *Corpus* é um ramo da Linguística? A resposta a essa pergunta é tanto ‘sim’ como ‘não’”. Os autores alegam que a LC não tem o mesmo estatuto da Semântica, Sintaxe ou Sociolinguística visto que essas disciplinas têm um objeto de investigação definido. Ao mesmo tempo, alegam que o termo *corpus* pode ser atrelado a cada uma das áreas da Linguística, gerando portanto a ‘Semântica de *corpus*’, a ‘Sintaxe de *corpus*’, por exemplo, em oposto à Semântica ou Sintaxe não baseadas em *corpora*. López (2009) afirma que se um dos princípios da pesquisa linguística é a observação e descrição de

fenômenos, então, a LC é mais do que uma metodologia, pois permite ao analista extrair dados, quantificá-los, observá-los e interpretá-los. Dessa forma, a natureza da pesquisa baseada em *corpus* não pode ser classificada nem como quantitativa e nem como qualitativa, pois já nasce híbrida em sua natureza. Sendo assim, López argumenta que a partir da extração de dados de grandes quantidades de textos através do auxílio de ferramentas de computador a pesquisa com base em preceitos da LC permite não apenas a extração desses dados, mas também dá ao linguista evidências irrefutáveis que podem ser analisadas com base no *corpus*. A LC permite que o pesquisador trabalhe com grandes quantidades de dados, sendo estudados tanto em termos de frequência quanto abrangência. Como resultado, o que deveria levar meses de trabalho através da análise manual, passa a levar segundos através do auxílio de programas de computador.

Segundo Tognini-Bonelli (2001, p. 223), ao fazer uso da LC como ferramenta de análise pode-se usar o *corpus* de duas maneiras: como sendo um princípio teórico, que viria a ser a abordagem dirigida pelo *corpus* (*corpus-driven*) e também como uma metodologia, ou seja, a baseada no *corpus* (*corpus-based*). Na primeira, o pesquisador não se atrela a nenhuma categoria linguística pré-estabelecida para analisar os dados. As hipóteses e teorias surgem a partir das evidências que emergem do texto. O analista parte do *corpus* em si, sem nenhum arcabouço teórico, pois são os dados que ditarão os rumos da pesquisa. A partir daí a pesquisa se inicia.

Na segunda abordagem, o linguista pode lançar mão de uma teoria e fazer uso de preceitos da LC para testar sua teoria, ou seja, o estudioso já parte de um pressuposto teórico e apenas utiliza o *corpus* para confirmar ou refutar sua ideia inicial.

Há ainda, segundo menciona Berber-Sardinha (2004, p. 263), um tipo de abordagem chamada de limitada ao *corpus (corpus-bound)* que utiliza o mesmo para compilação de materiais didáticos e dicionários.

Este estudo está atrelado à abordagem dirigida pelo *corpus* (*corpus-driven*), sendo a LC mais que uma metodologia de pesquisa, pois ao iniciar o trabalho não havia uma hipótese inicial a ser testada ou uma teoria para aplicar aos dados. Havia tão somente a suspeita de que os aprendizes estudados talvez não fizessem uso dos quadrigramas ensinados ao longo do curso avançado, e que o mesmo aconteceria em textos escritos em condições de avaliação.

4.2 *Corpus: corpus* de aprendiz

Palavra proveniente do latim, *corpus*, e em seu plural *corpora*, de acordo com o dicionário Michaelis, significa reunião de documentos, dados e informações sobre um assunto. Outra definição dada pelo dicionário, é que o *corpus* é um conjunto de material recolhido e bem delimitado no tempo e no espaço apto a servir para a descrição linguística.

Corpus, para esta dissertação, quer dizer *corpus* digitalizado e analisável por um programa de computador, seguindo preceitos da LC. Sinclair (2005) parece ser um dos teóricos que melhor aponta para uma definição que se atrela à visão deste estudo. Para o autor um *corpus* seria:

> Uma coleção de textos linguísticos em formato eletrônico, selecionados de acordo com critérios externos a fim de representar, ao máximo possível, uma língua ou uma variedade da língua como sendo uma fonte de dados para pesquisa linguística. (SINCLAIR, 2005, p. 16).

Berber-Sardinha (2004, p. 19) descreve quatro quesitos que caracterizariam a formação de um *corpus* analisável por ferramentas de computador: 1) o *corpus* seria formado por textos autênticos (sem o propósito de ser alvo de pesquisa linguística); 2) constituído de textos formados por falantes nativos; 3) o conteúdo do *corpus* deve ser escolhido criteriosamente e; 4) o conteúdo do *corpus* seria representativo de uma variedade linguística ou de um idioma.

O *corpus* deste estudo não preenche todas as categorias acima visto que não é constituído por falantes nativos (item 2),sendo assim, o mesmo se enquadra em outra categoria, *corpus* de aprendizes, que passo a tratar a seguir.

Um *corpus* de aprendiz seria um banco de dados de textos digitalizados da língua produzida por aprendizes de uma língua estrangeira (LEECH, 1998)[2]. Uma outra definição para um *corpus* de aprendiz seria uma quantidade de textos ou transcrições de fala de não nativos, não sendo necessariamente de alunos provenientes de um curso de línguas, podendo incluir falantes experientes e/ou que deixaram de ter aulas de língua (BERBER-SARDINHA, 2004). E uma última definição de um *corpus* de aprendiz é a de uma coleção eletrônica de dados textuais autênticos e digitais de uma segunda língua coletados de acordo com critérios explicitamente elaborados para um propósito específico na área de estudos de aquisição de

[2] O trecho original é: “A learner *corpus* is a computerized textual database of the language produced by foreign language learners.”

segunda língua/ensino de língua estrangeira. Tais dados são armazenados de uma maneira padronizada e homogênea e documentados de acordo com a sua proveniência (GRANGER, 2002).

Estando o *corpus* em questão em linha com os princípios mencionados anteriormente, ou seja, sendo produzidos por aprendizes de uma L2, elaborados de acordo com critérios e fins específicos, pode-se afirmar que o *corpus* deste estudo trata-se de um *corpus* de aprendiz.

Alguns exemplos de projetos bem sucedidos com *corpora* de aprendizes no exterior são: International Corpus of Learner English (ICLE), Cambridge Learner Corpus (CLC) e o Japanese English as a Foreign Language Learner (JEFLL). No Brasil há três projetos importantes como: o CABrI (Corpus de Aprendizes Brasileiros de Inglês) coordenado pela professora Deise Prina Dutra da UFMG, na USP a professora Stella Tagnin desenvolve o COMET (Corpus Multilíngue para Ensino e Tradução) desde 1999 e o Br-ICLE na PUC-SP coordenado pelo Professor Tony Berber-Sardinha.

Ao se analisarem blocos de palavras em *corpora* de escrita de aprendizes pode-se ter *insights* de como a aquisição da habilidade da escrita em L2 pode ocorrer. Assim também, erros provenientes da análise de *corpora* refletem os construtos internos de um aprendiz, os quais constituem um sistema linguístico independente chamado interlíngua, termo cunhado por Selinker (1972), e cuja conceituação já foi explicitada na seção 2.2 do capítulo 2. A análise dos blocos léxico-gramaticais na produção escrita de aprendizes pode ser crucial na área de aquisição de uma segunda língua por permitir a real observação da produção do aprendiz e possibilitar aos pesquisadores de aquisição de uma segunda língua explicar como a aprendizagem progride.

Embora haja estudiosos como Granger (2002) que promovam os estudos com *corpora* de aprendiz, algumas críticas são comumente levantadas no meio acadêmico acerca de tais estudos (WIDDOWSON, 2000). Entre elas estão a falta de autenticidade de um *corpus*, as limitações que o *corpus* impõe à análise linguística, já que esta será apenas um fragmento do que um falante seria capaz de produzir e também a falta de contexto, pois o *corpus* reflete uma realidade do contexto em que os textos foram coletados. Para se usar o *corpus* em sala de aula, a língua extraída precisaria ser recontextualizada, pois ela está fora de contexto original. O autor ainda afirma que o objetivo de um aprendiz de uma língua estrangeira (LE) é aspirar e se assemelhar à produção de um falante nativo, mas seria um erro afirmar que o que esta representado textualmente pelo *corpus* seria, de fato, a linguagem real e que o aprendiz deva se espelhar àquela língua.

No entanto, há de se convir que um *corpus* de aprendiz segue critérios particulares, conforme já mencionado por Berber-Sardinha (2004), que o distingue de um *corpus* tradicional. Entre eles, o fato de que os textos muitas vezes são escritos em situações irreais, como provas, avaliações com limitações de tempo, consulta ou não a fontes externas, são escritos como parte de uma tarefa e não textos criados por livre e espontânea vontade do autor. Além disso, soma-se o fato de que a língua do aprendiz seria uma fração da língua do falante nativo, crítica esta a que se apegam aqueles que julgam ser injusta a comparação de um *corpus* de aprendiz com um *corpus* de falantes nativos. Todavia, deve-se ressaltar que, ao contrastar ambos os *corpora*, não é objetivo do linguista de *corpus* menosprezar o aprendiz, pois o linguista sabe que haverá deficiências na produção linguística desse grupo, mas sim de fazer uma análise contrastiva com o intuito de sinalizar a professores e alunos que caminhos seguir, para onde o *corpus* de comparação aponta e que aspectos da interlíngua deste grupo precisa ser trabalhado.

4.3 **O *corpus* de estudo**

O *corpus* deste estudo consiste de 394 redações coletadas entre os anos de 2009 e 2012 em seis filiais do Rio de Janeiro de um dos maiores cursos livres de inglês do Brasil abrangendo as filiais da Zona Norte (aqui denominadas M e T), Zona Oeste (aqui denominadas F, B e CG) e Grande Rio (que recebeu as iniciais NI)[3].

O *corpus* compilado para este estudo é um *corpus* sincrônico, pois faz um recorte da língua dos aprendizes em um momento "X". O tamanho do *corpus* de análise é de 394 amostras, e contabiliza 91.333 *tokens*. Sendo assim, o *corpus* poderia ser classificado como de tamanho pequeno-médio, conforme afirma Berber-Sardinha (2004, p. 26), que estabelece essa categorização para *corpora* que têm entre 80 e 250 mil palavras. Também é importante mencionar que cada redação contém aproximadamente 250 palavras.

Este *corpus* de estudo pressupõe a inclusão por parte dos alunos dos quadrigramas estudados durante o semestre para dar conta especificamente do gênero resenha de filme (ver anexo A – Quadro 2).

[3] Os bairros das filiais foram omitidos para resguardar a confidencialidade dos dados. Com isto, siglas foram atribuídas para a identificação das mesmas.

Os textos coletados fazem parte da avaliação final que os alunos escrevem ao término do último semestre na instituição. Aos alunos são dadas as opções de escreverem três tipos de textos diferentes durante a avaliação, os quais já foram trabalhados em sala de aula e escritos em casa pelos mesmos como parte de tarefa extraclasse. São eles: narrativa pessoal, resenha de filme e artigo.

Ao escreverem os textos durante a avaliação os alunos dispõem de 1 hora e meia, tanto para a parte de conhecimento linguístico quanto para a redação, não consultam dicionário e já escreveram um texto semelhante em casa, o qual foi corrigido pelo professor e devolvido ao aluno com o devido *feedback*.

Os alunos estudados tiveram três aulas de composição escrita ao longo do semestre. O tema é basicamente trabalhado seguindo orientações do livro didático e com grande ênfase na prática oral do tópico a ser escrito em casa como tarefa extraclasse. Em termos práticos, isto significa dizer que os alunos tiveram pouca ou nenhuma oportunidade de praticar a habilidade escrita em sala, apenas ensaiando oralmente o conteúdo a ser reproduzido em casa ou recebendo o *input* acerca do conteúdo.

As redações foram colhidas de filial em filial, digitadas manualmente por este pesquisador e os erros de ortografia tiveram de ser corrigidos a fim de auxiliar a busca por palavras e quadrigramas pelo programa de computador, caso contrário o objetivo deste estudo seria prejudicado.

Os sujeitos desta pesquisa, à época da coleta, tinham estudado na instituição de ensino por cerca de sete anos, exceto alguma possível exceção de aluno de outra instituição que tenha sido submetido a um teste de nivelamento. São alunos que têm, em média, 17 anos de idade e estão na fase do pré-vestibular ou início da vida universitária; estão concluindo o curso com o nível que equivaleria ao intermediário-avançado, o que os equipararia ao nível do Quadro Comum Europeu B2 / Cambridge FCE, conforme a Figura 4 a seguir e as descrições para cada nível de proficiência.

CEFR Common European Framework	Levels	English Cambridge certificates
C2	Very advanced	CPE Certificate of Proficiency in English
C1	Advanced	CAE Certificate in Advanced English BEC Higher
B2	Upper intermediate	FCE First Certificate in English BEC Vantage
B1	Intermediate	PET Preliminary English Test BEC Preliminary
A2	Elementary	KET Key English Test
A1	Beginner	

Figura 4- Quadro Comum Europeu para referência de línguas.
Fonte: http://www.cambridgeenglish.org/images/126130-cefr-diagram.pdf

De acordo com o Quadro Comum Europeu para referência de línguas (CEF), os alunos que estão no nível B2 são classificados como “intermediário avançado” e aqueles que atingem esse nível são capazes de entender as ideias principais de textos complexos em tópicos concretos e abstratos, incluindo discussões técnicas em sua área de especialidade. Também podem interagir com um grau de fluência e espontaneidade que torna a interação regular com falantes-nativos bem possível sem dificuldades de ambas as partes. E por último, podem produzir textos claros e detalhados em uma ampla variedade de assuntos e explicando pontos de vista em um determinado tópico dando as vantagens e desvantagens de várias opções (COUNCIL OF EUROPE, 2011).

A instituição de ensino afirma sobre as habilidades adquiridas pelos alunos ao longo do curso avançado conforme consta no certificado de conclusão de curso, que os alunos tiveram 100 horas-aula em um período de 1 ano. No certificado também consta a seguinte afirmação acerca da habilidade linguística do aprendiz: “O aluno tem excelente domínio da língua inglesa falada e escrita. Interage com sucesso em situações sociais e acadêmicas, mesmo as mais complexas, como negociações, apresentações e debates. Tem habilidade para redigir textos de naturezas diversas tais como cartas formais e informais, ensaios e artigos”.

Ao longo dos anos de estudo de inglês, os alunos aprendem a escrever diversos tipos de texto, sempre recebendo o *input* em sala de aula e produzindo-os em casa e mais adiante

durante as provas. Aprendem a escrever textos como cartas (formais e informais), *e-mails*, relatórios, histórias, artigos, resenhas de filme, entre outros.

É mister ressaltar, que este estudo não fez o acompanhamento das aulas dadas pelos diversos professores ao longo dos anos, podendo cada professor ter trabalhado o conteúdo direcionado pelo programa curricular da instituição a seu próprio modo. No entanto, baseio meus comentários nos princípios metodológicos da instituição e o programa curricular que cada professor tem em mãos para preparar suas aulas.

A grosso modo, em uma aula de composição escrita, os alunos leem um texto típico do exemplar do gênero a ser trabalhado que é parte integrante de uma lição do livro didático do aluno, sugerido pelo programa curricular da instituição. Logo em seguida, respondem a perguntas de interpretação sobre o mesmo, oralmente, analisam que tipo de texto se trata e quais características constituem aquele gênero, como por exemplo, registro (formal, neutro ou informal), blocos léxico-gramaticais mais comuns, marcadores discursivos presentes, parágrafos e suas distribuições além do conteúdo dos textos. Embora muitas vezes haja espaço para um momento de prática individual do conteúdo, grande ênfase é dada ao trabalho oral em pares ou em pequenos grupos de alunos. Resumindo, dessa prática didática foram extraídos os quadrigramas ensinados.

Os dados coletados foram composições escritas de uma avaliação escrita que os alunos fazem durante o período. Ao longo do semestre, os alunos aprendem a trabalhar com, no mínimo, três gêneros diferentes em aulas de composição escrita. As avaliações são sempre divididas em prova A e prova B. Em cada prova existe um enunciado diferente pedindo a produção de um determinado gênero – um argumentativo ou um narrativo. Os alunos, então, escrevem um dos gêneros, caso tenham recebido a prova A ou B que são distribuídas aleatoriamente. As composições analisadas foram recolhidas de todas as turmas de último nível das seis filiais analisadas da instituição de ensino ao longo dos anos de 2009 a 2012. A partir daí, o material foi digitado manualmente, um a um por mim, a fim de ser elaborado o *corpus*, para que o mesmo fosse explorado com o auxílio do *WordSmith Tools 5.0*.

4.4 **A identificação de arquivos em um *corpus***

Nesta seção menciono a dificuldade que pode ser para um linguista de *corpus* encontrar informações específicas dentro de uma imensidão de arquivos que constitui um *corpus* digitalizado. Sendo assim, um *corpus* pode conter anotações de diversos tipos e/ou cabeçalhos[4], cada um com propósitos específicos, mas que facilitam ao analista buscar a informação que deseja de forma mais precisa e fazer estudos envolvendo tanto abrangência quanto frequência de uso (ou não uso) de certas ocorrências linguísticas, conforme passo a discutir a seguir.

Sendo um *corpus* formado por um grande número de textos seria muito difícil, dentro de uma imensidão de dados, localizar a informação desejada. Como uma biblioteca precisa de catalogação para que se encontre um livro seja pelo autor, nome da obra, editora, ano de publicação, etc., um *corpus* também necessita de uma identificação em forma de cabeçalho (*header*) para se buscar um texto específico e também para se analisar uma quantidade de textos agrupados de acordo com um critério de busca escolhido pelo analista. Um *corpus* bem identificado com critérios pré-estabelecidos "permite que se gerem estatísticas para um fenômeno linguístico" (BERBER-SARDINHA, 2004, p. 73).

Muito se fala sobre anotação de *corpus* como uma forma de refinar e facilitar a pesquisa com LC. Existem hoje em dia diversos tipos de anotações específicas que podem ter diversos objetivos.

O linguista pode fazer anotações sobre meta-dados que identifiquem o *corpus* propriamente dito, como e quando foi compilado, meio de compilação, modo oral ou escrito, mas pode também atribuir ao *corpus* um número de etiquetas para identificar erros, anotações morfossintáticas, figuras de linguagem, entre tantas outras. As anotações podem ser feitas tanto manualmente, o que daria um pouco mais de trabalho ou através de programas de computador específicos que reconhecem os formatos dos arquivos de textos submetidos (.txt) e os retornam com as etiquetas prontas.

Tanto a anotação com meta-dados quanto à etiquetagem auxiliam o analista de *corpus* a filtrar sua busca e com isso gerar listas de comparações, traçar perfis e estatísticas em seus estudos. A presente dissertação não faz uso de etiquetagem com categorias distintas, mas somente de anotação de meta-dados, incluindo cabeçalhos sobre o conteúdo de cada texto.

[4] Para exemplos de cabeçalhos ver Anexo E.

Nesta dissertação, entende-se por cabeçalhos, que foram denominados no *corpus* de estudo como <header>, uma espécie de ficha que detalha elementos de cada texto. Na dissertação foi importante identificar dados como: filial onde o texto foi escrito, se a redação foi cronometrada ou não (neste estudo todas elas foram), o sexo do aprendiz, o tipo de texto escrito (narrativa pessoal, artigo ou resenha de filme), a data em que o texto fora escrito (neste caso entre o primeiro semestre de 2009 e o segundo semestre de 2012), o nome do professor e por último, o título, campo este que em alguns casos fora deixado em branco por alguns alunos que não escreverem títulos nos textos.

Usar um cabeçalho auxilia as análises que seguem no próximo capítulo a partir do fato de que é possível traçar um perfil abrangente não apenas dos aprendizes envolvidos (sexo e faixa etária), mas também da instituição de ensino em termos de qual zona do Rio de Janeiro pertencem as filiais investigadas e as próprias filiais envolvidas na pesquisa.

5 **ANÁLISE DOS DADOS**

Neste capítulo, mostro os resultados da análise do *corpus* de estudo. Esta pesquisa, conforme anteriormente dito, é orientada pelo *corpus*, ou seja, não parte de nenhum pressuposto teórico para aplicação no *corpus*, mas nasce através da análise do mesmo. Isso significa dizer que através das evidências surgidas do *corpus* é que comecei a traçar o perfil da escrita dos sujeitos aprendizes.

O capítulo mostra as configurações dos dados em relação a vários pontos a saber: filial estudada, região do Rio de Janeiro onde as filiais se encontram, tipos de quadrigramas e quadrigramas errôneos nas redações dos sujeitos investigados

5.1 **Redações e suas distribuições por filiais, gêneros textuais e tipos de erro**

Minha primeira tentativa de compreender meus dados foi entender como eles estavam distribuídos. O fato de estarem em diferentes regiões do Rio de Janeiro tem significados específicos. Conforme será mostrado no Gráfico 2 a seguir, as filiais B e CG são filiais que se encontram na Zona Oeste da cidade do Rio de Janeiro, uma área mais carente de investimentos, com menor poder aquisitivo da população e menos acesso a diferentes meios culturais como teatros, bibliotecas e cinemas. Já a filial F, embora se encontre também na Zona Oeste, está em uma área em ascensão sócioeconômica com muitos lançamentos imobilários para a classe média e encontra-se bem próxima a região da Barra da Tijuca com maior acesso a meios de serviço e produções culturais. A filial T é uma das mais tradicionais do Rio de Janeiro, sendo uma filial de grande porte e em uma área com bom poder aquisitivo na Zona Norte do Rio de Janeiro. Os valores das mensalidades tanto desta filial quanto da filial T são maiores que as demais mencionadas neste estudo.

M encontra-se em uma área da Zona Oeste considerada subúrbio, uma área e filial decadentes com número de alunos em declínio e área marcada por violência, escassez de serviços culturais, embora seja uma área de grande densidade demográfica. Por último, a filial NI encontra-se na área que denomina-se Grande Rio, um município da Baixada Fluminense vizinho a cidade do Rio de Janeiro mas pertencente a área conurbada da metrópole fluminense. Esta filial esteve em declínio 10 anos atrás, mas destaca-se, hoje em dia, como

umas das 7 maiores da rede no estado do Rio de Janeiro com mais de 1800 alunos. A cidade onde a filial encontra-se é uma das maiores do estado, com poder econômico em crescimento graças ao forte setor de serviços e comércio e localiza-se a cerca de 30 km de distância da capital do Rio de janeiro. Embora a cidade possua boa gama de serviços, ainda é carente de atrativos culturais tendo a população que se deslocar a municípios vizinhos em busca de tais opções de entretenimento.

No Gráfico (1) a seguir, pode se ver a quantidade de redações que foi coletada para análise, isto é, o número de alunos de cada filial que conseguiu terminar o curso avançado após 7 anos de estudos. No gráfico trabalhou-se com porcentagens, e daí ser possível ver que a maioria das redações (43%) foram coletadas na filial NI, em segundo lugar com 29% aparece a filial T, em terceiro com 11,3% dos textos a filial F. 8, 6% dos textos coletados provem da filial M e logo em seguida com 4,3% e 3,8%, respectivamente, as filiais B e CG.

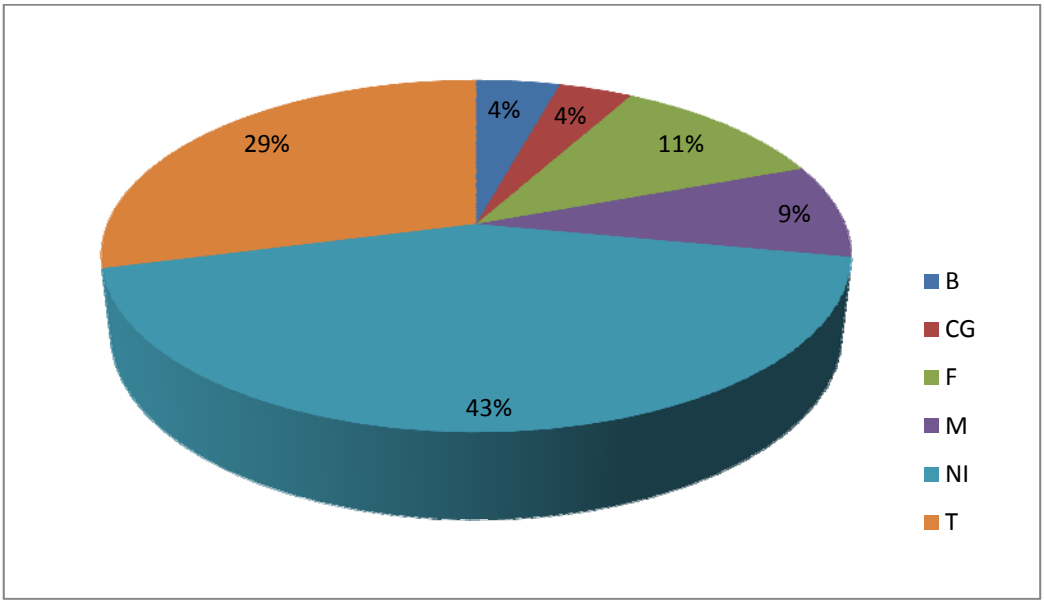

Gráfico 1- Quantidade de redações coletadas por filial.

A distribuição de redações permite que se faça um recorte da instituição de ensino no que tange ao Grande Rio (NI) com 43% dos textos, Zona Norte (M e T) 37,6% e Zona Oeste (B, CG e F) 19,4%. Logo, parece ficar mais homogênea a divisão das redações se analisarmos as mesmas pelas regiões que foram obtidas e não pelas filiais isoladamente conforme mostra o Gráfico 3 que se segue. Sendo assim, 80,6% dos textos coletados foram provenientes da capital fluminense e os restantes 19,4% do município vizinho. Se os dados fossem normalizados, com as filiais contribuindo com a mesma quantidade de textos cada, os resultados obtidos poderiam ter sido diferentes.

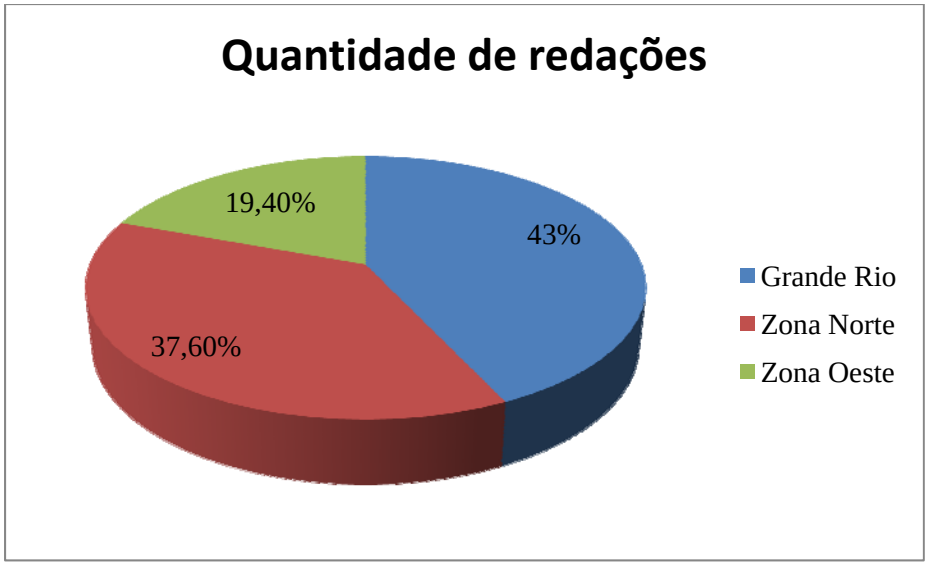

Gráfico 2- Quantidade de redações coletadas por região.

Como já mencionei, durante a atividade de avaliação, os alunos puderam escolher dois tipos de texto para escrever, gêneros esses que já haviam sido praticados oralmente em classe e escritos como trabalho de casa e posteriormente comentados pelo professor individualmente.

Dividi o total de redações colhidas por filial em relação a cada um dos três diferentes tipos de texto produzidos pelos alunos. Os aprendizes parecem ter ficado divididos igualitariamente entre a escrita de dois gêneros; a resenha de filme e a narrativa pessoal, com 48,7% e 44, 4%, respectivamente. Por outro lado, uma pequena minoria dos aprendizes, apenas 6,9% escolheu escrever um artigo. Os aprendizes que optaram por esse tipo de texto encontram-se esparsamente ao longo das seis filiais analisadas mas não em número igual. Segue adiante o Gráfico (3) com a divisão das redações coletadas por gênero textual.

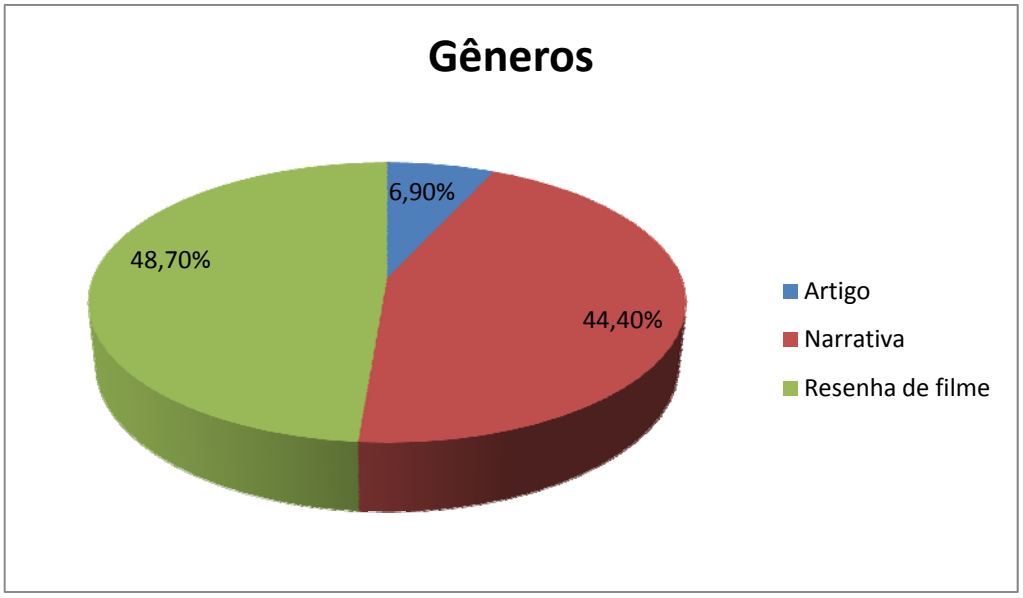

Gráfico 3- Gêneros textuais escritos pelos aprendizes.

5.2 **Quadrigramas em redações de curso livre**

Após um mapeamento inicial que permitiu saber a distribuição de redações por área do Rio de Janeiro e por gênero, o segundo passo foi o uso do conjunto de programas *Wordsmith Tools 5.0* e duas ferramentas que são a Wordlist (lista de palavras) e o Concordance (concordanciador).

Utilizei a lista de palavras para a extração dos quadrigramas mais frequentes e conforme demonstro na Tabela 6 (ver anexo B) os 150 quadrigramas mais frequentes encontrados no *corpus*.

A quantidade total de quadrigramas com um mínimo de 4 ocorrências foi de 293. Se somarmos aos quadrigramas que tiveram um mínimo de três ocorrências, o total aumenta para 580 quadrigramas encontrados. Decidi descartar os que tiveram um mínimo de 2 ocorrências pois, em alguns casos, o mesmo aprendiz reutilizava o mesmo quadrigrama. Estabeleci como ponto de corte para a análise deste estudo os primeiros 150 por um critério de importância, frequência (o ponto de corte seria cinco/quatro) e de distribuição – os blocos teriam de aparecer em pelo menos duas redações diferentes. Em outras palavras, este estudo se baseia em frequência e abrangência, ou seja, 150 parece ser um número que me permite ter uma boa frequência de quadrigramas e que abranja todas as filiais analisadas, inclusive as que tenham contribuído com menos textos. Embora dentro da LC o que não é usado, ou seja, o que não

ocorre também pode ter sua importância, neste estudo vou me ater apenas ao que o *corpus* mostra de forma explícita.

Na terceira coluna da lista do Anexo B, colocada no Anexo devido a sua extensão, pode-se ver a frequência de quadrigramas encontrados, sendo o quadrigrama “tells the story of” o mais frequente, usado 38 vezes no mesmo número de textos. No entanto, é evidente o fato de o quadrigrama de número 2 “the end of the” ter sido usado 24 vezes mas aparecer em apenas 22 textos, ou seja, 2 aprendizes repetiram o quadrigrama ao longo de suas composições, fato esse também presente nos quadrigramas de número 3, 5, 6, 20, 23, 27, 28, 30, 31, 32, 34, 37, 44, 46, 50, 51, 56, 68, 74, 76, 81, 97, 101, 103, 121, 129, 131, 135, 145 e 149. Isso significa dizer que, 31 quadrigramas, ou 21%, foram usados no mesmo texto mais de uma vez.

Em seguida, tracei um perfil dos quadrigramas produzidos pelos aprendizes de cada filial. Embora o número de redações tenha variado de filial para filial, o que analiso é a porcentagem de produção de quadrigramas encontrados em cada filial. Multiplicando-se os 150 primeiros quadrigramas utilizados pelo número de casos, obtive um número total de 1.220 ocorrências de quadrigramas, os quais estão divididos de acordo com as filiais da instituição de ensino, em porcentagem (Tabela 7).

Filial	Porcentagem
B	4.6%
CG	7.3%
F	11.3%
M	9.8%
NI	41.0%
T	26.0%
Total	100%

Tabela 7- Quadrigramas por filial.

A lista anterior foi organizada por ordem alfabética e conforme pode se constatar, a filial com o maior número de quadrigramas foi NI e também a que contribuiu com o maior número de textos para o *corpus* (43%) e a de menor uso de blocos foi B, com 4,3% de redações à frente apenas da filial CG que, embora tenha tido o menor número de redações presentes no *corpus* (3,8%) produziu 7,3% dos quadrigramas. NI e T, as duas filiais com o maior número de quadrigramas e redações, são grandes filiais e com um grande número de alunos (em torno de 1800 a 2000) e por esse fato, muitos alunos concluintes, isso significa um grande número de redações coletadas nessas duas filiais. Já B é uma filial mais nova e com menos alunos, daí a dificuldade de se coletar tantas redações do nível estudado.

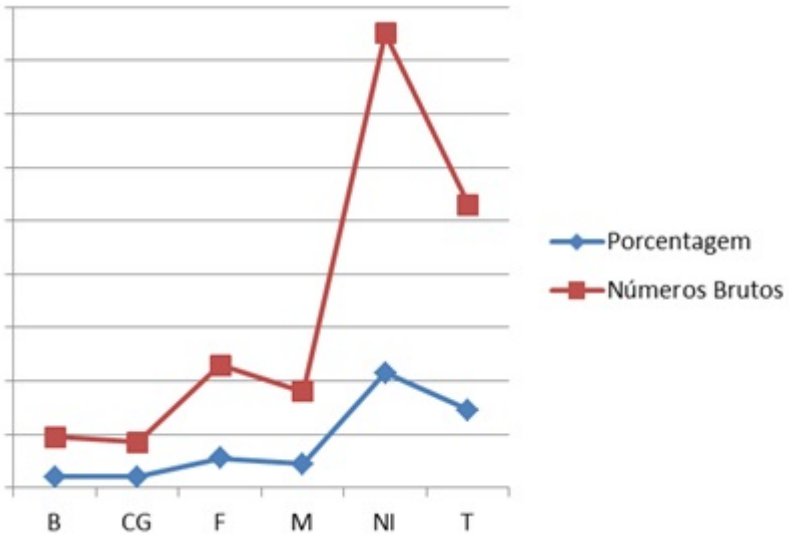

Gráfico 4- Números brutos *versus* porcentagem de quadrigramas nas filiais.

A análise do Gráfico 4 acima parece sugerir que existe uma relação entre o número total de redações e a quantidade de quadrigramas extraídos pelo programa de computador já que as quatro primeiras filiais em número de redações mantiveram a mesma ordem na relação de quadrigramas encontrados, apenas fugindo à regra as filiais B e CG. Nelas foi interessante perceber que não necessariamente mais textos significou mais quadrigramas. Outro fato foi que em uma menor quantidade de textos provenientes da filial CG foram encontrados mais blocos que na filial B. Observando-se a linha de números brutos CG teve menos textos, mas mais quadrigramas que a filial B.

O Gráfico 5 auxilia-nos a ter uma melhor visualização da divisão de quadrigramas extraídos pelo programa de computador e que são provenientes das seis filiais analisadas.

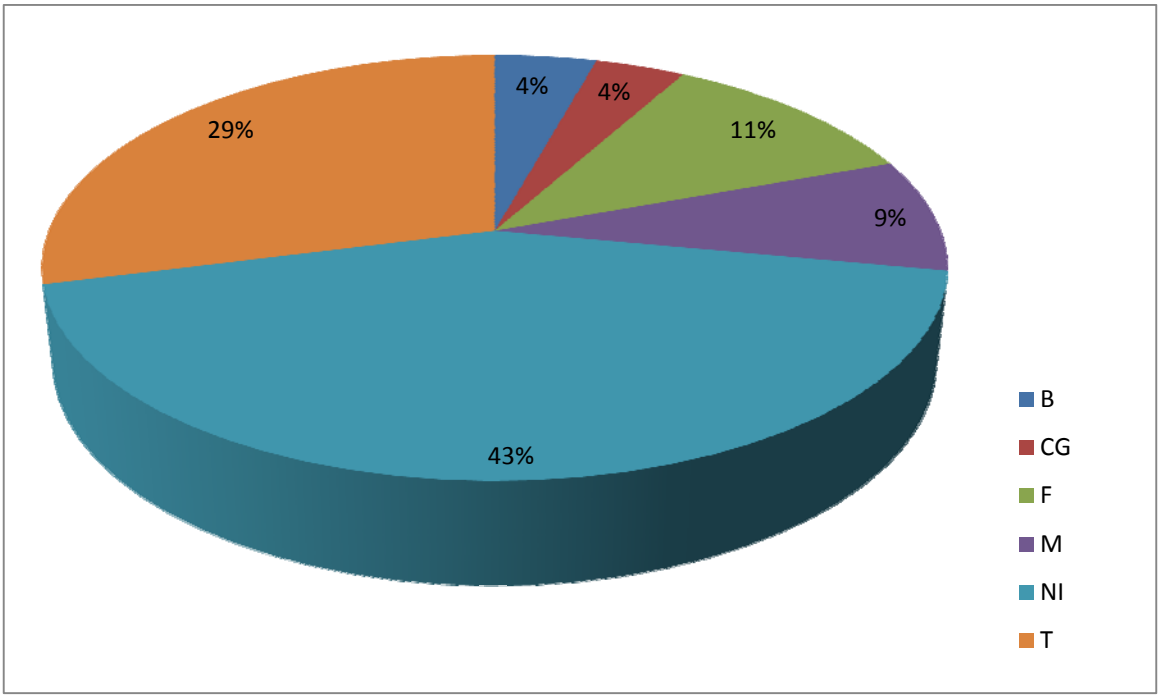

Gráfico 5- Divisão dos quadrigramas nas filiais.

Passo agora a abordar a relação entre o gênero e os blocos léxico-gramaticais. Bakhtin (2003) afirma que o gênero escolhe o léxico. Trazido para a realidade da sala de aula esse pensamento quer dizer que, dado um gênero, o aluno deveria utilizar o léxico de acordo com o gênero que está escrevendo. O léxico de uma receita de bolo não será o mesmo de uma carta pois o aluno tem moldado em sua mente quais escolhas léxico-gramaticais seriam mais apropriadas para o gênero em questão. Partindo dessa premissa, decidi fazer a divisão dos quadrigramas extraídos do *corpus* de acordo com o gênero em que os mesmos poderiam ser encontrados conforme começo a discutir a seguir.

Em primeiro lugar, há a resenha de filme, na qual 46% dos quadrigramas analisados estavam presentes única e exclusivamente. Tal resultado pode vir a ser explicado pelo fato de essa ter sido a aula com maior ênfase na aprendizagem de blocos lexicais característicos de cada movimento retórico de um exemplar prototípico deste gênero. Com isso, o aluno pode ter tido maior consciência da importância de replicar o conteúdo ensinado na redação da prova a fim de "soar" como um falante nativo ou de apenas garantir uma boa nota na avaliação. Em segundo lugar, estão os quadrigramas da narrativa. Muitos desses quadrigramas encontrados são quadrigramas que podem ser classificados como neutros, pois não seriam típicos de uma história, mas podem ser utilizados e foram encontrados em outros tipos de texto. Fato esse confirmado pelos 14% de quadrigramas encontrados tanto na resenha quanto

na narrativa mostrando mais uma vez a neutralidade do quadrigrama em questão e a capacidade do aprendiz de reciclar o mesmo em diferentes contextos.

Como os quadrigramas obtidos do artigo foram em quantidade ínfima (5%) decidi ater-me apenas à discussão das três maiores porcentagens, conforme mencionado até aqui. Segue abaixo o gráfico com a divisão dos quadrigramas extraídos do *corpus* de acordo com o gênero escrito pelos alunos (Gráfico 6).

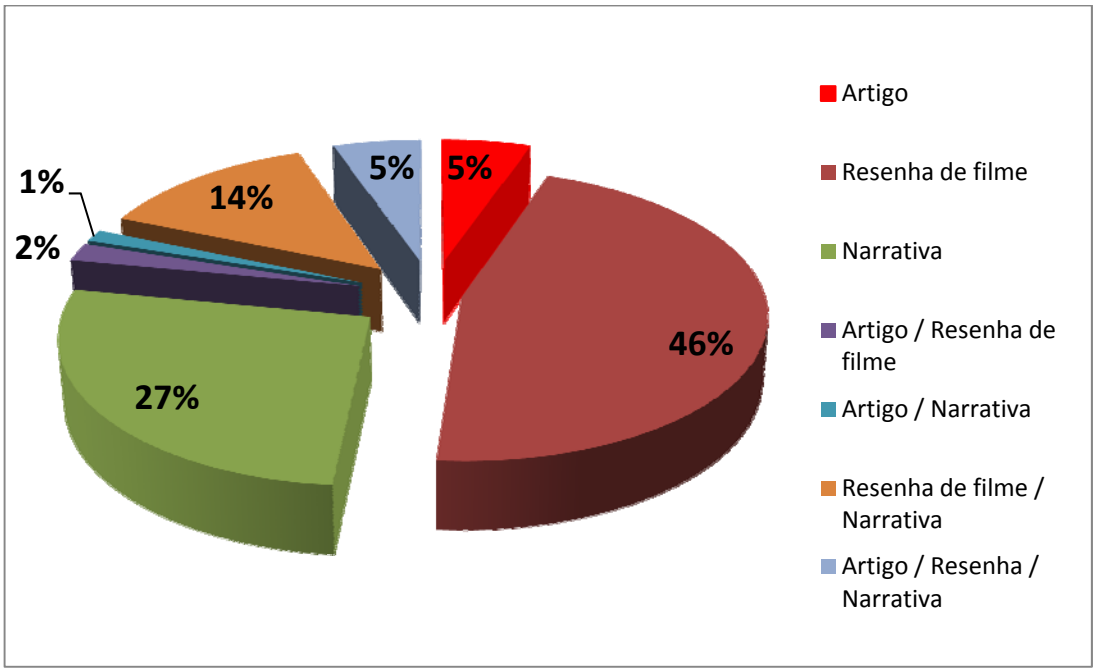

Gráfico 6- Divisão de quadrigramas e gêneros.

5.3 **Quadrigramas errôneos**

Antes de passar para a análise dos quadrigramas errôneos extraídos do *corpus*, seria interessante mencionar os quadrigramas ensinados para a produção da resenha de filme em específico (conforme Anexo C). Embora os quadrigramas errôneos extraídos do *corpus* não contemplem unicamente os ensinados para a escrita da resenha de filme, outros quadrigramas ensinados anteriormente foram extraídos com o auxílio do programa de computador, como foi o caso do quadrigrama “on the other hand”, tendo tido 3 ocorrências como “in the other hand”. Sejam eles erros por qualquer que seja a razão (influência da língua materna do aprendiz, da interlíngua ou de tentativa) os erros foram analisados e classificados sob dois vieses. O primeiro deles uma perspectiva pedagógica, voltada para a prática didática e de fácil

assimilação pelo professor de inglês e o aprendiz e o segundo, uma classificação voltada para os princípios da Linguística de *Corpus*, conforme mencionarei mais adiante. Analisando-se todo o *corpus*, ou seja, ultrapassando o limite de 150 quadrigramas foram encontrados 25 quadrigramas errôneos. O critério de busca foi que o quadrigrama tivesse sido produzido por, no mínimo, três aprendizes diferentes sendo da mesma filial ou não. Esse número pode ser considerado baixo, se levado em conta o total de quadrigramas produzidos por três aprendizes extraídos do *corpus*, um número bruto de 537. Ou seja, do total de quadrigramas extraídos, 4,7% são errôneos. Esse é um número aparentemente pequeno, mas que chama a atenção a partir do fato de que o mesmo erro ocorre não apenas entre aprendizes mas também filiais diferentes.

Passo a seguir a analisar a lista de quadrigramas errôneos, como pode-se ver na Tabela 7 a seguir. O quadrigrama errôneo mais frequente foi "have a lot of", que foi produzido 12 vezes, por 9 alunos diferentes de NI e 3 de M.

Os quadrigramas de maior dispersão nas filiais foram os de número 3 e 5, tendo sido produzidos por cinco filiais diferentes. Ou seja, não importa a origem do erro, o fato é que ele ultrapassa os muros da filial e parece ser compartilhado por grupos de aprendizes.

				Filiais Individuais					
Núm	**Quadrigrama**	**Frequência**	**Textos**	**B**	**CG**	**F**	**M**	**NI**	**F**
1	Have a lot of[5]	12	12	0	0	0	3	9	0
2	Tells the story about	8	8	0	1	0	2	2	3
3	Had a lot of	7	7	1	1	0	2	2	1
4	In this day I	4	4	0	0	0	0	1	3
5	Tells a story about	5	5	1	1	1	1	0	1
6	the history of a	5	5	1	0	0	0	1	3
7	A Oscar but should	4	4	0	1	1	0	2	0
8	Have won a Oscar	4	4	0	2	0	0	1	1
9	Not win a Oscar	4	4	0	1	1	0	2	0
10	Should have won a	4	4	0	2	0	0	1	1
11	Win a Oscar but	4	4	0	1	1	0	2	0
12	To enter in the	5	3	0	0	1	0	4	0
13	The final of the	4	3	0	1	0	1	2	0
14	A film that not	3	3	0	0	1	0	2	0
15	Cast is composed by	3	3	1	2	0	0	0	0
16	Film talk about a	3	3	0	0	0	1	2	0
17	Film tells the	3	3	1	0	0	0	1	1

[5] O quadrigrama em questão fora expandido e analisada toda a linha de concordância para, então, concluir-se que trata-se de um uso errôneo.

	history								
18	Film that not win	3	3	0	0	1	0	2	0
19	In love for a	3	3	0	0	0	1	2	0
20	In the other hand	3	3	0	0	1	0	0	2
21	In this moment I	3	3	1	0	0	0	0	2
22	The film talk about	3	3	0	0	0	1	2	0
23	The history and the	3	3	1	0	0	0	2	0
24	To insert an idea	3	3	0	0	1	0	0	2
25	In my point of (view)	4	4	0	0	0	0	3	1

Tabela 8- Lista de quadrigramas errôneos e filiais onde ocorreram.

A classificação para erros sugerida por Granger (2006), demasiado extensa (54 categorias), não se aplicaria aos itens desta dissertação de mestrado, visto que a autora trabalha com itens gramaticais individuais isto é, as categorias são anotadas em cada item lexical individualmente.

A classificação dos erros já listados na Tabela 8 é uma versão simplificada da categorização de Granger e dá conta dos blocos errôneos encontrados no *corpus* de estudo. Ela foi feita segundo 7 tipos diferentes de erros encontrados nos quadrigramas, sendo eles: preposição, artigo, 3ª pessoa do singular/escolha lexical errônea, falta de verbo auxiliar, escolha lexical errônea, erro de pronome demonstrativo e "there" (existencial). A seguir explico cada tipo de erro mencionado acima e exemplifico com casos retirados do *corpus*.

O primeiro deles é o que classifico como erro de preposição. Muitos alunos parecem ter dificuldade em assimilar qual preposição ou verbo vem junto a uma determinada palavra. Mas na verdade, pode se dizer que tais erros seriam de colocação. Por colocação entende-se o grau de uma palavra de atrair uma outra tão fortemente que ambas parecem ser inseparáveis. Esses blocos léxico-gramaticais sempre atrairão certas preposições não havendo outra possibilidade de alteração, adição ou exclusão. Conforme mostra a Tabela 9 adiante, foram encontrados sete erros classificados como erros de preposição. O primeiro quadrigrama errôneo foi encontrado em oito textos diferentes já a forma correta ocorreu 38 vezes, um número quase cinco vezes maior. Isso quer dizer que a cada seis tentativas de uso desses quadrigramas, uma é errônea.

No erro de número seis o aprendiz utilizou a preposição errônea "in" ao invés de "on/from", no entanto, esta também poderia ter sido omitida não sendo necessária na frase, podendo haver assim, uma exclusão da preposição. O exemplo de número 7 mostra um caso de sobreuso de preposição sendo esta desnecessária na frase.

Núm	Erro	Quantid. de quadrigramas errôneos	Forma correta
1	Tells the story about	8	Tells the story of
2	Tells a story about	5	Tells a story of
3	In love for a	3	In love with a
4	In the other hand	3	On the other hand
5	In my point of (view)	4	From my point of
6	In this day I	4	(On) this day I
7	To enter in the (university)	5	To enter university

Tabela 9- erros de proposição.

Embora os erros elencados a seguir possam ser considerados básicos para o nível estudado - o nível avançado, parece haver um descuido quanto ao uso do artigo indefinido "a/an" por alguns alunos já que foram encontrados cinco quadrigramas errôneos com tal ocorrência (Tabela 10).

Núm	Erro	Quant. de quadrigramas errôneos	Forma correta
1	A Oscar but should	4	An Oscar but should
2	Have won a Oscar	4	Have won an Oscar
3	Not win a Oscar	4	Not win an Oscar
4	Should have won a	4	Should have won an
5	Win a Oscar but	4	Win an Oscar but

Tabela 10- erros de artigo indefinido.

À primeira vista, o erro descrito a seguir parece estar no verbo da terceira pessoa do singular que tem a terminação "s" faltando. No entanto, também houve uma escolha lexical errônea porque o livro não "fala sobre" e sim, em inglês, "é sobre" algo. Logo, a forma correta do pentagrama seria: "the film is about a".

Núm	Erro	Quant. de quadrigramas errôneos	Forma correta
1	Film talk about a	3	Film is about a
2	The film talk about a	3	The film is about a

Tabela 11- Terceira pessoa do singular/escolha lexical errônea.

O erro listado adiante, como muitos outros, também foi extraído duplamente por ter sua origem em um pentagrama. No entanto, decidi também manter a divisão em dois quadrigramas a fim de manter o critério inicial de análise defendido por este estudo. Como

pode se ver na Tabela 12, o erro é devido a uma má construção da forma negativa. O único aprendiz não internalizou o uso do verbo auxiliar “did” necessário para expressar o passado na negativa.

Núm	Erro	Quant. de quadrigramas errôneos	Forma correta
1	A film that not	3	A film that did not win
2	Film that not win	3	Film that did not win

Tabela 12- Verbo auxiliar faltando.

Com seis ocorrências, as escolhas lexicais errôneas não variam muito. Em primeiro lugar pode-se observar o mau uso do substantivo “history” ao invés de “story”; logo depois, a possível influência da língua mãe ocasionando um erro na escolha do léxico “final” em detrimento de “end” e também um outro possível erro de interlíngua, no qual esse aluno demonstra não ter aprendido a expressão idiomática “put an idea into somebody’s head” fazendo uma tentativa errônea com base em sua língua materna “to insert an idea” (Tabela 13).

Núm	Erro	Quant. de quadrigramas errôneos	Forma correta
1	The history of a	5	The story of a
2	The final of the (game)	4	The end of the
3	Film tells the history	3	Film tells the story
4	The history and the	3	The story and the
5	To insert an idea	3	Put an idea into somebody’s head
6	Cast is composed by	3	Cast includes

Tabela 13- Escolha lexical errônea.

O quadrigrama que passo a analisar demonstra que houve um erro na escolha do pronome demonstrativo para referência anafórica e também da preposição que o acompanha, pois os aprendizes estavam se referindo a um momento anterior que havia mudado sua vida (Tabela 14).

Núm	Erro	Quant. de quadrigramas errôneos	Forma correta
1	In this moment I	3	At that moment I
2	At this moment I	3	At that moment I
3	In that moment I	3	At that moment I
4	On that moment I	2	At that moment I

Tabela 14- Erro de pronome demonstrativo e/ou preposição.

Em último lugar, um erro muito frequente entre os aprendizes brasileiros de inglês como língua estrangeira é o “there” com sentido existencial, muitas vezes confundido com o verbo “ter” em língua portuguesa, os alunos fazem mau uso do mesmo ao evitarem a expressão “there is/are” pelo verbo “have”, tanto no presente, passado, singular e plural (Tabela 15)[6]. O primeiro quadrigrama errôneo foi encontrado 12 vezes e o segundo teve 7 ocorrências.

Núm	Erro	Quant. de quadrigramas errôneos	Forma correta
1	Have a lot of	12	There is a lot of
2	Had a lot of	7	There were lots of

Tabela 15- Erro de “There” existencial.

Após listar cada uma das categorias que lancei mão para classificar os quadrigramas errôneos encontrados no *corpus*, decidi, então, agrupá-los em um gráfico que permitisse melhor visualização dos erros. De acordo com o Gráfico 7 a seguir, o erro mais frequente é o de preposição com 27% dos casos. Em segundo lugar com 19% foram as escolhas lexicais errôneas. O erro na escolha do artigo e o uso do “there” com sentido existencial tiveram 17% e 16% de ocorrências, respectivamente. Em 5º lugar dentre os erros mais frequentes foram os quadrigramas contendo erros tanto de pronome demonstrativo quanto de preposição com 10% dos casos. E por último, com 6% e 7% estão os erros de 3ª pessoa do singular e omissão do verbo auxiliar.

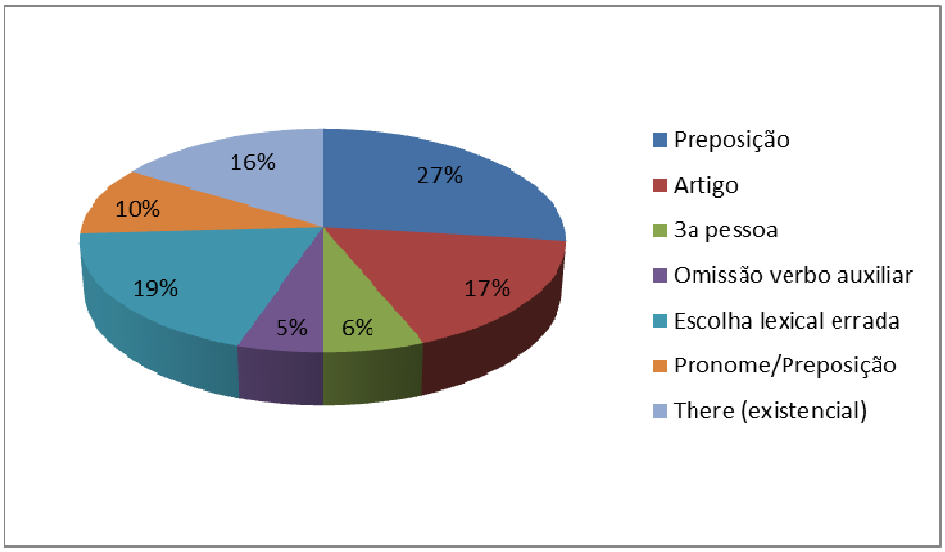

Gráfico 7- Tipos de erros encontrados no *corpus*.

[6] Para classificar o quadrigrama como errôneo foi preciso expandir a linha de concordância e analisar todo o seu entorno, para então, considera-lo como erro.

O mapeamento dos erros acima pode ser muito útil para um trabalho de correção e tratamento de erros, auxiliando o professor a lidar com alunos que possuem erros fossilizados e ampliando seu entendimento da interlíngua do aprendiz brasileiro. Com isso, os erros encontrados, embora relativamente poucos em relação ao tamanho do *corpus* podem contribuir para uma melhor prática didática do professor e uma melhor produção do aprendiz de L2.

No entanto, para o linguista de *corpus* a classificação descrita anteriormente parece fazer pouco sentido, sendo necessária uma análise que dê conta dos princípios que a Linguística de *Corpus* aborda como colocação, coligação, prosódia semântica, entre outros.

Sendo assim, decidi organizar os erros previamente mencionados segundo tais princípios e sugiro a classificação dos mesmos conforme erros de colocação e coligação apenas, demonstrados nas Tabelas 16 e 17, a seguir.

Erros de colocação	**Quant.**	**Forma correta**
Film talk about a	3	Film is about a
The film talk about	3	The film is about
Cast is composed by	3	Cast includes
The history of a	5	The story of a
The final of the (game)	4	The end of the
Film tells the history	3	Film tells the story
The history and the	3	The story and the
To insert an idea	3	Put an idea into somebody's head

Tabela 16- erros de colocação.

Erros de coligação	**Quant.**	**Forma correta**
Tells the story about	8	Tells the story of
In love for a	3	In love with a
In the other hand	3	On the other hand
In my point of (view)	4	From my point of
In this day I	4	(On) this day I
In this moment I	3	At that moment I
At this moment I	3	At that moment I
In that moment I	3	At that moment I
On that moment I	2	At that moment I
To enter in the (university)	5	To enter university
A Oscar but should	4	An Oscar but should
Should have won a	4	Should have won an
Win a Oscar but	4	Win an Oscar but

Have won a oscar	4	Have won an Oscar
Not win a oscar	4	Not win an Oscar
A film that not	3	A film that did not
Film that not win	3	Film that did not
Have a lot of	12	There was a lot of
Had a lot of	7	There were lots of

Tabela 17- erros de coligação.

No Gráfico a seguir (8), pode-se ver a divisão de erros segundo conceitos provenientes da Linguística de *Corpus*. Há uma predominância de erros de coligação com 76% dos casos sobre os erros de colocação com 24%. Isso significa que os aprendizes produzem mais erros referentes à estrutura gramatical dos blocos e em menor quantidade, erros de colocação.

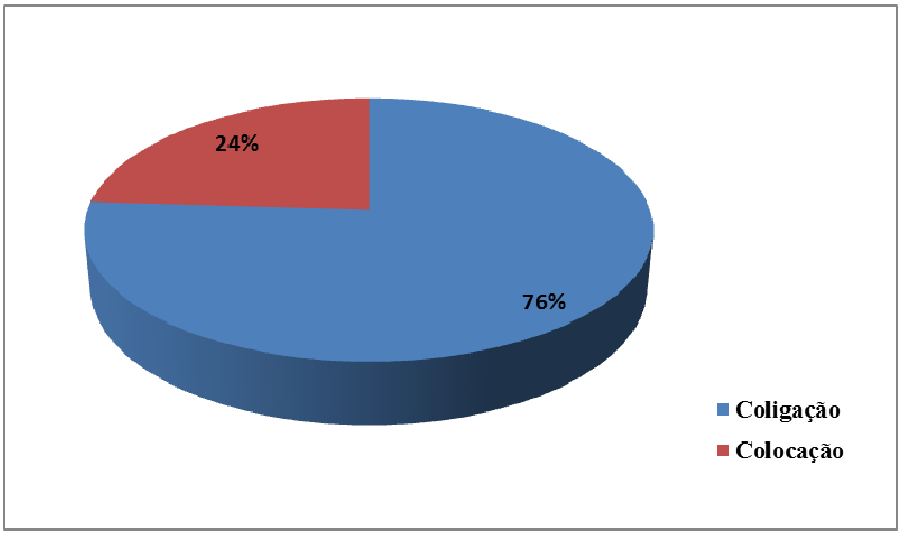

Gráfico 8- Erros de coligação e colocação.

Se comparada a proporção de produção de quadrigramas certos com errôneos, pode-se fazer uma analogia que quanto maior a quantidade de textos e palavras maior a probabilidade de se encontrarem erros, pois parece haver uma relação entre quantidade de textos, número de quadrigramas certos e errados. Os alunos erram muitas vezes porque estão tentando escrever certo. Dessa forma, seus erros são ocasionados por tentativas, isto é, quanto maior a tentativa maior a probabilidade do erro ou acerto. No caso deste estudo, pode-se dizer que os alunos acertam bastante, e que os mesmos escrevem razoavelmente bem, se analisados pelo viés da frequência de erros encontrados. Todavia, ainda há espaço para se analisar outras questões que dão conta do nível da escrita deste grupo de aprendizes como a densidade lexical dos textos, o nível de sofisticação léxico-gramatical das redações, entre outros filtros que poderiam ser aplicados em outras análises.

Adiante, demonstro, no Gráfico 9, a porcentagem de quadrigramas errôneos nas filiais estudadas. Embora a filial NI tenha mais de 40% dos erros produzidos, deve-se ressaltar que esta filial contribuiu com 41% de quadrigramas e 43% dos textos. Não se pode afirmar com isso que os alunos desta escola escrevam mal, mas sim que ainda não concluíram seu processo de aprendizagem. Logo, esses alunos ainda podem aprimorar sua escrita ao longo dos anos, caso continuem a fazer uso da língua inglesa em seus estudos.

Voltando aos percentuais de erros do Gráfico 9, as filiais T e M na Zona Norte do Rio de Janeiro possuem o 3º e 4 lugar respectivamente de quadrigramas errôneos. As três últimas colocações, então, são provenientes da Zona Oeste.

A filial F chama a atenção no Gráfico, pois parece ser uma exceção ao que foi dito até agora. No ranking de produção de quadrigramas a mesma aparece em 3º lugar com 11% dos quadrigramas extraídos do *corpus*. No entanto, ao analisarem-se apenas os errôneos esta filial cai para 5º lugar. Tal fato pode vir a significar que os alunos dessa filial tem maior consciência do uso de blocos lexicais e que também tenham tido maiores oportunidades de prática.

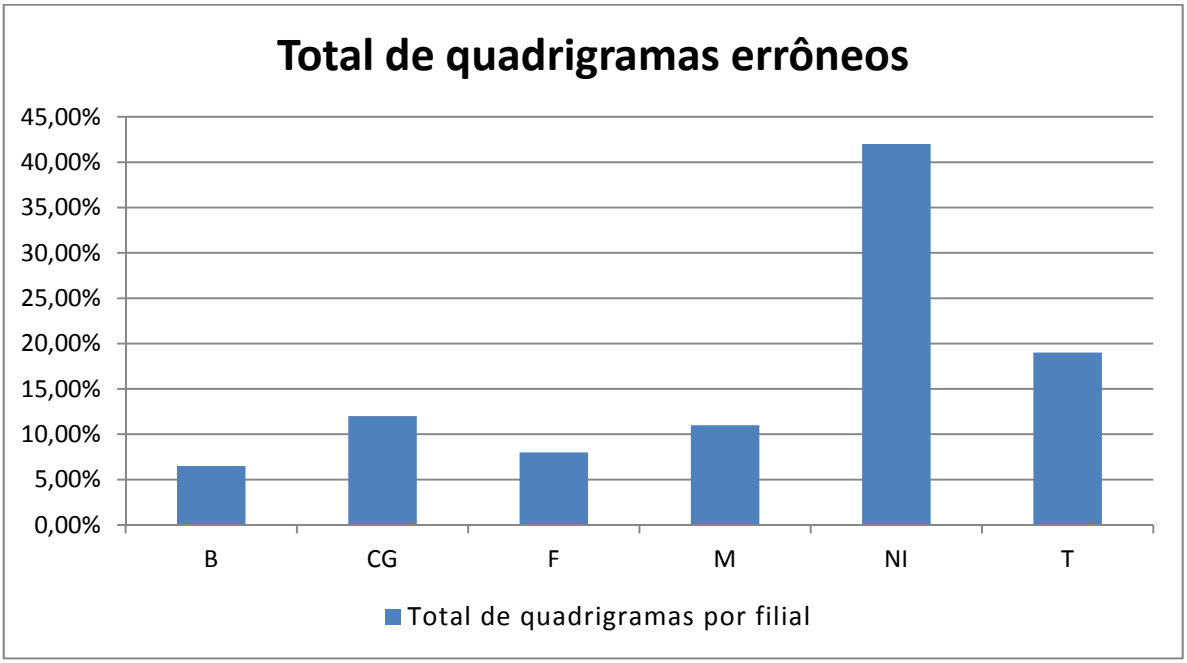

Gráfico 9- Total de quadrigramas errôneos por filial.

5.4 **Para uma categorização de quadrigramas: a proposta de Hyland**

Em um estudo de 2008, no qual analisou textos acadêmicos, Hyland fez uma proposta enxuta para lidar com a categorização de quadrigramas: sugeriu que os blocos lexicais extraídos do seu corpus fossem divididos de acordo com as três macrofunções que porventura poderiam estar desempenhando no discurso. Haveria então: blocos orientados para o tópico, para o texto e os orientados para o leitor.

O autor sugere que o uso extensivo de certas sequências pré-fabricadas como *it has been noted that* em gênero acadêmico, por exemplo, não apenas ajuda na identificação do registro do texto por parte dos leitores, mas também reduz o tempo de processamento pelo uso de padrões familiares para a junção de elementos de informação nova.

Em primeiro lugar os quadrigramas orientados para o tópico ajudariam aos escritores a estruturar suas atividades e experiências do mundo real. Esses quadrigramas poderiam ser ainda subdivididos em categorias como lugar, procedimento, quantificação, descrição e tópico propriamente dito, conforme listo no Quadro 3 abaixo.

	Propósito	**Exemplos**
Localização	Indicam tempo e lugar	at the beginning of, at the same time, in the present study
Procedimento	Sinalizam o procedimento ou papel de algo/alguém	the use of the, the role of the, the purpose of the, the operation of the
Quantificação	Relacionados à mensuração e valores	the magnitude of the, a wide range of, one of the most
Descrição	Possuem caráter descritivo	the structure of the, the size of the, the surface of the
Tópico	Voltados intrinsicamente para o tópico/ local	In the Hong Kong, the currency board system

Quadro 3- Quadrigramas orientados para o tópico.

Em segundo lugar, Hyland verificou a presença de uma categoria de blocos voltados para o texto. Esses estariam relacionados à organização textual e evidenciariam como seu significado pode atrelar valor às ideias e argumentos escritos relacionando, adicionando ou contrastando fatos, conforme pode ser visto no Quadro a seguir (4).

	Propósito	**Exemplos**
Marcador de transição	Estabelecem relações de adição ou contraste entre elementos e/ou ideias	On the other hand, in addition to the, in contrast to the
Marcador de resultados	Marcam relações de inferências ou causa entre elementos	As a result of it, it was found that, these results suggest that
Marcador estrutural	Marcadores reflexivos do texto que organizam pedaços do discurso ou direcionam o leitor a algum lugar no texto	In the present study, in the next section, as shown in figure
Marcador de suporte	Situam os argumentos especificando condições limitantes	In the case of, with respect to the, on the basis of, in the presence of, with the exception of

Quadro 4- Quadrigramas orientados para o texto.

A terceira e última categoria são os quadrigramas orientados para o leitor, cujo objetivo é estabelecer diálogo, uma espécie de direcionamento mais específico ao leitor do texto. Os quadrigramas orientados para o leitor estão subdivididos em dois tipos: os marcadores de postura e os de engajamento, conforme Quadro 5 a seguir.

	Propósito	**Exemplos**
Marcador de postura	Retratam a atitude a avaliação do autor	Are likely to be, may be due to, it is possible that
Marcador de engajamento	Dirige-se diretamente ao leitor	It should be noted that, as can be seen

Quadro 5: Quadrigramas voltados para o leitor.

Hyland (2008, p. 20) sugeriu com seu estudo que uma análise detalhada de *n-grams* pode oferecer 'insights' a uma dimensão crucial da análise de gênero muitas vezes ignorada. Com isso, tal análise pode vir a proporcionar um melhor entendimento das várias formas com que escritores fazem uso de recursos linguísticos em inglês em diferentes contextos acadêmicos.

Na tentativa de entender o perfil dos quadrigramas mais usados em nosso *corpus* e a função desempenhada, recorremos às categorizações maiores de Hyland (tópico, texto e leitor). Os quadrigramas mais frequentes encontrados, com 73% dos casos, foram aqueles voltados para o tópico, conforme pode ser visto na Tabela 19 abaixo.

Núm	**Quadrigrama**	**Frequência**
1	Tells the story of	38
2	The end of the	24
3	In the middle of	23
4	The story of a	22
5	One of the most	18
6	The most popular sport	17
7	The special effects are	17
8	One of the best	16

9	It tells the story	15
11	An unforgettable afternoon it	14
12	At the same time	14
14	Is a film that	13
15	Is one of the	13
16	The middle of the	13
17	To go to the	13
19	Have a lot of	12
20	I was going to	12
21	Unforgettable afternoon it was	12
23	In Rio de Janeiro	11
24	Life when I was	11
25	Most popular sport in	11
26	My life when I	11
28	A loto f things	10
30	Falls in love with	10
33	The film is set	10
34	The name of the	10
35	The plot is about	10
36	Film tells the story	9
37	For the first time	9
39	Plot is about a	9
40	The film tells the	9
42	With a lot of	9
43	Film is set in	8
45	Is a very good	8
46	Of the film is	8
48	Tells the story about	8
49	When I was a	8
50	Years ago I was	8
51	Are a lot of	7
52	Had a lot of	7
53	I was a child	7
56	It is a very	7
57	Sport in my country	7
59	The ball with the	7
60	The movie tells the	7
61	The resto f the	7
63	To go to a	7
64	Was a child I	7
65	Win the Oscar but	7
66	A lot of fun	6
67	A lot of people	6
68	A walk to remember	6
69	And in the end	6
70	And one of them	6
71	And tells the story	6
72	As the story unfolds	6
74	Beginning to the end	6
75	Changed my life because	6
76	Didn’t win an Oscar	6
77	End of the film	6

78	Event changed my life	6
79	Fall in love with	6
81	Harry Potter and the	6
85	I was in the	6
86	I went to the	6
87	In front of the	6
91	Movie tells the story	6
92	Name of the film	6
93	Of the most famous	6
96	Story of a man	6
97	The beginning to the	6
98	The main character is	6
99	The Oscar but should	6
101	There are a lot	6
103	Went to the beach	6
104	A child I used	5
105	A story about a	5
106	Afternoon it was a	5
107	Afternoon when I was	5
108	An unforgettable afternoon I	5
109	An unforgettable afternoon when	5
112	And it is a	5
113	And the most important	5
114	And you have to	5
115	Ball with the hands	5
116	Based on a book	5
117	Because it is a	5
118	Changed my life was	5
119	Child I used to	5
120	Decided to go to	5
122	Did not win the	5
123	Film is directed by	5
128	In love with a	5
129	In love with her	5
130	In my country is	5
132	In this day I	5
133	Is a movie that	5
134	is based on a	5
136	It was an unforgettable	5
138	Lot of special effects	5
141	My mother and my	5
142	Not win the Oscar	5
143	Of a group of	5
144	Of a man who	5
145	Of Rio de Janeiro	5
146	Of the movie is	5
147	Out of the blue	5
148	Players in each team	5
149	Plays the role of	5
150	Popular sport in Brazil	5

Tabela 19- Quadrigramas voltados para o tópicos.

Em segundo lugar, com 23% dos quadrigramas, estão os voltados para a interação entre escritor e leitor, conforme a Tabela 20 abaixo:

Núm	Quadrigrama	Frequência
10	Should have won the	15
13	Have won the Oscar	14
18	Changed my life when	12
22	I highly recommend it	11
27	Really changed my life	11
29	Changed my life I	10
31	Have won an Oscar	10
32	Should have won an	10
38	My point of view	9
41	What strikes you most	9
44	I would like to	8
54	I will never forget	7
58	Strikes you most is	7
62	This film should have	7
73	Asked me if I	6
80	Film should have won	6
82	I didn’t want to	6
83	I don’t know why	6
84	I have ever seen	6
88	Is a must see	6
89	Is the most popular	6
90	Me to go to	6
94	Recommend this film because	6
95	So I decided to	6
110	And I decided to	5
111	And I had to	5
121	Did a great job	5
124	I didn’t know what	5
125	I had to do	5
126	I had to go	5
127	I highly recommend this	5
135	It should have won	5
137	It’s a must see	5
139	My life because I	5
140	My life I was	5

Tabela 20- Quadrigramas voltados para o leitor.

E por último com apenas 4% os quadrigramas organizacionais, conforme podem ser vistos na Tabela 21 a seguir:

Núm	Quadrigrama	Frequência
47	On the other hand	8
55	In the end of	7
100	The other hand the	6
102	To sum up the	6
131	In the end he	5

Tabela 21- Quadrigramas voltados para a organização do texto.

Decidi elaborar um Gráfico (10) com base nas três categorias de classificação dos quadrigramas proposta por Hyland (2008), que contempla os quadrigramas voltados para o tópico do texto, para a organização textual e a interação.

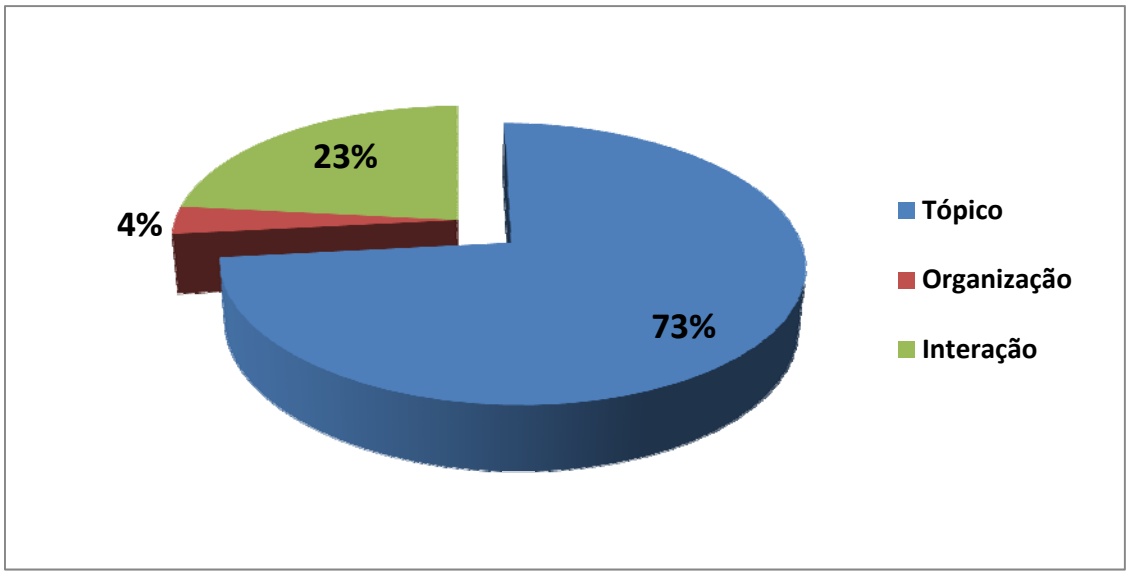

Gráfico 10- categorização dos quadrigramas proposta por Hyland.

As tabelas e gráficos parecem sugerir que os aprendizes estudados preferem os quadrigramas voltados para o tópico, ainda que grande parte do percentual a mais desses quadrigramas possa ser explicado porque alguns quadrigramas se superpõem a outros, como por exemplo, *the film tells the, film tells the story, tells the story about,* .

Neste capítulo, fiz a análise dos dados do *corpus* de estudo seguindo uma abordagem direcionada pelo *corpus*. Com isso, meu objetivo era traçar um perfil do grupo de aprendizes ao longo de seis instituições de ensino de um dos mais importantes cursos de inglês do Rio de Janeiro, a análise contemplou cerca de 400 alunos concluintes e com isso creio ter sido possível fazer um considerável aporte sobre o nível de produção linguística que se pode esperar de um aluno graduando da instituição estudada.

Passo, no próximo capítulo, a responder às perguntas de pesquisa que guiaram este estudo e a tecer as conclusões que a análise do *corpus* me proporcionaram chegar através dos dados obtidos e aqui mencionados como fruto desta pesquisa de dissertação.

CONCLUSÃO

Inicio este capítulo de conclusão com uma tentativa de responder às perguntas que deram o pontapé inicial a este estudo. Em seguida, discorro sobre as limitações do mesmo e sugiro algumas possíveis aplicações para ele. Logo depois, levanto a possibilidade de extensões futuras do mesmo.

Respondendo às perguntas de pesquisa

Antes de responder às cinco perguntas de pesquisa que foram levantadas no primeiro capítulo desta dissertação, faço uma pequena ressalva. A conclusão deste estudo deve levar em consideração que avaliar o nível de proficiência de um aprendiz de L2 apenas pela produção escrita de quadrigramas pode não ser inteiramente justo já que aprender uma língua envolve a aquisição de quatro habilidades (escrever, falar, compreensão escrita e auditiva). Como foi lembrado por Lewis (2009), mesmo o uso de numerosas expressões cristalizadas típicas da língua escrita não é indicativo de proficiência se o aprendiz as usa repetidamente. As conclusões deste trabalho, portanto, devem ser entendidas com reservas. Seguem abaixo as perguntas feitas no capítulo inicial desta dissertação e as respostas que apresento.

a) Os textos dos alunos quando submetidos a um programa de análise revelam que os aprendizes fazem uso do ‘principio idiomático’ ou do ‘princípio da livre escolha’? Dos quadrigramas que porventura aparecem, quantos foram ensinados formalmente, quantos não foram?

Como a única aula de composição escrita em que houve ênfase no ensino formal de blocos léxico-gramaticais como característicos do gênero tenha sido a resenha de filme, esta pergunta só pode ser respondida no que tange a esse gênero. Foram ensinados 20 quadrigramas na aula de resenha como sendo esses característicos do gênero e esses deveriam ser reproduzidos pelos aprendizes em seus textos durante a prova. Foram encontrados 92 quadrigramas pelo programa de computador no total. Dentre os quais somente 13 quadrigramas haviam sido ensinados. Isto quer dizer que do universo de 92 quadrigramas

produzidos somente 14% deles haviam sido ensinados. Os 76% remanescentes ou eram errôneos ou criados de acordo com o princípio da livre escolha.

Embora o grupo de aprendizes tenha começado em sua grande maioria a estudar inglês desde criança, e de acordo com o estudo de Wray (2002), aprendizes sob estas condições supostamente teriam mais facilidade em adquirir e produzir sequências multi-lexicais, parece que, no que tange à população estudada, o passar do tempo não lhes ajudou a produzi-las num contexto escrito. Os sujeitos do estudo fizeram uso do princípio da livre escolha, tendo o seu texto escrito resultado em alguns quadrigramas errôneos, quadrigramas não aprendidos e quadrigramas não prováveis de serem utilizados por falantes nativos.

Os aprendizes estudados tiveram contato com a língua alvo nas aulas de inglês através da prática do conteúdo ensinado, nesse caso os blocos léxico-gramaticais. No entanto, parece que a quantidade ou talvez a qualidade dessa prática não tenha sido o suficiente para promover as oportunidades necessárias para que houvesse a aprendizagem dos quadrigramas ensinados. Talvez os aprendizes precisariam ter tido mais tempo e também oportunidades extras tanto dentro quanto fora de sala de aula para praticar os quadrigramas e com isso adquirir a idiomaticidade da língua que se esperava com o uso de tais blocos.

Conforme já foi dito, muitos dos quadrigramas analisados não pertencem ao princípio idiomático porque este lida com expressões fixas e muitos dos quadrigramas analisados parecem ter sido criados livremente de acordo com o princípio da livre escolha. Mas dentre os aprendizes que fizeram uso do princípio idiomático não é possível afirmar se realmente houve aquisição dos quadrigramas por parte dos aprendizes ou se foram apenas memorizados a fim de serem reproduzidos durante uma avaliação. No caso desta dissertação, pôde notar-se que parece haver um vão entre a quantidade de blocos léxico-gramaticais e expressões idiomáticas que são ensinados ao longo dos anos de estudo na instituição de ensino e o que foi encontrado na versão final dos textos pesquisados. Parece ainda ser possível sugerir que os aprendizes estudados não são tão avançados quanto a instituição promete, pois supostamente estariam no nível intermediário-avançado, comparáveis ao nível do quadro comum europeu B2 / Cambridge FCE, mas apresentam ainda problemas básicos. Em estudos semelhantes ao meu, como o de Granger (2004), foi necessário salientar que as categorizações do nível de proficiência dos alunos tendem a ser vagas e a se encontrarem em descrições de cunho institucional ao invés de resultados de testes específicos. Esse parece ser um problema comum, segundo a autora, para os trabalhos com *corpora* de aprendiz.

b) Os quadrigramas ensinados formalmente são usados exatamente como foram ensinados, ou sofrem algum tipo de elaboração/transformação?

Conforme mencionado anteriormente, muitos aprendizes utilizam o princípio da livre escolha, seja transformando os quadrigramas ensinados, seja elaborando quadrigramas próprios.

Dos 20 quadrigramas ensinados formalmente na aula de resenha de filme, foram encontradas 34 variações de cinco quadrigramas ensinados (ver Tabela 7 – anexo E). Tal fato pode sugerir tanto que o aprendiz recicla os blocos léxico-gramaticais em outros contextos fazendo combinações dos blocos de formas diversas, como também pode significar uma falta de criatividade do aprendiz em fazer uso de outros blocos ou até mesmo indicar uma preferência idiossincrática dos aprendizes por certos quadrigramas em detrimento de outros. Esse último fato nos remete a questão do sobre uso de certos blocos e pouco uso de outros pelos aprendizes.

Foram também encontrados quadrigramas errôneos, sejam eles erros colocacionais ou erros coligacionais. Dessa forma, parece ser possível dizer que muitos aprendizes ainda estavam cometendo erros de tentativa ao errarem as palavras que se justapunham umas a outras formando o quadrigrama.

c) Usando-se o sistema próprio de categorização de blocos léxico-gramaticais desenvolvido por Hyland (2008), como podemos identificar que quadrigramas são usados pelos aprendizes?

De acordo com as três categorias propostas por Hyland os quadrigramas usados em sua grande maioria pelos aprendizes são aqueles voltados para o tópico do texto.

A falta de blocos léxico-gramaticais voltados para o texto causa truncamento de ideias e menos suavidade na leitura, o que provavelmente não ocorreria em um texto de falantes nativos, conforme estudo mostrado por Lewis (2009).

Era de se esperar que os aprendizes utilizassem mais quadrigramas voltados para a organização do texto. Por se tratarem de alunos avançados, supostamente teriam ciência de que um texto coeso necessita de marcadores de discurso e conectivos. Mas foram encontrados apenas 4% de quadrigramas voltados para a organização. No entanto, os aprendizes utilizaram em maior quantidade os quadrigramas voltados para o tópico (73%). Isso demonstra que os aprendizes ainda não dominam a técnica de se escrever um texto com coerência e coesão,

resultando muitas vezes em um texto repetitivo e com ideias truncadas. Parece muitas vezes ao leitor que falta fluência na leitura de muitos exemplares produzidos. Já com o uso de quadrigramas organizacionais o leitor conseguiria fazer uma leitura mais fluida, antecipando informações por vir, conectando ideias e contrastando-as. No entanto, pode-se abrir um parêntese e levantar-se duas hipóteses: 1ª) Talvez o tipo de tarefa escrita não demande do aprendiz o uso de conectivos; 2º) Pode ser que o grupo de aprendizes estudados tenham feito uso de bigramas, trigramas ou até mesmo pentagramas como marcadores de discurso. Sendo assim, um futuro estudo seria necessário para abranger essas questões que o recorte da presente pesquisa não nos permitiu analisar.

d) Considerando que os quadrigramas foram contextualizados a partir de gêneros, e que esses foram também ensinados, há alguma diferença numérica no uso dos quadrigramas nos diferentes gêneros? Em caso positivo, que gênero propicia condições maiores de utilização dos quadrigramas ensinados?

Muitos blocos léxico-gramaticais foram ensinados aos aprendizes ao longo dos sete anos de estudo na instituição, já que o ensino dos já citados blocos são uma crença da instituição e uma prática corrente durante as aulas. No entanto, cabe ao escopo deste estudo mensurar apenas os blocos léxico-gramaticais de quatro palavras produzidos neste último semestre na escrita de três gêneros textuais estudados (artigo esportivo, narrativa pessoal e resenha de filme).

Embora a LC não priorize o trabalho com números brutos mas com porcentagens, listei os 150 quadrigramas mais frequentes encontrados no *corpus* (ver anexo B – Tabela 6) a fim de fazer as comparações em números percentuais. A grande maioria dos quadrigramas encontrados, 46% foram ensinados na aula de resenha de filme. Em segundo lugar com 27% foram os quadrigramas característicos da aula de narrativa pessoal. Os quadrigramas obtidos do artigo esportivo foram minoria, apenas 5% no total. Se somarmos os três tipos de textos temos um total de 78% quadrigramas, restanto 22% de quadrigramas que não foram utilizados em um único gênero exclusivamente. Esses blocos foram reciclados, ou seja, encontrados em diversos tipos de texto ao longo de diversas filiais e aprendizes. Esse fato sugere que alguns quadrigramas não são característicos de um gênero em si, podendo ser utilizados em vários tipos de escrita ou também pode indicar que os aprendizes usam sua criatividade reciclando os blocos em diversos tipos de texto. Se observar mais a fundo esses 22% de quadrigramas que

são “reciclados”, 5% deles aparecem nos três tipos de textos escritos, conforme já fora mencionado no Gráfico 6 (página 72).

O gênero no qual os aprendizes utilizam mais blocos lexicais que foram ensinados foi a resenha de filme, mas este fato não deve ser levado em consideração pois é a aula em que mais constam blocos lexicais que devem ser utilizados como forma de se construir tal gênero.

Os motivos por esta preferência fogem ao alcance deste estudo, porém alguns questionamentos podem ser feitos em relação aos dados encontrados. Em primeiro lugar, familiaridade pessoal com o tópico narrativa pode ter influenciado os alunos a escolherem tal tema, pois desde criança sempre contamos e ouvimos histórias. Logo, os alunos podem imaginar que escrever uma história seria apenas uma réplica de contar uma história de forma oral. Em relação ao fato de a maioria dos aprendizes ter escolhido a resenha de filme, isso pode vir a ser explicado porque muitos dos alunos parecem ter entendido a estrutura retórica da resenha de filme como um simples resumo do enredo, isto é, uma outra narrativa sem a parte opinativa (ver anexo F – Quadros 6 e 7).

Com relação ao artigo, o tema da redação era: “escreva um artigo sobre o esporte favorito dos brasileiros”. Como era de se esperar, quem optou pelo artigo escreveu sobre futebol, por ser este a ‘paixão nacional’, mas somente 6.9% optaram pelo artigo. Pode-se supor que escrever um artigo requer uma habilidade maior, não sendo tão simples quanto uma narrativa ou uma resenha de filme. Talvez não seja parte da prática social dos aprendizes estudados escrever artigos, daí a grande objeção pelo genero em questão. Em segundo lugar, exposição a experiências com o assunto em língua materna pode servir de base para a escrita de um determinado gênero, em especial gêneros narrativos, já que os aprendizes têm bastante conhecimento prévio do assunto em L1 para escrever a narrativa em L2 (WANG e WEN, 2002). Com isso, o conhecimento prévio de um determinado assunto e a exposição a experiências em língua materna podem ter influenciado também a escolha dos alunos pelos dois principais gêneros escritos – a resenha de filme (já que esta em muitos casos se assemelhou a uma narrativa do filme) e a narrativa pessoal.

E em terceiro lugar, pode-se dizer que maior ênfase fora dada ao ensino de narrativas e resenhas pessoais durante as aulas de produção escrita na instituição. Como professor da mesma e tendo feito a análise do livro didático, do programa curricular da instituição para as aulas de composição escrita e tendo ministrado aulas para algumas turmas do nível analisado, posso afirmar que os alunos praticam como escrever uma história usando os tempos verbais no passado, que tipo de informação desenvolver em cada parágrafo, assim como o léxico que devem utilizar. Assim como nessa aula sobre narrativas, na aula onde aprendem a escrever a

resenha de filme, os alunos aprendem os cinco movimentos retóricos que compõem a resenha e os blocos lexicais que devem vir em cada um dos movimentos. Já em relação ao artigo, os alunos apenas leem um exemplar prototípico do gênero e o professor pede que escrevam um exemplar semelhante em casa, não tendo oportunidades de prática em sala de aula. Esses fatos mencionados anteriormente podem ter vindo a colaborar para uma maior produção de dois dos gêneros coletados para esta análise.

e) A partir dos resultados finais, é possível se formular alguma conclusão sobre a pertinência de ensinar quadrigramas para posterior testagem?

Parece ter ficado claro com este estudo que ensinar não necessariamente significou aprender. Long e Doughty (2005) corroboram a ideia que a instrução formal auxilia o processo de aprendizagem em uma L2, através do estímulo dado e ao *feedback* recebido. No entanto, o aprendiz não irá produzir se não houver maiores oportunidades de prática e uma motivação intrínseca para aprender.

Com isso, pode se sugerir que ensinar para testar o conhecimento não demonstrou ser uma prática válida. Não adiantou, no caso deste, estudo ensinar e testar. Os aprendizes precisariam de mais tempo e oportunidades (e menos ansiedade) para adquirir os quadrigramas ensinados.

Também é importante salientar que 79% do total de quadrigramas encontrados no *corpus* de análise foram usados apenas uma vez pelos aprendizes. 21% dos quadrigramas foram usados duas ou mais vezes por aprendizes diferentes. Esse fato revela que o uso se concentra em um reduzido número de quadrigramas pelos quais os aprendizes parecem ter certa preferência.

Ainda sobre a aprendizagem dos quadrigramas ensinados, não é de se negar que seria de grande valia o ensino de blocos léxico-gramaticais como forma de fazer com que o texto do aprendiz ganhe maior fluência na L2. No entanto, conforme já mencionado, não basta apenas ensinar para testar. Pois é preciso que haja, entre o ensino e o teste, oportunidades de prática e que o conteúdo a ser aprendido seja significativo para o aprendiz. Será que os aprendizes em questão se sentem motivados a aprender a escrever uma resenha de filme a ponto de adquirirem a língua necessária para escrever tal gênero?

As limitações do estudo

Esta pesquisa não é definitiva porque não tive total controle sobre as condições de aprendizagem e repetição, mas somente sobre os números que mostram que parece não ter adiantado muito ensinar os quadrigramas mencionados neste estudo.

Todavia, ensinar pode sim ser algo bom e prazeroso contanto que o aluno encontre uma razão para aprender. A aprendizagem precisa ser significativa tanto no nível micro (quando falo dos quadrigramas) quanto no macro (ao analisar os gêneros textuais).

As questões discutidas anteriormente me fazem voltar à Figura 3 (p. 26) de Spolsky (1989) que demonstra que, no fim de tudo, as oportunidades de aprendizagem (pelo menos no que diz respeito a uma língua estrangeira) ainda irão desembocar nas razões que o aprendiz tem para aprendê-la.

Isso me fez questionar, por exemplo, sobre a pertinência de se ensinar determinados gêneros textuais como a resenha de filme e porque maior ênfase não fora dada aos blocos léxico-gramaticais que são característicos de uma narrativa pessoal, já que esta parece ser uma prática social mais comum entre os aprendizes.

Aplicabilidade do estudo

Este estudo contribui para auxiliar professores de inglês a preparar suas aulas de escrita em L2. Assim, professores podem ter uma maior compreensão da importância de se ensinar blocos léxico-gramaticais para garantir fluência e acuracidade no idioma. Foi notável a importância de se dar mais oportunidades de prática ao aluno. Tempo maior para assimilação e prática devem ser dados ao aprendiz tanto dentro quanto fora da sala de aula. Além disso, o estudo sugere que ensinar a escrita de gêneros textuais que não fazem parte da prática social do aprendiz pode resultar em textos com características de outros gêneros, com os quais os alunos estejam mais familiarizados. Com isso, pode haver um questionamento sobre a validade de se ensinar para adolescentes de cerca de 17 anos o porquê de se escrever um gênero textual que os mesmos não mais terão que fazer uso em oportunidades futuras. Diferentemente da narrativa pessoal, todos nós contamos e escrevemos histórias com muito

mais frequência que artigos esportivos ou resenhas de filme. Ou seja, a história é um arqui-gênero que permeia muitas outras escolhas textuais.

O estudo também contribui para o treinamento de professores, não apenas da instituição de ensino investigada, mas outros coordenadores e treinadores de professores podem se beneficiar dos direcionamentos apontados por esta pesquisa a fim de treinar educadores e otimizar a prática docente em aulas de composição escrita.

Pesquisas futuras

Entre possíveis pesquisas futuras que este estudo poderia se estender estão: aumentar o *corpus* de estudo com textos escritos em um tempo futuro já que a instituição de ensino está adotando mudanças no material didático. Com tal análise, poderia se investigar quais gêneros textuais estão sendo ensinados e se os blocos léxico-gramaticais ensinados são os mesmos.

Outro estudo possível seria comparar os resultados obtidos nesta dissertação com outros *corpora* de aprendizes brasileiros de diferentes instituições de ensino. Assim, talvez fosse possível avaliar a "brasilidade" do inglês dos alunos em questão. Poderia haver traços que todos os aprendizes brasileiros carregam ao aprender inglês e que sejam visíveis na produção escrita.

Também aponto para um possível estudo contrastando o uso de blocos léxico-gramaticais com a nota que o aluno tirou na redação. Com isso, poderia ser averiguado se o professor considera o uso dos blocos como sendo fruto de uma redação bem escrita atribuindo a esta uma nota maior ou não.

Seria também interessante, uma maior investigação sobre os gêneros textuais que os alunos preferem escrever e aqueles que os mesmos têm maior facilidade e dificuldade em fazer uso. Com isso, maior ênfase poderá vir a ser dada às aulas de produção escrita na instituição de ensino estudada, a fim de que os alunos tenham mais oportunidades de prática em sala de aula dos gêneros textuais em questão.

Todas as sugestões mencionadas acima não visam apenas à descrição linguística, mas também a auxiliar o professor a melhorar sua prática docente.

Considerações finais

Para concluir, faço algumas reflexões pessoais acerca do estudo. A primeira delas é acerca do método escolhido. Comecei o anteprojeto de pesquisa desta dissertação visando analisar os erros produzidos pelos aprendizes, pois tinha uma suposição inicial que os alunos avançados da instituição em questão produziam muitos blocos léxico-gramaticais errôneos, já fossilizados e que não bastaria o professor corrigi-los. Para mim, os erros já estariam internalizados e esses seriam evidentes nas produções escritas.

No entanto, a LC, a partir da compilação e extração de uma grande quantidade de textos, me fez perceber que minha percepção inicial não se confirmou. Os erros encontrados foram isolados e não podem ser generalizados como sendo constituintes do perfil do aluno da instituição de ensino. Alguns erros podem ser encontrados com mais frequência em alguns aprendizes que em outros, mas ao olharmos o global, vê-se que o aprendiz acerta mais que erra, ou seja, ele aprende a L2, mesmo não produzindo tantos blocos quanto os ensinados.

Com isso, posso dizer que o método de pesquisa utilizado para análise dessas redações foi apropriado, pois, embora tenha mirado nos erros, acertei nos quadrigramas, o que me permitiu visualizar a produção dos aprendizes sob outro viés. Espero ter feito uma contribuição significativa para os estudos sobre a léxico-gramática na escrita de aprendizes brasileiros de inglês como L2.

REFERÊNCIAS

ALMEIDA, M. I. *Prosa argumentativa em língua inglesa:* um estudo contrastivo sobre advérbios em *corpora* digitais. 2010. 152 f. Dissertação (Mestrado em Linguística) - Instituto de Letras, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2010.

BAKHTIN, M. *Estética da criação verbal*. São Paulo: Martins Fontes, 2003. p. 279-326.

BERBER-SARDINHA, T. *Linguística de corpus*. Barueri, SP: Manole, 2004.

_____; SHEPHERD, T. M. G. An online system for error identification in Brazilian learner English. In: TEACHING AND LANGUAGE CORPORA CONFERENCE, 8., 2008, Lisboa. *TaLC8*. Lisboa: Associação de estudos e de investigação do ISLA, 2008. p. 257-263.

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

COUNCIL OF EUROPE. Common European framework of reference for languages: learning, teaching, assessment. [S.l.]: Council of Europe, 2011.

COOK, Vivian; LONG, J.; McDONOUGH, Steven. First and second language learning. In: PERREN, G. E. (Org.). *The mother tongue and other languages in education.* London: CILTR, 1979. p. 1-14.

CUNHA, A. F. da. Funcionalismo. In: MARTELOTTA, M. E. (Org.) *Manual de linguística*. São Paulo: Contexto, 2010. p. 157-174.

DUTRA, D. P. *Agrupamentos lexicais na escrita de aprendizes brasileiros de inglês*: um estudo baseado em *corpus*. Plano de trabalho apresentado ao Programa Pesquisador Mineiro. Edital FAPEMIG 03/2010.[S.l.:s.n., 200-].

_____. BERBER-SARDINHA, T. *Pacotes lexicais em corpora de aprendizes*.Trabalho apresentado no ELC. Porto Alegre: [s.n.], 2010.

ELLIS, R. *Second language acquisition*. Oxford: Oxford University Press, 2000.

ERDOGAN, V. Contribution of error analysis to foreign language teaching. *Mersin University Journal of Faculty Education*. Turquia, v. 1, ed. 2, p. 261-270, dez. 2005.

FERRIS, D. R. *Treatment of error in second language student writing.* Michigan: The University of Michigan Press, 2002.

FIRTH, J. R. A synopsis of linguistic theory, 1930-1955. In: STUDIES in linguistic analysis: special volume of the philological society. (1-32). Oxford: Blackwell. Reimpresso em Palmer, F. R. (Ed.). Selected papers of J. R. Firth 19-52-59. p. 168-205. Harlow: Longman, 1957-1968.

GARDNER, R.; LAMBERT, W. *Attitudes and motivation in second language learning*. Rowley, Mass: Newbury house, 1972.

GRANGER, S. Prefabricated patterns in advanced EFL writing: collocations and Formulae. In: COWIE, A. P. (Ed.) *Phraseology: theory, analysis and applications.* Oxford: Oxford University Press, 1998, p. 145-160.

_____. HUNG, J.; PETCH-TYSON, S. (Ed.). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam/Philadelphia: Benjamins, 2002.

_____. Error-tagged learner *corpora* and CALL: a promising synergy. *CALICO Journal*. San Marcos, TX, v. 20, n. 3, p. 1-16, 2003.

_____. Computer learner *corpus* research: current status and future prospects. In: CONNOR, U. *Applied corpus linguistics:* a multidimensional perspective. Amsterdam and Atlanta: Rodopi, 2004. p. 123-145.

_____. Using error-tagged learner corpora to create English-specific CEF descriptors. In: CONFERENCE OF EALTA, 3., 2006. [S.l.: s.n., 2006].

GURGEL, M. C. L. A linguagem: história, língua e natureza. In: CARNEIRO, M. (Org.). *Pistas e travessias*. Rio de Janeiro: EdUerj, 1999. p. 11-28.

HARMER, J. *How to teach English*. Essex, UK: Pearson, 2007.

HOEY, M. *Lexical priming:* a new theory of words and language. London: Routledge, 2005.

HYLAND, K. As can be seen: lexical bundles and disciplinary variation. *ESP,* v. 27, p. 4-21, 2008.

KENEDY, E. Gerativismo. In: MARTELOTTA, M. E. (Org.). *Manual de linguística*. São Paulo: Contexto, 2010. p.127-139.

KRASHEN, S. *Principles and practice in second language acquisition.* California: University of Southern California, Pergamon Press, 1982.

_____. *Language acquisition*. Disponível em: <www.goo.gl/nw0rg>. Acesso em: 26 dez. 2012.

LEECH, G. Preface. In: Granger. S. (Ed.). *Learner English on computer.* London & New York: Addison Wesley Longman, 1998. p. 14-20.

LEWIS, M. *The idiom principle in L2 english*. Stockholm: Stockholm University, 2009.

LEWIS, M. *The lexical approach*. Hove: Language Teaching Publications, 1993.

LONG, M.; DOUGHTY, C. *The handbook of second language acquistion.* Australia: Blackwell, 2005.

LÓPEZ, W. C. Error analysis in a learner corpus: what are the learners' strategies? In: A SURVEY on corpus-based research. Panorama de investigaciones bas**e**adas en *corpus*. AELINCO 2009. Disponível em: < http://www.um.es/lacell/aelinco/contenido/index >. Acesso em: 05 jul. 2013.

McENERY, T.; WILSON, A. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

PETTER, M. Linguagem, Língua, Linguística. In*:* FIORIN, J. L. (Org.). *Introdução à linguística. I:* objetos teóricos*.* São Paulo: Editora Contexto, 2011. p. 11-24

SPRATT, M.; PULVERNESS, A.; WILLIAMS, M. *The TKT course*. Cambridge: Cambridge University Press, 2005.

SCHUTZ, R. *Interlíngua e Fossilização*. English made in Brazil, 2005. Disponível em: <http://www.sk.com.br/sk_interfoss.html >. Acesso em: 28 dez. 2012.

SELINKER, L. Interlanguag*e*. *IRAL*. p. 209-231, 1972.

SHEPHERD, T. M. G. Panorama da linguística de corpus. In: _____; BERBER-SARDINHA, T.; PINTO, M. V. (Org.). *Caminhos da linguística de corpus*. Campinas, SP: Mercado de Letras, 2012.

_____. Corpora de aprendiz de língua estrangeira: um estudo contrastivo de n-gramas. *Veredas on-line*, Juiz de Fora, MG, ed. 2, p. 100-116, 2009.

_____; ZYNGIER, S.; VIANA, V. Feixes lexicais e visões de mundo: um estudo sobre *corpus*. *Matraga*, ano 13, n. 19, p. 125-140, 2009.

SINCLAIR, J. *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991.

SINCLAIR, J. Preface. In: LEWANDOWSKA-TOMASZCZYK, B. (Ed.). *Practical applications in language and computers*. Frankfurt: Peter Lang, 2004. p. 7-11.

_____. *Corpus* and text-basic principles. In: WYNNE, M. (Ed.) *Developing linguistic corpora:* a guide to good practice. Oxford: Oxbow Books, 2005, p. 1-16. Disponível em: < http://goo.gl/JA6uFX > Acesso em: 04 ago. 2013.

SPOLSKY, B. *Conditions for second language learning*: introduction to a general theory. Oxford: Oxford University Press, 1989. p. 28.

SKINNER, B. F. *Sobre o behaviorismo*. São Paulo: Cultrix, 1974.

THORNBURY, S. *How to teach grammar*. Essex, UK: Pearson, 2008.

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

UR, P. *A course in language teaching*. Cambridge: Cambridge University Press, 1996.

WANG, Wenyu; WEN, Qiufang. L1 use in the L2 composing process: an exploratory study of 16 Chinese EFL writers. *Journal of second language writing.* v. 11, ed. 3, p. 225-246, 2002.

WIDDOWSON, H. The limitations of linguistics applied. *Applied Linguistics*, v.21, n.1, p. 3-25, 2000.

WILKINS, D. A. *Linguistics and language teaching*. London: Edward Arnold., 1972.

_____. *Second language learning and teaching*. London: Edward Arnolds, 1976.

WRAY, A. *Formulaic language and the lexicon.* Cambridge UK: Cambridge University Press, 2002.

YIP, V. *Interlanguage and learnability / from Chinese to English*. USA: John Benjamins Publishing ,1995.

YORIO, C. A. Idiomaticity as an indicator of second language proficiency. In: HYLTENSTAM, K.; OBLER, L. (Ed.). *Bilingualism across the lifespan*. Cambridge: Cambridge University Press, 1989. p. 55-72.

ANEXOS

Anexo A – Quadro 2: programa curricular da instituição de ensino para o ensino de resenha de filme.

Habilidade	Tipo de texto	Tarefa	Léxico	Quadrigramas ensinados
Produção escrita	Resenha de filme	Escreva uma resenha sobre um filme que deveria ter ganhado um Oscar mas não ganhou. / Escreva uma resenha que você enviará a uma locadora de filmes. Você poderá mencionar: o enredo, o elenco, os efeitos especiais (se necessário), porque você gostou do filme, etc.	visual imagery, atmosphere, acting, plot, performance, ending, cast, scene, slow, special effects, set, soundtrack, screenplay, stunning, first-rate, convincing, predictable, adapted, reveal, suspend, direct, strike, stars.	The latest film by The Oscar-winning film Adapted from a novel In the title role The film is set Tells the story of The story takes place As the story unfolds What strikes you most Creates an atmosphere of Acting is first-rate I was impressed by Not a gripping film The plot is quite The ending is predictable What the film lacks Characters fail to move It's well-worth seeing I highly recommend it It doesn't quite deliver

Anexo B – Tabela 6: os 150 quadrigramas mais frequentes do *corpus* de estudo.

Núm	Quadrigramas	Frequência	Textos
1	Tells the story of	38	38
2	The end of the	24	22
3	In the middle of	23	21
4	The story of a	22	22
5	One of the most	18	17
6	The most popular sport	17	16
7	The special effects are	17	17
8	One of the best	16	16
9	It tells the story	15	15
10	Should have won the	15	15
11	An unforgettable it	14	14
12	At the same time	14	14
13	Have won the Oscar	14	14
14	Is a film that	13	13
15	Is one of the	13	13
16	The middle of the	13	13
17	To go to the	13	13
18	Changed my life when	12	12
19	Have a lot of	12	12

20	I was going to	12	11
21	Unforgettable afternoon it was	12	12
22	I highly recommend it	11	11
23	In Rio De Janeiro	11	9
24	Life when I was	11	11
25	Most popular sport in	11	11
26	My life when I	11	11
27	Really changed my life	11	10
28	A lot of things	10	9
29	Changed my life I	10	10
30	Falls in love with	10	9
31	Have won an Oscar	10	9
32	Should have won an	10	9
33	The film is set	10	10
34	The name of the	10	9
35	The plot is about	10	10
36	Film tells the story	9	9
37	For the first time	9	7
38	My point of view	9	9
39	Plot is about a	9	9
40	The film tells the	9	9
41	What strikes you most	9	9
42	With a lot of	9	9
43	Film is set in	8	8
44	I would like to	8	7
45	is a very good	8	8
46	of the film is	8	7
47	on the other hand	8	8
48	tells the story about	8	8
49	When I was a	8	8
50	Years ago I was	8	7
51	are a lot of	7	6
52	had a lot of	7	7
53	I was a child	7	7
54	I will never forget	7	7
55	In the end of	7	7
56	It is a very	7	5
57	Sport in my country	7	7
58	Strikes you most is	7	7
59	The ball with the	7	7
60	The movie tells the	7	7
61	The resto f the	7	7
62	This film should have	7	7
63	To go to a	7	7
64	Was a child I	7	7
65	Win the Oscar but	7	7
66	A lot of fun	6	6
67	A lot of people	6	6
68	A walk to remember	6	5
69	And in the end	6	6
70	And one of them	6	6

71	And tells the story	6	6
72	As the story unfolds	6	6
73	Asked me if I	6	6
74	Beginning to the end	6	5
75	Changed my life because	6	6
76	Didn’t win an Oscar	6	4
77	End of the film	6	6
78	Event changed my life	6	6
79	Fall in love with	6	6
80	Film should have won	6	6
81	Harry Potter and the	6	5
82	I didn’t want to	6	6
83	I don’t know why	6	6
84	I have never seen	6	6
85	I was in the	6	6
86	I went to the	6	6
87	In front of the	6	6
88	Is a must see	6	6
89	Is the most popular	6	6
90	Me to go to	6	6
91	Movie tells the story	6	6
92	Name of the film	6	6
93	Of the most famous	6	6
94	Recommend this film because	6	6
95	So I decided to	6	6
96	Story of a man	6	6
97	The beginning to the	6	5
98	The main character is	6	6
99	The Oscar but should	6	6
100	The other hand the	6	6
101	There are a lot	6	5
102	To sum uo the	6	6
103	Went to the beach	6	5
104	A child I used	5	5
105	A story about a	5	5
106	Afternoon it was a	5	5
107	Afternoon when I was	5	5
108	An unforgettable afternoon I	5	5
109	An unforgettable afternoon when	5	5
110	And I decided to	5	5
111	And I had to	5	5
112	And it is a	5	5
113	And the most important	5	5
114	And you have to	5	5
115	Ball with the hands	5	5
116	Based on a book	5	5
117	Because it is a	5	5
118	Changed my life was	5	5
119	Child I used to	5	5
120	Decided to go to	5	5

121	Did a great job	5	3
122	Did not win the	5	5
123	Film is directed by	5	5
124	I didn't know what	5	5
125	I had to do	5	5
126	I had to go	5	5
127	I highly recommend this	5	5
128	In love with a	5	5
129	In love with her	5	4
130	In my country is	5	5
131	In the end he	5	4
132	In this day I	5	5
133	Is a movie that	5	5
134	Is based on a	5	5
135	It should have won	5	4
136	It was an unforgettable	5	5
137	It's a must see	5	5
138	Lot of special effects	5	5
139	My life because I	5	5
140	My life I was	5	5
141	My mother and my	5	5
142	Not win the Oscar	5	5
143	Of a group of	5	5
144	Of a man who	5	5
145	Of Rio de Janeiro	5	4
146	Of the movie is	5	5
147	Out of the blue	5	5
148	Players in each team	5	5
149	Plays the role of	5	4
150	Popular sport in Brazil	5	5

Anexo C - Quantidade de quadrigramas ensinados e encontrados no *corpus*.

Quadrigramas ensinados	**Quant. de quadrigramas encontrados**
Tells the story of	38
I highly recommend it	11
The film is set	10
What strikes you most	9
it's well worth seeing	4
The latest film by	4
The ending is predictable	3
I was impressed by	2
acting is first rate	2
in the title role	2
Adapted from a novel	2
The story takes place	2

Anexo D - Tabela 7: Quadrigramas ensinados *versus* modificados pelos aprendizes.

Ensinado	**Variação**	**Quantidade**
In the title role	The main role is	2
-	Playing the main role	2
The film is set	The movie is set	5
-	It is set in	4
-	The plot is set	3
-	The story is set	2
Tells the story of	Tells the history of	3
-	Tells the history about	2
It's well-worth seeing	A well Worth seeing	2
	It is well worth	2
	Say it's well Worth	2
I highly recommend it	I highly recommend this	5
	TOTAL	34

Anexo E - Exemplo de resenha de filme como resumo do enredo.

Resenha 1:

<header>

<branch>T</branch>

<timed>yes</timed>

<sex>F</sex>

<type of text>filmreview</type of text>

<date>2010.1</date>

<teacher>YM</teacher>

<grade>8.5 </grade>

<title>Moulin Rouge</title>

</header>

Moulin Rouge is a film that was set in Paris during the beauty epic, on a house of shows called the Moulin Rouge. Tells a story about a courtesan named Satine that falls in love with a poor writer. Even she could not fall for anyone, they live the most beautiful story of love, a prohibited love. It's a music with a lots of music of Elton John, Madonna. There is a lot of dancing and the set has every little detail to show how it really was. Satine is played by Nicole Kidman. The atmosphere is absolutely stunning! The film holds you from the moment that it starts till the end. It's not a happy story about love, because Satine has a disease that

kills her on the end. What I lost like about that film is even he knowing she was a courtesan and couldn't be with him, she had to work on that because of money, he didn't give up of her. He fight for love till the end, just like Romeo and Juliet story. All the special effects about the place where they dance and sing are amazing and incredible. Want a film about love? Definitely, Moulin rouge is the one.

Quadro 6 : Exemplo 1 de resenha de filme com características de uma narrativa.

Resenha 2:

Resenha 1:

<header>

<branch>N</branch>

<timed>yes</timed>

<sex>M</sex>

<type of text>filmreview</type of text>

<date>2010.2</date>

<teacher>GS</teacher>

<grade>7.0 </grade>

<title>A film that that did not win an Oscar but should have</title>

</header>

The film I would like to talk about is the fast and Furious #1, the first film of the series. There are a total of four films, but the first one is the best. The main actor, Paul Walker, as Brian, did a great job in this film. Not only exciting scenes, but funny ones too. His story is about a cop that investigates street race drivers. During the film, Brian gets involved with Mia, Don's sister. Don (Vin Diesel) is a car stealer and Brian has to catch him, so, now you can imagine how the film goes on. My favourite part is the sports cars used in the film, models like Toyota Supra, Nissan Skyline, Mitsubishi Eclipse and even Ferraris - bringing a special taste to this film. So, if you doesn't know the end of the film, go see it at our school on our cinema night. I'm waiting for you guys, come see it.

Quadro 7: Exemplo 2 de resenha de filme com características de uma narrativa.