PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO (PUC-SP)

Wendel Mendes Dantas

ERROS DE ESCRITA EM INGLÊS POR BRASILEIROS: IDENTIFICAÇÃO, CLASSIFICAÇÃO E VARIAÇÃO ENTRE NÍVEIS

MESTRADO EM LINGUÍSTICA APLICADA E ESTUDOS DA LINGUAGEM

SÃO PAULO

2012

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO (PUC-SP)

Wendel Mendes Dantas

ERROS DE ESCRITA EM INGLÊS POR BRASILEIROS: IDENTIFICAÇÃO, CLASSIFICAÇÃO E VARIAÇÃO ENTRE NÍVEIS

Dissertação apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de MESTRE em Linguística Aplicada e Estudos da Linguagem, sob orientação do Prof. Dr. Antonio Paulo Berber Sardinha

MESTRADO EM LINGUÍSTICA APLICADA E ESTUDOS DA LINGUAGEM

SÃO PAULO

2012

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO

2012

BANCA EXAMINADORA

_____

_____

_____

Dedico este trabalho a todos os que lutam solitários em busca dos seus sonhos e à minha sogra Maria Benedicta de Oliveira (*in memoriam*).

"On to wake up he said to the woman: Listen, my girl: today is day of to pay the instalment of television, comes here the subject with the bill, on certain. But happens that yesterday I no brought money from the city, am on none [slang].

– Explain this to the man – answered the woman. – No like of these things. Gives an air of swindling, I like of to fulfil rigorously the my obligations."

(SABINO, F. O Homem Nu, In: Shepherd, 2001)

AGRADECIMENTOS

A Deus, por ter me dado saúde e a oportunidade de estudar.

Ao Prof. Dr. Antonio Berber Sardinha (Tony) pelo apoio desde o início, num momento de falta de autoconfiança, críticas e desvalorização no mercado de trabalho. Agradeço também pela paciência e prontidão nas orientações e por ter me introduzido à Linguística de Córpus.

À Profª Drª Rosinda de Castro Guerra Ramos, que me ajudou com preciosas dicas de leitura e estruturação, e à Prof. Drª Maria Cecília Pérez de Souza e Silva (Cecilinha) pela simpatia, simplicidade, acolhimento e introdução à Análise do Discurso.

Às Professoras Drª Maria Cecília Lopes (Ciça) e Drª Patrícia Bértoli-Dutra (Pat) pela simpatia e por terem aceitado fazer parte da banca examinadora.

À colega e tutora Marcia Veirano Pinto, orientanda de doutorado do Prof. Tony, por ter sempre se mostrado disponível a me ajudar nos momentos de dúvida, por ter revisado este trabalho e me direcionado durante meu mestrado. Agradeço também pelo trabalho como co-avaliadora para a verificação de minha metodologia de análise.

Ao colega Etelvo Ramos Filho pela ajuda com o projeto, pela primeira avaliação de concordância entre avaliadores e pelas conversas sempre agradáveis.

Ao Prof. Dr. Douglas Biber por ter sido um modelo de simplicidade e acessibilidade, e pelas dicas dadas ao assistir minha apresentação no ELC 2010.

Ao grupo de orientandos do Prof. Tony pelo ambiente descontraído e de colaboração mútua.

À Debora Schisler, diretora do departamento de engenharia de educação da escola na qual trabalho, pelo apoio irrestrito ao projeto e pela permissão de uso do sistema *online*.

À Capes pela bolsa, que possibilitou a materialização deste sonho tão importante.

Às funcionárias do CEPRIL e LAEL, Márcia e Maria Lúcia, e aos atendentes da secretaria de pós-gaduação pelo auxílio constante.

A todos os amigos que sempre me deram forças e acreditaram em mim e a todos aqueles que, direta ou indiretamente, colaboraram com este trabalho.

E, por fim, a todos aqueles que não acreditaram em mim e tentaram impor barreiras à realização deste projeto, pois foram rochas a serem transpostas, sobre as quais pude vislumbrar um mar de novas possibilidades.

RESUMO

DANTAS, Wendel Mendes. **Erros de escrita em inglês por brasileiros: identificação, classificação e variação entre níveis**. 2012. 164f. Dissertação (Mestrado) – Pontifícia Universidade Católica de São Paulo, São Paulo, 2012.

O trabalho tem como objetivo identificar e classificar os erros na escrita de aprendizes brasileiros de inglês. As perguntas que norteiam a pesquisa são: "Quais os erros mais comuns no córpus COBRA-7_recorte?"; "Qual a variação de erro entre os níveis de curso dos aprendizes no córpus COBRA-7_recorte?" e "Qual nível de curso apresenta maior diversidade de erros no córpus COBRA-7_recorte?". Esta pesquisa encontrou suporte teórico na Linguística de Córpus, área que se dedica à coleta e análise criteriosa de dados de textos em formato digital, e especificamente nas pesquisas dedicadas a córpus de aprendizes. Os córpora empregados na pesquisa foram o *Corpus of Contemporary American English* (COCA) (córpus de consulta) e uma amostra do COBRA-7 (córpus de estudo), compilado a partir de redações de aprendizes adultos matriculados em uma rede de escolas de inglês como língua estrangeira do estado de São Paulo, produzidas entre 2009 e 2010. Os dados foram coletados de um servidor *online* da própria instituição em 2011. Os resultados indicaram que os erros mais comuns encontrados no córpus de análise referem-se a: má escolha lexical, uso de tempo e aspecto verbal, uso de determinantes, e uso inadequado de questões, negações ou auxiliares. Revelaram também que o nível de curso pré-intermediário apresenta as maiores quantidade e diversidade de erros, provavelmente por se tratar de um nível no qual os aprendizes são expostos a tempos verbais diversos. Por fim, mostraram que embora a má escolha lexical, sobretudo a substituição de preposições ou conjunções por outras ou pelas mesmas classes gramaticais constitua um problema para os aprendizes, essa dificuldade diminui ao longo do curso, ao contrário do erro no uso dos tempos e aspecto verbais, que tende a aumentar.

Palavras-chave: córpus de aprendizes, ensino de idiomas, análise de erros, concordância entre avaliadores, linguagem como sistema probabilístico

ABSTRACT

The aim of this study was to identify and classify errors found in Brazilian English learners's written tasks. The questions which guide this research are: "Which are the most common errors in COBRA-7_recorte?"; "What is the error variation among course levels for learners found in COBRA-7_recorte?" and "Which course level shows the highest error diversity in COBRA-7_recorte?". The main theoretical underpinning for the research is provided by Corpus Linguistics, an area devoted to the collection and criterious analysis of data collected from texts in electronic formar, and, specifically, by research on learner corpora. the corpora used in this research were the Corpus of Contemporary American English (COCA) (consultation corpus) and a sample of COBRA-7 (COBRA-7_recorte), the study corpus, compiled from the writings of adult learners enrolled in a network of schools which teach English as a foreign language in the state of São Paulo, and which have been produced between 2009 and 2010. The data have been collected from the institution's online server in 2011. Results have shown that the most common errors found in the analysis corpus are: wrong lexical choice, tense and aspect use, use of determiners, and wrong use of questions, negatives or auxiliaries. They have also revealed that pre-intermediate course level has the highest quantity and diversity of errors, probably because it is a level at which learners are exposed to diverse verbal tenses. Finally, this study has also shown that although wrong lexical choice, particularly the replacement of prepositions or particiles by words from the same or other grammatical categories, seem to be a problem for learners, this difficulty tends to decrease along the course, unlike errors of verbal tense and aspect use, which tend to increase.

Keywords: learner corpus, language learning, error analysis, inter-rater reliability, language as probabilistic system

SUMÁRIO

LISTA DE QUADROS, FIGURAS E TABELAS

Quadros

Figuras

Tabelas

Introdução

Há aproximadamente dez anos anos venho atuando na sala de aula de cursos de idiomas como professor. Uma das tarefas mais comuns nessa atividade, como se sabe, é a solicitação e a correção de redações sobre temas diversos. Aguçava-me a curiosidade, tanto sob um quadro teórico estruturalista (baseado em exercícios do tipo *fill-in-the-blanks*) quanto sob uma abordagem comunicativa (com atividades com foco em funções, por exemplo) o fato de encontrar erros muito semelhantes, naturalmente relacionados a níveis de curso[1] específicos, em qualquer das escolas nas quais eu tenha atuado profissionalmente. Tais repetições, como é o caso da troca do *there is* por *have*, o uso da preposição *of* após verbos como *like* e *need*, e mesmo o uso de *same* no lugar de *really* para indicar ênfase, levaram-me a questionar a eficiência das correções dessas redações e a absorção das orientações que eram dadas aos aprendizes com relação às suas produções. Constatei que as redações, mesmo sendo uma importante fonte de dados linguísticos, em geral não eram revisadas pelos aprendizes após terem sido corrigidas e entregues. Desse modo, a correção parecia, em muitos casos, perder seu objetivo, que é mostrar ao aprendiz em que deve melhorar.

Ao revisar a literatura, descobri estudos anteriores com foco nos erros cometidos por aprendizes, dentre os quais posso citar Lado (1957), Selinker (1972), Corder (1981), Granger (1998, 2002, e 2008), e Shepherd (2001).

Como estava interessado em analisar um grande número de textos, era-me necessário um arcabouço teórico que me permitisse categorizá-los e analisá-los. Por isso, este trabalho encontra suporte teórico também na Linguística de Córpus[2]. A Linguística de Córpus é uma área que

> ocupa-se da coleta e da exploração de córpora, ou conjuntos de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou uma variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador.
>
> (Berber Sardinha, 2004, p.3)

[1] Chamo de nível de curso o nível de "proficiência" adotado internamente pelas escolas de idiomas, como básico 1, básico 2, pré-intermediário etc. Optei por essa denominação porque o termo nível de proficiência parece indicar um conceito geral e uniforme que revelaria o nível de um determinado aprendiz em qualquer parte do mundo, o que, de fato, não ocorre.

[2] Neste trabalho optei por utilizar esta ortografia (sem itálico e com acento), recomendada por Sacconi (1996, p. 204) e usada por Cassemiro (2009) em sua dissertação de mestrado por considerá-la mais condizente com as regras de acentuação do português.

Mais especificamente, pelo fato de ter sido necessária a criação de um córpus (o Córpus Multinível de Aprendizes Brasileiros de Inglês como Língua Estrangeira, doravante COBRA-7), para verificar tais erros o trabalho aqui proposto fundamenta-se na vertente da Linguística de Córpus que se ocupa com córpora de aprendizes, que "podem ser rudemente definidos como coleções de textos eletrônicos produzidos por aprendizes de língua..."[3] (Granger, 2002, p.7). Os trabalhos centrais da área com relevância para o projeto são os de Granger (1998, 2002, 2008), O'Keeffe, McCarthy e Nesselhauf (2004).

Embora haja publicações voltadas aos erros produzidos pelos aprendizes de idiomas estrangeiros – dentre as quais pode-se citar Berber Sardinha e Shepherd (2011), Dutra e Silero (2010), Delegá-Lucio (2006) e Balbás (2003) – parece haver necessidade de mais pesquisas que utilizem um córpus de produções escritas de aprendizes brasileiros de inglês provenientes de escolas de ensino de língua estrangeira e forneça aos professores que lidam com esses aprendizes ferramentas para a classificação e identificação de erros. O presente projeto buscará preencher essas lacunas.

Dado o contexto geral, a justificativa pessoal deste projeto reside no fato de que os professores de idiomas lidam frequentemente com erros em seu cotidiano e muitas vezes não se dão conta de que esses erros se repetem, ou não sabem o que fazer para diminuir suas incidências. Acredito que com esta pesquisa poderei, indiretamente, não apenas compreender melhor minha própria atuação, mas também, diretamente, fornecer recursos para que outros professores possam observar suas próprias realidades em sala de aula, identificar erros nas produções dos seus aprendizes segundo o método proposto neste trabalho e criar atividades que visem a melhorar as produções escritas dos seus estudantes.

A relevância deste trabalho deve-se à investigação dos erros mais comuns para os aprendizes brasileiros em cada nível de curso. Além disso, deve-se também à criação de uma metodologia de classificação e identificação de erros para auxiliar os professores de inglês em sua atuação com aprendizes brasileiros.

Em segundo lugar, embora o foco da pesquisa nos erros pareça à primeira vista remeter ao estruturalismo, acredito que eles devam ainda ser objetos de estudo, pois são questões que permanecem no cotidiano do professor de inglês, ainda que inseridos em um quadro teórico funcionalista, como é o caso da abordagem comunicativa, adotada por grande parte das escolas de idiomas hoje. Portanto, esta pesquisa não pode ser considerada estruturalista, pois acredito que léxico e gramática são inseparáveis e definidos no uso.

[3] "*can be roughly defined as electronic collections of learner data*" (tradução minha).

Em terceiro lugar, os resultados desta pesquisa podem trazer dados importantes para o Grupo de Estudos de Linguística de Córpus (GELC), do qual faço parte, pois esse grupo tem interesse na publicação de livros voltados para professores de idiomas.

Por fim, mas não menos importante, a relevância do projeto está no fato de o córpus COBRA-7 ser inédito em sua categoria, pois além de possuir mais de 570.000 palavras (2260 redações), é um dos poucos córpora brasileiros de aprendizes que utilizam a produção escrita de estudantes de inglês matriculados em uma rede de escolas de idiomas. Dentre os córpora de aprendizes nacionais, os mais próximos ao COBRA-7 seriam o CABrI (36.187 em 2010) (UFMG) e o Br-Icle (40.834 palavras) (LAEL-PUC-SP) – sub-córpus do *The International Corpus of Learner English* (ICLE), ambos compostos de textos produzidos por universtários estudantes de Letras.

Isso posto, passarei à descrição do objetivo e dos questionamentos que nortearam esta pesquisa.

Esta pesquisa tem como objetivo identificar e classificar os erros na escrita de aprendizes brasileiros de inglês. Um dos desdobramentos possíveis desta pesquisa seria a possibilidade de prover aos professores e pesquisadores um sistema de identificação e classificação de erros, com vistas a auxiliá-los em seu trabalho e informar a produção de materiais didáticos locais, isto é, voltados a aprendizes brasileiros.

Dados esse objetivo, as perguntas de pesquisa a serem investigadas no trabalho são elencadas abaixo:

1- Quais os erros mais comuns no córpus COBRA-7_recorte[4]?

2- Qual a variação de erro entre os níveis de curso dos aprendizes no córpus COBRA-7_recorte?

3- Qual nível de curso apresenta maior diversidade de erros no córpus COBRA-7_recorte?

A dissertação está organizada da seguinte maneira. O capítulo 1 discute a fundamentação teórica da pesquisa, mostrando os principais conceitos e trabalhos prévios na Linguística de Córpus, córpora de aprendizes e estudo de erros. O capítulo 2 apresenta em detalhes a metodologia empregada na pesquisa, incluindo a descrição do córpus e seu processo de coleta e criação, bem como a especificação dos procedimentos de análise dos dados, criação da metodologia de classificação e identificação de erros aqui proposta. O capítulo 3 mostra e interpreta os resultados

[4] O COBRA-7_recorte é uma amostra do COBRA-7, o córpus de estudo desta pesquisa. Tal recorte será justificado posteriormente no capítulo Metodologia.

obtidos para cada uma das questões de pesquisa. Por fim, o capítulo ‘Considerações Finais’ faz um fechamento do estudo. A bibliografia encerra a dissertação.

Capítulo 1: Fundamentação Teórica

Esta pesquisa, como visto anteriormente, tem como objetivo identificar e classificar os erros na escrita de aprendizes brasileiros de inglês. Um dos desdobramentos possíveis desta pesquisa seria a possibilidade de prover aos professores e pesquisadores um sistema de identificação e classificação de erros, com vistas a auxiliá-los em seu trabalho e informar a produção de materiais didáticos locais, isto é, voltados a aprendizes brasileiros. Para atingir essa finalidade irei expor, neste capítulo, o arcabouço teórico sobre o qual a pesquisa se fundamenta. Primeiramente são apresentados os trabalhos referentes à Linguística de Córpus e que definem a área. A seguir, os trabalhos específicos da área conhecida como Córpora de Aprendizes, que fornecem os parâmetros tipológicos para a criação de um córpus. Por fim, justifico o conceito de erro nesta pesquisa.

1.1 Linguística de Córpus

Conforme dito na Introdução, o trabalho aqui proposto tem como fundamentação teórica principal a Linguística de Córpus, que pode ser definida como uma área que

> ocupa-se da coleta e da exploração de córpora, ou conjuntos de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou uma variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador.
>
> (Berber Sardinha, 2004, p. 3)

Os principais expoentes na Linguística de Córpus são: Sinclair (1966, 1987, 1988, 1991, 2004), que assentou a pedra fundamental da área, Halliday (1991, 1992), Hoey (1991, 1997), Biber (1993, 1999), Biber, Conrad e Reppen (1998), Berber Sadinha (1998, 1999, 2000, 2001, 2004), Granger (1998, 2002, 2008), O'Keeffe, McCarthy e Carter (2007) e Nesselhauf (2004). Estudos essenciais desta área para o presente projeto são Sinclair (1987b, 1991, 1996, 2004 e 2004b), que fornecem os conceitos fundamentais da área e os critérios de coleta, armazenamento e pré-processamento de córpora; Berber Sardinha (2000, 2004), que resume os aspectos supracitados, e Granger (1998, 2002, 2008), O'Keeffe, McCarthy e Neselhauf (2004), que apresentam conceitos relativos aos córpora de aprendizes, como definição, tamanho, tipologia e *design*, e resumem a literatura da área.

A Linguística de Córpus trabalha dentro de um quadro teórico empirista. Segundo essa perspectiva, os dados devem provir da linguagem em uso (cf. Sinclair, 1991; Bennet, 2010, p. 7). Sinclair (1996) chama tal particularidade de autenticidade. Esse enfoque se contrapõe ao gerativista, que defende o "estudo da linguagem por meio da introspecção, como forma de verificar modelos de funcionamento estrutural e processamento cognitivo da linguagem" (Berber Sardinha, 2004, p. 30). Assim, enquanto que os adeptos do gerativismo se interessam em estudar a estrutura linguística dentro da mente (cognitiva) e "quais agrupamentos sintáticos são possíveis (permissíveis) dado o conhecimento que um falante nativo possui de sua língua" (Berber Sardinha, 2004, p. 30), a perspectiva empirista se interessa pela observação e estudo da linguagem em sociedade e pela "probabilidade dos sistemas linguísticos" (Berber Sardinha, 2004, p. 30), isto é, quais palavras co-ocorrem com outras na língua em uso e com qual frequência (linguagem como sistema probabilístico, cf. Halliday, 1991). Segundo Sinclair (1991), a introspecção não permite ao falante distinguir fatores como o que mais ocorre na língua e as escolhas lexicais feitas. Por isso, a língua deve ser estudada colhendo-se material produzido em "comunicações genuínas de pessoas realizando suas atividades cotidianas"[5] (Sinclair, 1996, p. 7).

A abordagem empirista, portanto, pressupõe que diversos traços linguísticos, embora possíveis teoricamente, podem não ocorrer na prática com grande frequência. Isso acontece porque é o uso que estabelece a predileção de ocorrência (grau de atração) para uma determinada combinação de palavras. A Linguística de Córpus estuda essa ocorrência em produções escritas (ou orais transcritas) reais e as analisa (cf. Biber et al., 1998, p. 3) por meio de programas computacionais que gerem resultados que possam ser usados em finalidades diversas, dentre as quais pode-se citar:

a) a Lexicografia (criação de dicionários) como no projeto COBUILD, liderado por Sinclair na Universidade de Birmingham em 1980 (cf. O'Keeffe, McCarthy e Carter, 2007, p. 17);

b) a criação de gramáticas como a *Longman grammar of spoken and written English*, de Biber *et al* (1999);

c) o Processamento de Linguagem Natural (PLN) e à Linguística Computacional (LC) (cf. Berber Sardinha, 2004), embora essas áreas em geral operem dentro do quadro teórico gerativista;

[5] "*genuine communications of people going about their normal business*" (p. 7, tradução minha).

d) o setor editorial, onde há o interesse em se compilar grandes córpora linguísticos com a finalidade de se utilizarem sentenças reais em exemplos que componham livros didáticos voltados a cursos de idiomas;

e) a área de tradução (cf. O'Keeffe, McCarthy e Carter, 2007, p. 19);

f) a Línguística Forense, área de análise criminalística na qual os córpora "podem ser usados para comparar padrões linguísticos"[6] (O'Keeffe, McCarthy e Carter, 2007, p. 20);

g) a pesquisa com a produção de aprendizes. Tal enfoque é conhecido como Córpora de Aprendizes.

1.1.1 Organização e coleta de córpus

Um córpus deve estar necessariamente em formato eletrônico. Todavia, nem sempre os dados são coletados nesse formato. Há casos em que são transcritos a partir de outros tipos de mídia, como revistas, jornais e mesmo composições manuscritas. Naturalmente, quanto mais próximos os dados estiverem dos formatos eletrônicos utilizados atualmente, mais rápida será a organização do córpus. O formato de texto mais utilizado na linguística de córpus para usuários da plataforma (sistema operacional) *Windows* é o de extensão .TXT (*plain text*, como definido por Sinclair, 1996, p. 8), ou seja, um tipo de arquivo de texto que permite uma formatação mínima e aceita apenas caracteres alfanuméricos e demais símbolos disponíveis no teclado do computador, como "*", "%", "@", ",". Esse tipo de arquivo é utilizado tanto pela questão da simplicidade de formatação quanto pelo fato de ocupar pouco espaço no disco rígido do computador, o que acelera o processamento quando se analisam grandes quantidades de arquivos por meio de programas computacionais como o *Wordsmith Tools 5.0* (Scott, 2008).

Berber Sardinha (2004) e Bergh e Zanchetta (2008) discutem como a Internet pode ser uma grande fonte de dados. De fato, Giulli e Signorini (2005) estimaram na época um número de mais de 10 bilhões de páginas na Internet, o que demonstrava uma quantidade dez vezes maior do que em

[6] "*...can be used to compare language patterns*" (tradução minha, p. 20).

1999. Siemens (2005) afirma que, por causa da Internet, o conhecimento dobra a cada 18 meses. Parte desse conteúdo está disponível gratuitamente, o que facilita muito o trabalho do linguista de córpus. Porém, uma considerável parte – 5 a 10 vezes maior do que é publicamente acessado na rede mundial de commputadores (cf. Bergh e Zanchetta, 2008, p. 312) – está na "*web* invisível"[7] (Bergh e Zanchetta, 2008, p. 311-312), ou seja, um ambiente virtual acessado por meio de um nome de usuário e senha e que é invisível a sistemas de busca como o *Google*. Os autores também relatam três formas possíveis de busca na Internet: "busca, pastoreio e folheio" (p. 18)[8]. "Buscar" seria procurar em sítios como o *Google* ou outros um termo de busca específico. Encontra-se, desse modo, somente o que se procura. "Pastorear" significa colher textos em um provedor que os reúne em quantidade, como, por exemplo, o sítio de um jornal a partir do qual se pode copiar diversos exemplares ao mesmo tempo. Analogamente, o espaço virtual seria como uma fazenda, os dados como o gado e o provedor como o fazendeiro. "Folhear", por sua vez, significa encontrar dados relevantes por acaso, enquanto se navega na Internet. Um exemplo de "folheio" é a procura de gírias diversas em *blogs*.

Nem todo conjunto de textos escritos constitui um córpus. Para que seja considerado dessa forma é preciso que seja coletado e organizado segundo critérios específicos. Berber Sardinha (2004) sugere que páginas da Internet, ao serem copiadas, passem por um processo de "limpeza" utilizando-se um *script* desenvolvido em PERL, uma linguagem de programação estável e multiplataforma criada por Larry Wall em 1987. Esse processamento é importante para se remover as formatações dos textos e os elementos gráficos, como figuras e vídeos. Porém, esse procedimento pode ser complexo para quem tem pouco conhecimento dessa linguagem ou de computadores. Uma alternativa a isso é se copiar o conteúdo da página da Internet com um comando simples de cópia e colá-lo em um arquivo .TXT. Isso será feito nesta pesquisa, conforme será visto no capítulo Metodologia.

Após a cópia do texto em um arquivo com extensão .TXT, ele deve ser salvo no disco rígido do computador em uma pasta chamada *corpus* ou *corpora*[9], dentro de uma pasta interna que dê nome ao projeto realizado, como "mestrado"[10].

O passo seguinte para a organização de um córpus é inserir informações a respeito dele para facilitar a localização do arquivo e busca interna por dados específicos. Há diversas formas de

[7] Tradução minha para *invisible web* (p. 311-312).
[8] Tradução minha para *hunting*, *grazing* e *browsing*, respectivamente (p. 18).
[9] Este tipo de notação foi incorporado a partir da área da computação e por isso não utiliza acentuação. Tal critério é recomendável para que seja facilitada – ou possibilitada – a manipulação das pastas em qualquer sistema operacional.
[10] Deve-se evitar acentuação, e palavras compostas devem ser unidas com um sinal de "_". Isso deve ser feito porque nem todos os programas computacionais e plataformas operacionais (como *DOS*, por exemplo) são capazes de interpretar outros padrões de nomeação.

inseri-las. Uma delas é a criação de arquivos paralelos em formato *XML* (*Extensible Markup Language*), um formato de marcação para a descrição de dados. Esse tipo de arquivo pode ser criado paralelamente, com o mesmo nome do arquivo .TXT correspondente a cada arquivo que compõe o córpus, de modo que não interfira nos dados coletados.

Uma outra forma de se associar informações sobre os textos de um córpus é inserindo cabeçalhos no início de cada arquivo .TXT que o compõe. Cabeçalhos

> são uma parte do arquivo de cada texto do córpus que contém informações sobre o texto, tais como a origem, a data de coleta, o grupo de pesquisa responsável, o tamanho do texto, sistema de transcrição, detalhes do *copyright*, a autoria, os participantes. Essas informações deixam explícitos vários pontos importantes do texto, de tal modo que o computador possa processá-las.
>
> (Berber Sardinha, 2004, p. 73-74)

Berber Sardinha (2004) discute dois tipos de cabeçalhos: *SGML* (*Standard Generalized Markup Language*) e *COCOA* (*Count and Concordance on Atlas*). O primeiro, de modo semelhante à marcação *XML*, utiliza etiquetas de abertura e fechamento do tipo <abrir>TEXTO</fechar>.

O segundo tipo de cabeçalho, *COCOA*, não possui etiquetas de abertura e fechamento, mas apenas uma etiqueta do tipo <autor Shakespeare> dentro da qual a informação é escrita. Este tipo de cabeçalho será utilizado na organização do córpus COBRA-7, como será visto no capítulo Metodologia.

Os cabeçalhos são importantes para que haja um registro claro que identifique o córpus (cf. Sinclair, 1996): fonte, tipologia, pesquisador responsável, data de compilação etc. Além de facilitar a manipulação do córpus pelo próprio criador e demais pesquisadores interessados, o uso do cabeçalho é um padrão sugerido e valorizado pela Linguística de Córpus (cf. Berber Sardinha, 2004, p. 76), pois serve como lembrete do que cada arquivo etiquetado contém, o que beneficia não somente o autor da pesquisa como outros interessados.

De modo geral, programas computacionais para a análise de córpus, como o *Wordsmith Tools 5.0* (Scott, 2008) ignoram informações que estejam dentro do córpus entre os sinais de menor e maior "<informação>". Dessa forma, o conteúdo do córpus não é comprometido pelas informações do cabeçalho.

1.1.2 Tipologia, tamanho de córpus e abordagens de análise

Sinclair (1991) fornece alguns critérios para a criação de um córpus. Primeiramente, especifica que um córpus é eletrônico. Em seguida, menciona a necessidade de se observarem os direitos autorais ao compilá-los. No caso do córpus de estudo deste trabalho, o COBRA-7, o uso das composições de aprendizes está de acordo com as normas do Comitê de Ética em Pesquisa da PUC-SP (Protocolo de Pesquisa nº 230/2011, cf. Anexos). O resumo abaixo, baseado em Sinclair (1991), traz a tipologia de córpus. Segundo ela, um córpus pode ser:

a) falado: contém transcrições de textos falados, ou seja, de diálogos ou falas espontâneas;

b) escrito: contém textos escritos de naturezas diversas (literários, acadêmicos etc.);

c) *quasispeech*: contém textos que simulam a oralidade, como no caso de roteiros de filmes.

Para Berber Sardinha (2004), os córpora devem ser eletrônicos, criteriosamente escolhidos, representar uma língua ou variedade e ser compostos a partir de dados autênticos com a finalidade "de ser um objeto de estudo linguístico" (p. 18). Para o autor, a representatividade do córpus está intimamente associada ao seu tamanho: quanto maior, mais representativo o córpus pode ser. A tabela a seguir apresenta, em números, os tamanhos e as classificações de córpora observados na época e que podem servir de base:

Tamanho em palavras	Classificação
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Tabela 1.1.1: quadro-resumo dos tamanhos de córpora (formatação adaptada de Berber Sardinha, 2004, p. 26).

O tipo de córpus deve ser estabelecido considerando-se alguns critérios específicos. O quadro a seguir, adaptado de Berber Sardinha (2004, p. 20-21), resume esses critérios:

a) **modo**: falado (contém transcrição de fala);
escrito (contém textos escritos);

b) **tempo:** sincrônico: (compreende um período de tempo);
diacrônico (compreende vários períodos de tempo);
contemporâneo (representa o tempo atual);
histórico (representa um tempo passado);

c) **seleção:** de amostragem (é uma pequena amostra de uma variedade);
monitor (é uma amostra considerável de uma variedade);
dinâmico (seu tamanho pode aumentar ou diminuir);
estático (seu tamanho é fixo);
equilibrado (os números de textos são distribuídos igualmente);

d) **conteúdo:** especializado (textos de tipos específicos);
regional (textos provêm de uma ou mais variedades);
multilíngue (abrange idiomas diferentes);

e) **autoria:** de aprendiz (os autores não são falantes nativos);
de língua nativa (os autores são falantes nativos);

f) **disposição:** paralelo (contém textos comparáveis);
alinhado (tradução abaixo da linha do original);

g) **finalidade:** de estudo (o córpus que se pretende descrever);
de referência[11] (serve como contraste ao córpus de estudo);
de treinamento (feito para se desenvolver aplicações e ferramentas).

A tabela a seguir resume essa classificação:

Modo:	Falado	escrito				
Tempo:	sincrônico	diacrônico		contemporâneo	histórico	
Seleção:	de amostragem	monitor		dinâmico	estático	equilibrado
Conteúdo:	especializado	regional/dialetal	multilíngue			
Autoria:	de aprendiz	de língua nativa				
Disposição:	paralelo	alinhado				
Finalidade:	de estudo	de referência	de treinamento			

Tabela 1.1.2: tipologia de córpus (quadro-resumo baseado em Berber Sardinha, 2004, p. 20-21).

[11] "O tamanho recomendado de um córpus de referência é cinco vezes o tamanho do córpus de estudo" (Berber Sardinha, 2004, p. 102).

As pesquisas em Linguística de Córpus podem ser abordadas de duas formas: "guiada pelo córpus"[12], ou seja, baseada no que o córpus revelar ao pesquisador (*corpus-driven*, cf. Tognini-Bonelli, 2001), ou "baseada em córpus"[13], isto é, quando o pesquisador tenta demonstrar por meio de córpus uma hipótese prévia (*corpus-based*, cf. Tognini-Bonelli, 2001). Nesta pesquisa, trabalharei com ambos os casos, pois aplicarei um sistema de classificação de erros baseado em Shepherd (2001) (baseada em córpus) e desenvolverei um outro sistema a partir da observação do recorte do córpus COBRA-7 (guiada pelo córpus).

1.1.3 Léxico-gramática[14]

Sinclair (1991) justifica que há duas características organizacionais da linguagem: a primeira é o que o autor chama de princípio de escolha livre e a segunda é chamada de princípio idiomático[15]. O primeiro princípio pressupõe que a comunicação se constroi por meio do uso de estruturas fixas que determinam a organização das palavras nas orações. Segundo esse pressuposto, se a estrutura da língua portuguesa permite que em geral um adjetivo suceda um substantivo, então qualquer palavra que pertença a essa classe morfológica atende a essa necessidade, como no caso de "grande", que sucederia o substantivo "amigo", cuja combinação resultaria em "amigo grande". Esse princípio considera que a formação das orações acontece preenchendo-se lacunas com palavras que correspondam às classes gramaticais pertinentes. O uso da linguagem, portanto, nessa visão, é randômico.

O segundo princípio (idiomático), pressupõe que a língua deve ser estudada a partir de seu uso em situações cotidianas, que define quais palavras sofrerão um maior ou menor grau de atração. Nesse caso, as combinações de palavras não ocorrem de forma randômica (cf. Sinclair, 1991): é o uso que determina qual o agrupamento e o significado que virá desse processo. Esse princípio defende, portanto, que o falante tem à sua disposição um conjunto de combinações lexicais pré-construídas, dentre as quais usará aquela que melhor convier à situação. Dentre essas expressões pode-se citar *of course*, o quantificador *a pint of*, *the first time* e outros. Aplicando esse princípio ao exemplo do parágrafo anterior, ver-se-á que não necessariamente o posicionamento do adjetivo "grande" após o substantivo "amigo" atenderia à necessidade de um usuário da língua. Seria possível fazer uma inversão ("grande amigo"), a qual alteraria por completo o significado dessa expressão, pois o substantivo "grande" deixaria de ter signficado literal (tamanho) e ganharia um significado ligado a importância. Isso acontece porque, para Sinclair e Teubert (2007), combinações

[12] Tradução usada por Condi (2005, p. 24).
[13] Idem.
[14] *Lexicogrammar* (tradução usada por Berber Sardinha, 2004).
[15] Estas traduções para os termos *open-choice principle* e *idiom principle* são usadas por Berber Sardinha (2004, p. 33).

lexicais como a que usei acima não têm relação com a classe morfológica e sua posição, pois a gramática não atribui significado, mas apenas o gerencia. A escolha pela expressão "amigo grande" ou "grande amigo" é é definida pelo uso e traz simultaneamente ambas as palavras que a compõem.

Assim, o princípio de livre escolha pressupõe que a linguagem é constituída em boa parte por combinações típicas de palavras. Tais combinações são constituídas de modos diferentes de acordo com a língua. Em português, por exemplo, esse princípio nos permite afirmar que em geral os adjetivos sucedem os substantivos, como no caso da combinação "homem grande". O princípio idiomático, por sua vez, pressupõe que as palavras se combinam devido ao grau de atração que exercem umas com relação às outras, o que é estabelecido pelo uso. Esse princípio justificaria, por exemplo, a ocorrência em português da combinação "grande homem", a qual não somente inverte a ordem das palavras do exemplo anteriormente utilizado para ilustrar o pressuposto do princípio da livre escolha mas também lhe altera o significado. Esse exemplo pode indicar que o princípio idiomático de Sinclair anula o princípio da livre escolha. Porém, Sinclair (1991), ao falar sobre os princípios de livre escolha e idiomático, não diz que aquele é falso. Na verdade, os discursos "são criados pela combinação dos princípios idiomático e de livre escolha"[16] (Partington, 1998, p. 21), o que torna difícil identificar quando cada um deles está operando. Por isso a Linguística de Córpus analisa o "comportamento" de determinadas palavras dentro de seus contextos gramaticais.

Baker *et al.* (2006, p. 7) afirmam que os "itens lexicais devem ser caracterizados em termos de suas distribuições em padrões gramaticais"[17]. Isso significa que é difícil se fazer uma separação entre léxico e gramática, motivo pelo qual, na Linguística de Córpus, léxico e gramática são analisados conjuntamente: "as considerações semânticas estão associadas às escolhas gramaticais de modo complexo" (Sinclair, 2004b, p. 276)[18]. Por isso mesmo, as ocorrências de erro encontradas no córpus COBRA-7 são formadas por padrões léxico-gramaticais.

Como léxico e gramática são interdependentes, o foco em um não exclui o outro. O princípio idiomático citado anteriormente, também chamado por Partington (1998) de princípio colocacional, é composto pelo que Sinclair (1991) chama de colocação e Bolinger (1976) chama de *prefabs* ou idiomacidade, que consiste no significado que uma palavra assume por causa de outras que a acompanham (cf. Leech, 1974).

Entende-se por colocação "a ocorrência de duas ou mais palavras com um pequeno espaço entre elas dentro de um texto"[19] (Sinclair, 1991, p. 170). O termo foi introduzido por Firth (1957) e

[16] "*are created by a combination of the idiom and open choice principles*" (p. 21, tradução minha).
[17] "*lexical items must be characterised in terms of their distributions in grammatical patterns*" (p. 7, tradução minha).
[18] "*semantic considerations are intricately associated with grammatical choices*" (p. 276, tradução minha).
[19] "*Collocation is the occurrence of two or more words within a short space of each other in a text*" (p. 170, tradução minha).

pressupõe que o significado é colocacional, ou seja, é o agrupamento das palavras que revelam seus significados: "Um dos significados de noite é sua colocabilidade com escura, e, de escura, naturalmente, sua colocação com noite" (p. 196)[20].

Enquanto que para Firth e Leech o termo está relacionado à alteração que o significado do nódulo recebe por causa das palavras que o acompanham, para Sinclair e Teubert (2007) cada combinação de palavras cria um significado específico. Com relação ao excerto supracitado, a palavra "noite" seria uma palavra com um significado específico, ao passo que "noite escura" constituiria outro significado, uma espécie de redução no significado de noite, a saber: o de uma noite particularmente escura:

> Uma vez que aceitamos que as palavras podem ser co-selecionadas, não escolhidas sempre uma por vez, então não há mais um problema com noite escura; noite não distingue um dos significados de escura, e nem escura distingue um dos significados de noite. A frase noite escura tem seu próprio significado; *grosso modo*, a noção de 'escuro' já está presente na noção de 'noite' (embora nem todas as noites sejam escuras, é característica de uma noite ser escura). Então o adjetivo escuro não está selecionando dentre todas as possíveis noites aquelas que são escuras, mas está reforçando o elemento escuro já presente em noite[21] (Sinclair e Teubert, 2007, cap. 12).

Tais significados estão pré-ativados na mente do falante. Isso significa que, para o falante nativo, os significados são definidos pelo uso e apreendidos localmente por meio da exposição.

Nesselhauf (2005) divide as colocações em quatro grupos: a) "combinações livres"[22] (p. 14): todas as palavras que compõem a expressão são usadas com sentido literal, como no caso de *drink tea* (beber chá, em português); b) "colocações restritas"[23] (p. 14): pelo menos uma das palavras possui significado não literal, como no caso de *perform a task* (realizar uma tarefa); c) "expressões idiomáticas figurativas"[24] (p. 15): a substituição das palavras que compõem a colocação é raramente possível. O significado da combinação é figurado, mas há a preservação de uma interpretação literal, como no caso de *do a U-turn* (mudar radicalmente o comportamento); d) "expressões

[20] "*One of the meanings of night is its collocability with dark, and of dark, of course, collocation with night*" (p. 196, tradução minha).
[21] "*Once we accept that words can be co-selected, not chosen always once at a time, then there is no longer a problem with dark night; night does not distinguish one of the meanings of dark, nor does dark distinguish one of the meanings of night. The phrase dark night has its own meaning; roughly speaking, the notion 'dark' is already present in the notion 'night' (though not all nights are dark, it is characteristic of a night to be dark). So the adjective dark is not selecting among all possible noghts, the dark ones, but is reinforcing the dark element already in night*." (cap. 12, tradução minha).
[22] *Free combinations* (tradução minha, p. 14).
[23] *Restricted collocations* (tradução minha, p. 14).
[24] *Figurative idioms* (tradução minha, p. 15).

idiomáticas puras"[25] (p. 15): a substituição das palavras é impossível e o significado é figurado, isto é, não há marca de interpratação literal. Exemplo: *blow the gaff* ("abrir o bico", contar o segredo de alguém).

Sinclair e Teubert (2007) nos permitem dividir as colocações em dois grupos: "colocações simples" (mais previsíveis em seu uso e estrutura, como *of course*, *on the strength of*, *good morning*, *in the middle of* e outras) e "expressões idiomáticas" (cap. 9)[26] (combinações de palavras abstratas e definidas pelo uso, como *raining cats and dogs*, *cost the earth* e outras).

Nem toda combinação lexical constitui uma colocação. Para que seja considerada uma colocação é em geral necessário que seja verificado o grau de atração das palavras que a compõem por meio de cálculos estatísticos. Neste trabalho não farei tal verificação, pois esta pesquisa não foca no estudo das colocações, mas sim em um método para a identificação de palavras problemáticas dentro de uma combinação de palavras. Por isso, usarei nesta pesquisa o termo colocação indiscriminadamente sempre que me referir a combinações de palavras, sem diferir se são muito ou pouco literais.

1.1.4 Coligação

Além das colocações, intimamente ligadas ao princípio idiomático, a literatura de córpus traz também o termo coligação (eg. Firth, 1968, p. 182), que se refere à co-ocorrência de palavras entre categorias gramaticais específicas (cf. Hoey 2000, p. 234).

Enquanto que o termo colocação pode ser mais diretamente ligado ao príncípio idiomático e contempla a combinação de palavras como "pronto" e "socorro" (pronto socorro, i.e., um centro de atendimento de urgência), o termo coligação contempla tanto conceitos do princípio idiomático quanto do princípio de livre escolha, isso porque indica a combinação de uma determinada palavra com uma classe gramatical, como, por exemplo, "preposição + carro", a qual pode representar um termo de busca para se descobrir quais preposições aparecem com a palavra "carro". Nesta pesquisa o termo coligação será empregado sempre que houver menção a um termo de busca que englobe explicitamente léxico e gramática, como no exemplo acima.

Tal conceito é importante porque servirá de base para a metodologia de investigação de erros proposta nesta pesquisa, como será visto posteriormente.

---

[25] *Pure idioms* (tradução minha, p. 15).
[26] Traduções minhas para *collocations* e *idioms*, simultaneamente. Cap. 9.

1.2 Córpora de aprendizes

Por se tratar de pesquisa com produção de aprendizes, este trabalho utiliza as definições e dos pressupostos teóricos de Granger (1998, 2002, e 2008), O'Keeffe, McCarthy, Carter (2007), Nesselhauf (2004) e outros para a criação de um córpus de aprendizes. Desse modo será possível a compilação do córpus COBRA-7 segundo os padrões esperados dentro da Linguística de Córpus.

O trabalho com córpus de aprendizes é recente dentro da Linguística de Córpus. Granger (2008) atribui o seu aparecimento ao final dos anos 80 e/ou início dos anos 90.

Baker, Hardie, McEnery (2006, p. 103) definem Córpus de Aprendiz como a "produção linguística criada por aprendizes de uma língua. Boa parte dos córpora consiste de ensaios escritos usando tópicos pré-estabelecidos produzidos em salas de aula de ensino de idiomas[27]" (p. 103).

Para Granger (2002),

> Córpora de aprendizes em computador são coleções eletrônicas de dados textuais autênticos de LE/SL reunidos de acordo com critérios de *design* explícitos para um propósito de ASL/ELE em particular (p. 7)[28].

Nesselhauf (2004) os define como "sistemáticas coleções computadorizadas de textos produzidos por aprendizes de língua[29]" (p. 125).

Como a Linguística de Córpus trabalha dentro de um quadro empirista, a autenticidade é importante no seu objeto de estudo. Segundo Sinclair (1996), o termo autenticidade indica que "O material todo é reunido a partir de comunicações genuínas de pessoas em situação natural[30]" (p. 7).

Granger (2002), todavia, reconhece que essa autenticidade seja discutível no que concerne aos córpora de aprendizes, pois dificilmente as comunicações e tarefas produzidas pelos estudantes são espontâneas.

---

[27] "*language output produced by learners of a language. Most learner corpora consist of written essays using pre-set topics produced in language-teaching classrooms*" (tradução minha, p. 103).
[28] "*Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose*" (tradução minha, p.7). As siglas *FL/SL* e *SLA/FLT*, traduzidas como LE/SL e ASL/ELE, significam, respectivamente, *Foreign Language/Second Language* (Língua Estrangeira/Segunda Língua) e *Second Language Acquisition/Foreign Language Teaching* (Aquisição de Segunda Língua/Ensino de Língua Estrangeira).
[29] "*systematic computerized collections of texts produced by language learners*" (tradução minha, p. 125).
[30] "*All the material is gathered from the genuine communications of people going about their normal business*" (p. 7, tradução minha).

Nesta pesquisa considerarei as produções dos aprendizes como autênticas, pois a despeito das afirmações de Sinclair (1996) e Granger (2002) supracitadas, a produção escrita de aprendizes possui um caráter de comunicação verdadeiro, pois destina-se a um leitor. Não se trata, portanto, de um conjunto de sentenças provenientes da cabeça do aprendiz e que servirão como exemplos abstratos para uma análise de erros, mas sim de um conjunto de informações e colocações produzidas para a apreciação de outros falantes, ou seja, para o uso escolar. Portanto, os córpora de aprendizes são legítimos no sentido de que não são apenas uma coletânea de textos avulsos.

Granger (2008) diz que os estudantes cujas composições se enquadram nas pesquisas com córpora de aprendizes podem ser considerados como pertencentes a um dos seguintes grupos[31]:

a) "Inglês como Segunda Língua (ISL)" (p. 260)[32]: trata-se do inglês aprendido em um país no qual essa é a língua oficial, em geral enquanto imigrante.

b) "Inglês como Língua Estrangeira (ILE)" (p. 260)[33]: engloba o idioma inglês aprendido em sala de aula em um país no qual não é o idioma oficial. É o caso do inglês aprendido no Brasil, por exemplo e, desse modo, dos aprendizes cujas redações irão compor o COBRA-7.

Para que sejam considerados dados de um córpus é necessário que os textos sejam integrais, isto é, que não tenham sido recortados aleatoriamente. Pequenos trechos ou frases não se qualificam como dados de córpus (cf. Granger, 2008). Dessa forma, um córpus de aprendizes irá conter os erros e os usos corretos da língua (cf. Granger 2008, p. 49-50).

Granger (2002) também fala sobre os critérios de coleta e organização de córpus. Com relação à coleta, a captura dos dados que irão compor o córpus pode ser feita por meio de "*download* eletrônico, escaneamento ou digitação"[34] (p. 174), sendo esta a forma mais comum (cf. Granger, 1998). No caso dos dados que formarão o córpus COBRA-7, como será visto, as composições dos aprendizes foram digitadas diretamente por eles na Internet. Essa possibilidade

[31] Nesselhauf (2004, p. 128) menciona esses mesmos grupos.
[32] *English as a Second Language* (*ESL*), (p. 260, tradução minha).
[33] *English as a Foreign Language* (*EFL*), (p. 260, tradução minha).
[34] *"Downloading of electronic data, scanning and keyboarding*" (p. 174, tradução minha). O'Keeffe, McCarthy e Carter (2007, p. 2) falam do *download* e da digitação ao abordar o tema.

não foi prevista por Granger. Porém, Sinclair (1991) a prevê quando propõe que um córpus pode ser formado por meio da "adaptação de material que já esteja em formato eletrônico" (p. 14)[35].

Granger (1998) justifica que os córpora de aprendizes devem ser organizados mediante critérios específicos como, por exemplo, o uso de cabeçalhos *SGML* (*Standard Generalized Markup Language*) ou outros, como visto anteriormente, já que tal informação fornece um registro preciso do histórico dos dados (origem, tipologia, pesquisador responsável etc.), o que facilita pesquisas futuras e alinha a pesquisa aos critérios da Linguística de Córpus.

Sobre o propósito de um córpus de aprendizes, Granger (2002) diz que, entre outros, está a contribuição "para a produção de melhores ferramentas e métodos de ELE" (p. 50)[36], que é um dos objetivos desta pesquisa. Tal propósito é confirmado por Nesselhaudf (2004), que o divide em dois:

a) "identificar o que é particularmente difícil para para um certo grupo de aprendizes e dar uma ênfase especial nesses pontos em materiais diferentes" (p. 126)[37];

b) "tirar das análises do córpus de aprendiz ideias a respeito da aquisição de segunda língua (por exemplo sobre sequências de desenvolvimento) e desenvolver implicações para o ensino a partir dessas percepções" (p. 126)[38].

Granger (2002) fornece uma tipologia de córpus semelhante à de Berber Sardinha (2004). Segundo a autora, a tipologia de um córpus é sempre dicotômica, pois as classificações se opõem, conforme mostra a tabela a seguir, adaptada de Granger (2002 e 2008):

[35] "*adaptation of material already in electronic form*" (p. 14, tradução minha).
[36] *"...to the production of better FLT tools and methods*" (p. 50, tradução minha). A sigla *FLT* representa *Foreign Language Teaching*) foi traduzida como ELE (Ensino de Língua Estrangeira).
[37] "*to identify what is particularly difficult for a certain group of learners and to put special emphasis on these points in the different materials*" (tradução minha, p. 126).
[38] "*to derive insights about second language acquisition (for example about developmental sequences) from learner corpus analyses and to draw implications for teaching from these insights*" (tradução minha, p. 126).

Monolíngue		⇔		Bilíngue
Geral		⇔		Técnico
Sincrônico		⇔		Diacrônico
Escrito		⇔		Falado
Comercial		⇔		Acadêmico
Grande		⇔		Pequeno[39]
Em inglês		⇔		Em outro idioma
Longitudinal	⇔	quasilongitudinal	⇔	Crosseccional[40]
Uso imediato		⇔		Uso posterior[41]

Tabela 1.2.1: tipologia de córpus de aprendiz (adaptado de Granger, 2002, p. 51 e 2008, pp. 261-263)[42].

Para Granger (2002), a grande maioria dos córpora de aprendizes pertencem às categorias abaixo:

a) são monolíngues;

b) são escritos;

c) possuem amostras de linguagem produzidas por não especialistas;

[39] A autora não diz precisamente o que entende por córpora pequenos e por córpora grandes, mas indica (1998, p. 174) que o ICLE (*International Corpus of Learner English*), com 200 mil palavras por subcórpus nacional, é um córpus pequeno.

[40] Berber Sardinha (2004) chama a classificação "longitudinal" de "diacrônica" e a "crosseccional" de "sincrônica" (p. 20). O autor não trata da "quasilongitudinal", que é a coleta de dados em um único ponto no tempo, de aprendizes com diferentes níveis de proficiência.

[41] De acordo com esta classificação, no caso do "uso imediato" (p. 263) (*immediate use*) os aprendizes são produtores e usuários dos dados ao mesmo tempo. No caso do "uso posterior" (*delayed use*), os dados serão usados por outros aprendizes.

[42] Tradução e organização minhas.

d) tendem a ser sincrônicos, ou seja, "descrevem o uso feito pelos aprendizes em um ponto específico no tempo" (p. 51)[43]. Granger (2008) introduz a categoria "quasilongitudinal" (p. 263)[44], que se refere à coleta de produções de aprendizes em níveis diferentes de proficiência em um mesmo ponto no tempo.

1.3 O conceito de erro nesta pesquisa

A identificação e a classificação dos erros nas produções escritas de aprendizes não é algo recente na Linguística Aplicada. Desde os anos 1950, alguns linguistas passaram a se interessar especificamente pelo estudo desses erros amparados em aportes teóricos diversos (e.g. Lado, 1957; Selinker, 1972).

Para esta pesquisa, erro é uma falha em se produzir colocações compatíveis com as utilizadas pelos falantes nativos de inglês em contexto de comunicação formal[45]. Como as colocações são definidas pelo uso, entendo o erro como parte do processo de aprendizagem, pois uma determinada colocação pode não ter sido aprendida pelo estudante como tal, quer dizer, como uma sequência típica da linguagem, mas sim como palavras isoladas que ele combina livremente. Entretanto, essa preferência pelas colocações no entendimento de erros não impede que se olhe para ele do ponto de vista gramatical, isto é, a partir da morfologia ou sintaxe, desde que reconheçamos que são duas perspectivas diferentes. Na verdade, nesta dissertação codifico o erro por meio dessas duas perspectivas: gramatical e colocacional, nessa ordem, por meio de dois procedimentos distintos de identificacão de erro (vide metodologia).

O erro, portanto, deve ser analisado colocacionalmente[46], observando-se as colocações usadas tipicamente em um córpus de consulta. Desse modo, o erro não se trata de uma deficiência, mas de um passo necessário ao aprendizado[47], que depende da orientação do professor e do empenho do aprendiz em desenvolver técnicas para adaptar seu conhecimento colocacional nativo àquele da língua estrangeira.

---

[43] *"...describe learner use at a particular point in time*" (p. 51, tradução minha).
[44] *Quasilongitudinal* (p. 263, tradução minha).
[45] Chamo de formal a comunicação escrita ensinada em escolas de idiomas porque visa preparar o aprendiz para se comunicar na variedade padrão da língua inglesa.
[46] Isto é, considerando-se não somente como as palavras da colocação se combinam, mas também quais palavras as antecedem ou sucedem.
[47] Nesta pesquisa emprego o termo aprendizado num sentido neutro, sem entrar na discussão de que se trata de algo que se encerra somente com a morte (cf. Siemens, 2005) e outros.

A seguir, apresento a metodologia utilizada nesta pesquisa.

Capítulo 2: Metodologia

Neste capítulo é apresentada a metodologia empregada na pesquisa, incluindo o contexto de produção, a especificação dos procedimentos de coleta para a criação do córpus COBRA-7, seu recorte, e a identificação e classificação de erros aqui propostas. Primeiramente, são reiterados os objetivos da pesquisa e elencadas as questões que a nortearam. A seguir é detalhado o contexto de produção das redações que compuseram o córpus COBRA-7, seguido do seu procedimento coleta, criação e recorte, o que é seguido da metodologia de identificação, classificação e anotação de erros proposta nesta pesquisa, bem como do calculo de concordância (*agreement*) e do teste estatístico de concordância entre avaliadores.

2.1 Objetivo e questões de pesquisa

A pesquisa tem como objetivo identificar e classificar os erros na escrita de aprendizes brasileiros de inglês. Um dos desdobramentos possíveis desta pesquisa seria a possibilidade de prover aos professores e pesquisadores um sistema de identificação e classificação de erros, com vistas a auxiliá-los em seu trabalho e informar a produção de materiais didáticos locais, isto é, voltados a aprendizes brasileiros.

Dado esse objetivo, as questões de pesquisa a serem investigadas no projeto são elencadas abaixo:

1- Quais os erros mais comuns no córpus COBRA-7_recorte?

2- Qual a variação de erro entre os níveis de curso dos aprendizes no córpus COBRA-7_recorte?

3- Qual nível de curso apresenta maior diversidade de erros no córpus COBRA-7_recorte?

Antes de respondê-las, todavia, convém apresentar o contexto de produção.

2.2 Contexto de produção

A escola de idiomas da qual procedem as redações que compõem o córpus de estudo desenvolvido nesta pesquisa possui quinze unidades, dentre as quais treze funcionam em sistema de franquias. Essa escola de idiomas está disponível em diversos pontos da cidade de São Paulo, além de Guarulhos, as três cidades que compõem o chamado ABC (Santo André, São Bernardo e São Caetano), e Vinhedo, e possui na época da pesquisa (dezembro de 2011) um total de 4.708 aprendizes matriculados (crianças, adolescentes e adultos). Trata-se de uma escola fundada em 1987 e cuja metodologia se baseia na teoria das inteligências múltiplas proposta por Gardner (1985).

A escola oferece aulas de inglês, espanhol e alemão (unidade Santo Amaro somente) a pessoas de idades variadas. Há vários modelos de curso disponíveis, como ilustra o quadro a seguir:

Adolescentes:	Aulas sempre às tardes (horários fixos[48]); a) Duas aulas semanais com duração de 1h e 15 minutos cada até o nível intermediário (*Teen*); b) Duas aulas semanais com duração de 1h e 30 minutos cada a partir do nível intermediário superior (*Teen*).
Adultos:	Aulas de manhã, à noite ou aos sábados (horários fixos); a) Duas aulas semanais com duração de 1h e 30 minutos cada (*Twice*); b) Duas aulas semanais com duração de 3 horas cada (*Vertical*); c) Aulas aos sábados com duração de 3 horas e 30 minutos (*Saturday*); d) Aula individual de 1h e 30 minutos às sextas-feiras (*Flexi*).

Quadro 2.2.1: tipos de curso oferecidos pela escola de idiomas da qual foram extraídas as redações que compuseram o córpus COBRA-7[49].

A franqueadora tenta exercer uma forte política de padronização. Todos os professores a serem contratados pelas unidades devem passar pelo processo seletivo da franqueadora e por

[48] Por "horário fixo" entenda-se a não mobilidade do aprendiz em assistir as aulas em horário que não seja aquele no qual se matriculou. Um aprendiz matriculado, por exemplo, em uma turma às terças e quintas das 20h às 21h30 não poderá frequentar outra turma. Faltas justificadas são repostas fora do horário de aula, individualmente, não necessariamente com o professor da turma.

[49] As letras "a)", "b)" etc. indicam as opções de curso nas quais os aprendizes podem se matricular.

capacitações frequentes. Tal procedimento visa fazer com que o realizado em sala seja muito próximo do prescrito pela escola.

A rede disponibiliza um espaço *online* de acesso restrito ("*web* invisível", cf. Bergh e Zanchetta, 2008, p. 311) no qual os professores inserem as notas e a presença dos aprendizes e os demais colaboradores postam informações pessoais e/ou financeiras.

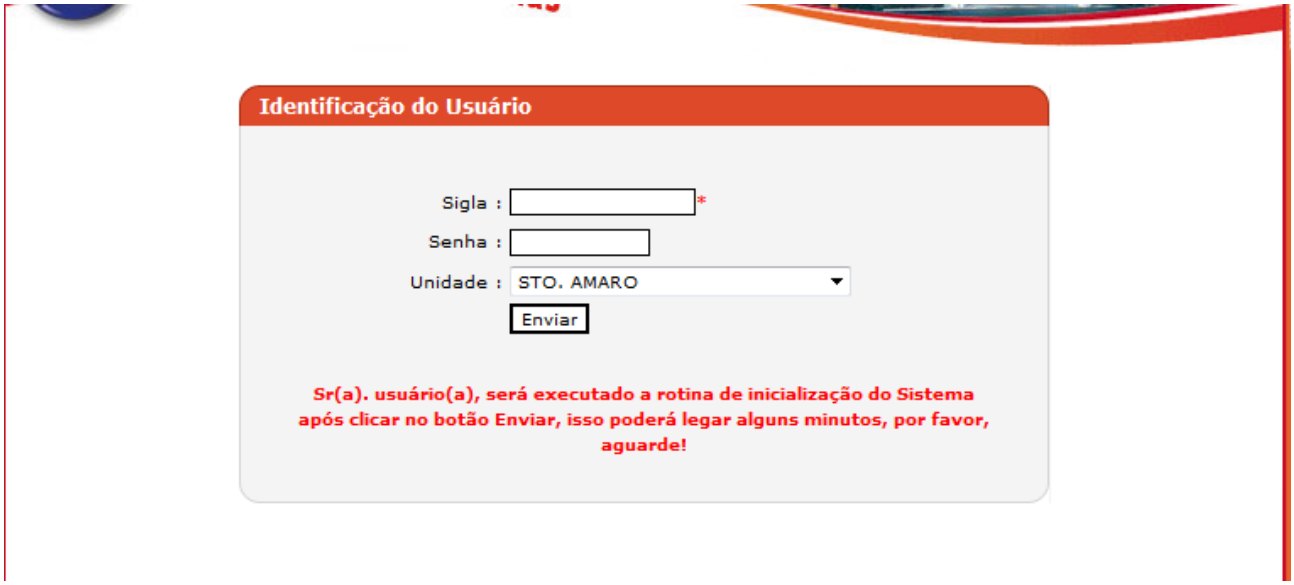


Figura 2.2.1: recorte de tela da área de acesso restrito a colaboradores da escola de idiomas da qual foram coletadas as redações que compuseram o córpus COBRA-7.

Desde 2009 foi englobada a esse sistema uma ferramenta de composição e armazenamento de redações cuja interface se assemelha à do programa computacional de edição de textos *Microsoft Word*, do pacote do *Microsoft Office*. A partir de então a franqueadora passou a recomendar que os professores não mais exigissem que os aprendizes escrevessem as composições em papel no momento da prova. Ao invés disso os docentes deveriam postar o tema da redação no sistema *online* com antecedência mínima de uma semana para que os aprendizes, com seus nomes de usuário e senha, pudessem acessá-lo e realizar a tarefa, isto é, a redação que irá compor parte da nota da prova. Tal sistema foi chamado de *e-portfolio*. As redações compostas nesse sistema permanecem armazenadas por tempo indeterminado no servidor. Esse armazenamento e uso são autorizados pelos aprendizes ou responsáveis legais no ato da assinatura do contrato de matrícula.

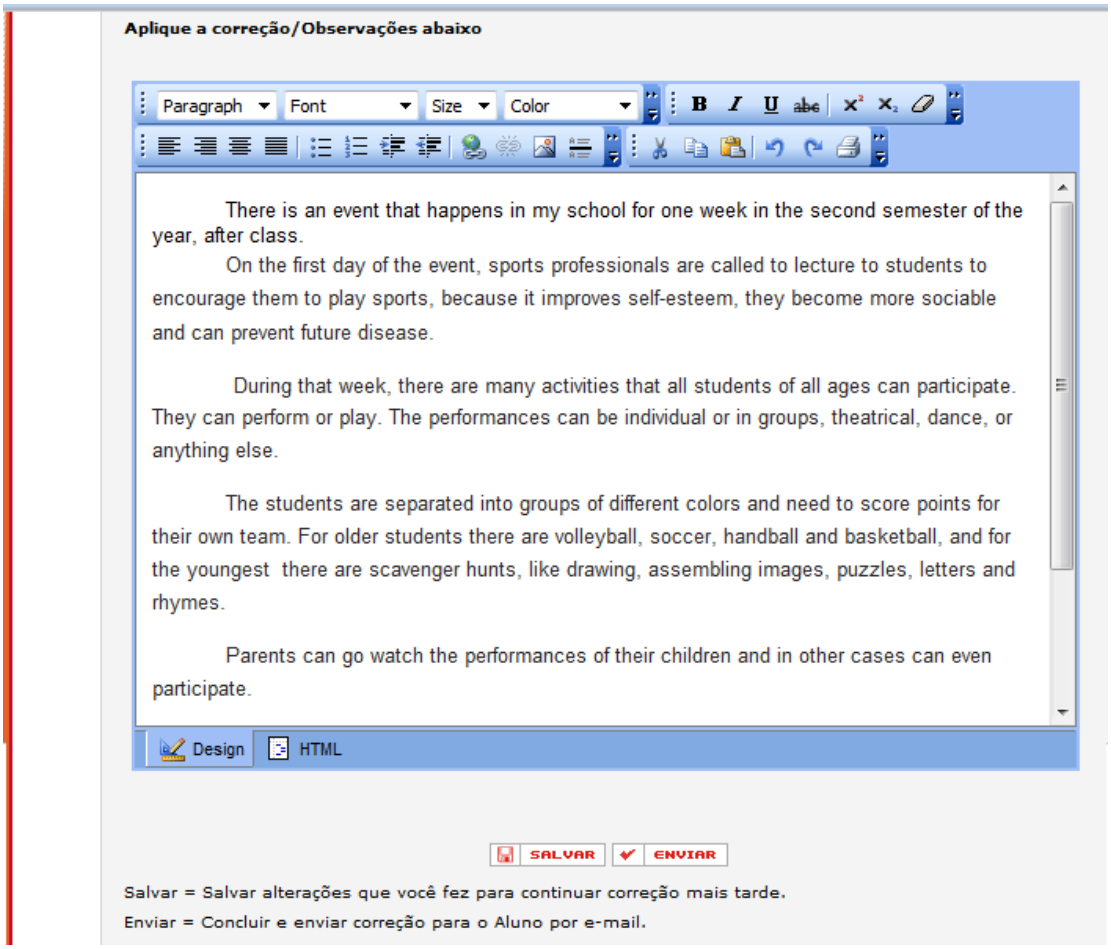


Figura 2.2.2: recorte de tela mostrando a área de composição de textos, dentro do sistema *online* de acesso restrito[50].

Em cada nível de curso nessa escola de idiomas os aprendizes fazem de três a quatro avaliações. Isso significa que fazem também de três a quatro *process writings*[51], cada um composto de pelo menos duas versões[52], todas com mediação do professor. Assim, considerando que a escola tem seis níveis de curso com três ou quatro avaliações cada, é possível dizer que, no final do nível avançado, um aprendiz que tenha estudado desde o básico em uma escola da rede terá feito entre 18 e 24 redações como parte das tarefas e provas exigidas, cada uma delas (teoricamente) com as

[50] Nessa área o aprendiz escreve suas redações, abstendo-se, assim, de escrevê-las durante as provas
[51] Sistema que avalia os processos de criação da redação e considera não somente o produto mas a evolução do aprendiz no processo.
[52] A escola chama de versão o número do rascunho enviado no sistema de *process writing*, como será visto a seguir. O prescrito pela escola é o de que haja três versões de cada redação feita pelos aprendizes, porém foi observado na coleta dos dados que há um número decrescente entre as versões 1, 2 e 3 (total de 2.533 arquivos para a versão 1, 102 arquivos para a versão 2 e 25 arquivos para a versão 3), o que indica que nem todos os professores ou aprendizes seguem o prescrito. Por experiência, posso afirmar que alguns aprendizes enviam a versão 1 somente na data reservada à versão 3.

versões 1, 2 e 3. Mais adiante será explicado o funcionamento do sistema de *process writing* e as versões das redações.

As avaliações feitas pelos aprendizes são compostas de:

a) prova escrita (a qual engloba uma média de três unidades vistas no livro didático correspondente e contempla léxico, gramática, compreensão auditiva e redação);

b) prova oral (na qual são avaliados critérios como fluência, clareza, uso de funções e estrutura gramatical, empatia, entre outros).

Como mencionado anteriormente, é recomendação da franqueadora que a redação (item "a" acima) seja composta no *e-portfolio* em sistema de *process writing*. Esse processo é feito da seguinte maneira:

a) o professor acessa o sistema *online* da rede por meio de seu código funcional e senha, seleciona a turma desejada (necessariamente uma de suas próprias turmas) e posta o tema da redação, as datas de correção e a data final. Há três datas diferentes para correção, uma para cada versão da redação. Por exemplo: versão 1 até 19/03, versão 2 até 21/03, e versão 3 até 25/03. Essas datas são escolhidas pelo professor. As redações podem ser dissertativas, narrativas ou descritivas e cada professor tem liberdade para escolher tanto o tema quanto o tipo de redação, de modo que os temas e os tipos de redação irão variar comparando-se turmas de vários professores que estejam no mesmo nível de curso, ainda que estejam dentro de uma mesma unidade dessa rede de escolas;

b) o aprendiz acessa o sistema por meio do seu código de estudante e senha ou clicando no *link* disponível no e-mail que recebe informando-o da nova tarefa, localiza a atividade, faz a primeira versão da redação e a envia pelo próprio sistema;

c) o sistema envia ao professor da turma um e-mail para avisá-lo de que há redações a serem corrigidas. O docente então acessa novamente o sistema de acesso restrito, localiza e lê cada redação e, utilizando recursos como negrito, mudança de cor, tamanho de fonte, ou ainda

sublinhado, destaca no texto do aprendiz as partes que apresentam erros[53]. Em seguida, o professor reenvia pelo próprio sistema a redação ao aprendiz;

d) o aprendiz, após ser notificado por e-mail da correção, acessa novamente o sistema, observa os destaques feitos pelo professor em seu texto e tenta corrigir os problemas por si só, sem que o professor tenha dito precisamente qual era o erro. Em seguida, o aprendiz reenvia a redação pelo próprio sistema. Essa é a segunda versão da redação;

e) o professor acessa o sistema como feito anteriormente e, novamente, destaca no texto os erros, reenviando-o ao aprendiz pelo próprio sistema;

f) embora nem todos o façam, o aprendiz tem direito a acessar a redação mais uma vez, fazer os ajustes finais com relação aos apontamentos do professor e reenviar a terceira versão da composição, que é a definitiva;

g) o professor acessa o sistema uma última vez, avalia o progresso do aprendiz em corrigir seus próprios erros e atribui uma nota à redação.

O sistema de *process writing* parece permitir ao aprendiz uma maior liberdade para compor seus textos, pois não há a pressão do professor e dos colegas de sala de aula, uma vez que a redação pode ser feita remotamente, a partir de casa ou de qualquer outro local adequado para a execução dessa tarefa. Além disso, pode haver menor limitação e pressão do tempo comparando-se à execução dessa tarefa em sala de aula durante uma prova. Uma outra vantagem desse sistema apresentada pela escola é a de que o professor levaria em conta a evolução e a autonomia do aprendiz ao avaliá-lo. Ademais, como a nota de aprovação nesta escola é 8,0, alcançar a nota máxima na redação (em geral 3,0) pode ser importante na média correspondente à prova realizada e ter grande impacto no final do nível de curso correspondente, determinando a aprovação ou reprovação do aprendiz.

Não há por parte da rede um padrão claro para correção dessas redações e verificação do processo de crescimento dos aprendizes, de modo que os professores recorrem a sistemas diversos que lhes possibilitem executar essa tarefa. Alguns preferem as orientações de Brown (2007, pp. 413-414), que recomenda que na correção de uma redação sejam considerados os seguintes aspectos[54]:

[53] Não é recomendável que, nas verificações, o professor dê a resposta que corija o erro do aprendiz.
[54] Os dados foram adaptados da obra do autor.

a) **conteúdo:** desenvolvimento da tese (em caso de composições dissertativas), uso de causa e efeito, foco;

b) **organização**: eficiência da introdução, sequência lógica das ideias, tamanho apropriado e desenvolvimento da conclusão;

c) **discurso:** uso de *topic sentences* e marcadores discursivos (*well*, *now*, *you know*, *I believe that*), referências, variação;

d) **sintaxe:** como é feita a escolha lexical (escolha de palavras) e a estruturação das sentenças;

e) **vocabuário:** se as palavras usadas correspondem ao nível de curso do aprendiz e se as colocações se aplicam;

f) **mecânica:** grafia, pontuação, organização visual.

Sistemas de correção como o proposto por Brown (2007) são interessantes porque permitem ao professor uma uniformização do critério de avaliação, além de não contemplarem apenas questõees estruturais.

Brown (2007, p. 414) recomenda a seguinte atribuição de nota para cada item: conteúdo (0 a 24); organização (0 a 20); discurso (0 a 20); sintaxe (0 a 12); vocabulário (0 a 12); mecânica (0 a 12), cuja totalização das notas máximas seria 100. A nota adequada para cada redação da escola em referência pode ser encontrada fazendo-se uma regra de três simples a partir dos dados numéricos acima. Por exemplo, em uma prova cuja redação vale entre 0,0 e 3,0 um professor da rede que utilize o sistema de atribuição de notas proposto pode Brown (2007) pode usar a seguinte tabela como modelo:

conteúdo:	0,0 a 0,72
organização:	0,0 a 0,60
discurso:	0,00 a 0,60
sintaxe:	0,0 a 0,36
vocabulário:	0,0 a 0,36
mecânica:	0,0 a 0,36
TOTAL:	0,0 a 3,00

Tabela 2.2.1: tabela de notas para redação com nota máxima 3,0, baseada na proposta de Brown (2007).

Nessa escola o relacionamento professor-aprendiz é bastante informal. Essa particularidade é notada em todos os ambientes das escolas, tanto na decoração quanto na interação com os demais colaboradores. Por isso, as aulas devem ser descontraídas e repletas de jogos ou atividades extra-livro. Nessa escola, espera-se, por exemplo, que crianças, jovens ou adultos participem de atividades como correr até a lousa e tocar uma palavra específica ou ainda montar quebra cabeças, competindo ou não com outro grupo criado a partir da divisão dos membros da mesma sala da aula. Atividades como essas estão diretamente relacionadas ao método no qual a escola acredita (sete inteligências).

Do ponto de vista linguístico a instituição se afilia à abordagem comunicativa, pois preza a simulação de situações cotidianas e dá importância ao ensino de funções de linguagem e colocações. Enquadra-se, portanto, dentro da visão empírica da linguagem.

Os professores da rede precisam constantemente preparar as aulas preenchendo um formulário específico criado pela instituição (figura a seguir)[55] e trazer materiais como tiras de papel para atividades interativas, notícias de jornais ou revistas *online* e outros recursos. Do mesmo modo, as atividades escritas, como as redações, em geral também possuem um caráter informal.

[55] Esse procedimento é controlado mediante a solicitação de planos de aula e a observação de aulas por parte da coordenação de cada unidade.

[redacted]

CLASS PLAN

Date: _____ Book: _____ Unit: _____ Page: _____ Lesson: _____ Teacher: _____

Class profile	
Timetable fit	
Aim	

TIME	STAGE	OBJECTIVES	PROCEDURES	ANTICIPATED PROBLEMS & SOLUTIONS	INTER ACTION	MATERIAL	MI

Figura 2.2.3: modelo de planejamento de aulas para professores da rede de escolas na qual trabalho[56].

A seguir falarei sobre a compilação do córpus COBRA-7, que estará disponível em breve no sítio do CEPRIL, do LAEL (PUC-SP).

2.3 Compilação do córpus COBRA-7

Como visto, o objetivo deste trabalho era identificar e classificar os erros na escrita de aprendizes brasileiros de inglês. Um dos desdobramentos possíveis desta pesquisa seria a possibilidade de prover aos professores e pesquisadores um sistema de identificação e classificação de erros, com vistas a auxiliá-los em seu trabalho e informar a produção de materiais didáticos locais, isto é, voltados a aprendizes brasileiros. Para atingir esse objetivo era necessária uma análise em um córpus de estudo que contivesse redações de aprendizes brasileiros de inglês como língua estrangeira matriculados em uma escola de idiomas e estivessem em níveis de curso diversos. Por trabalhar em uma escola com essas características e que armazena em formato digital as composições dos seus aprendizes optei por coletá-las e criar o córpus de COBRA-7.

Para a compilação do córpus COBRA-7 foram buscadas[57] no sistema *online* (*e-portfolio*) da escola de idiomas na qual trabalho as redações de aprendizes postadas entre 2009 e 2010. Para isso,

[56] Por razões legais, a quadrado preto cobre a logomarca da escola. A última coluna (MI) se refere a qual inteligência é utilizada na atividade (visual, lógica, corporal etc.)

[57] Como visto na Fundamentação Teórica, Bergh e Zanchetta (2008) propõem os termos "busca", "pastoreio" ou "folheio" para a localização de arquivos na Internet para compor um córpus. Como nenhum dos termos descreve o processo de localização, seleção e cópia executado para a compilação do córpus COBRA-7, optei por usar aleatoriamente os termos "buscar", "localizar" e "acessar".

primeiramente acessei o servidor *online* de acesso restrito da escola por meio de nome de usuário e senha funcional.

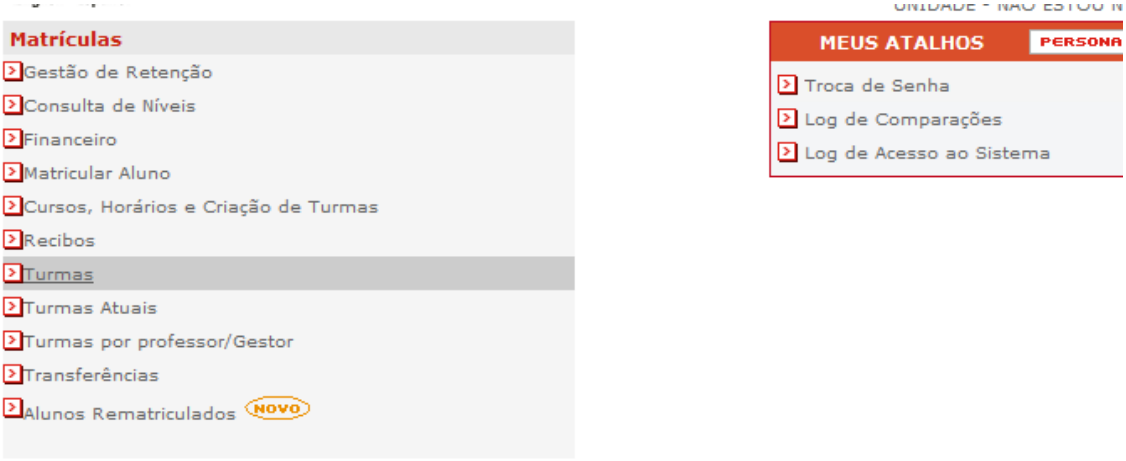


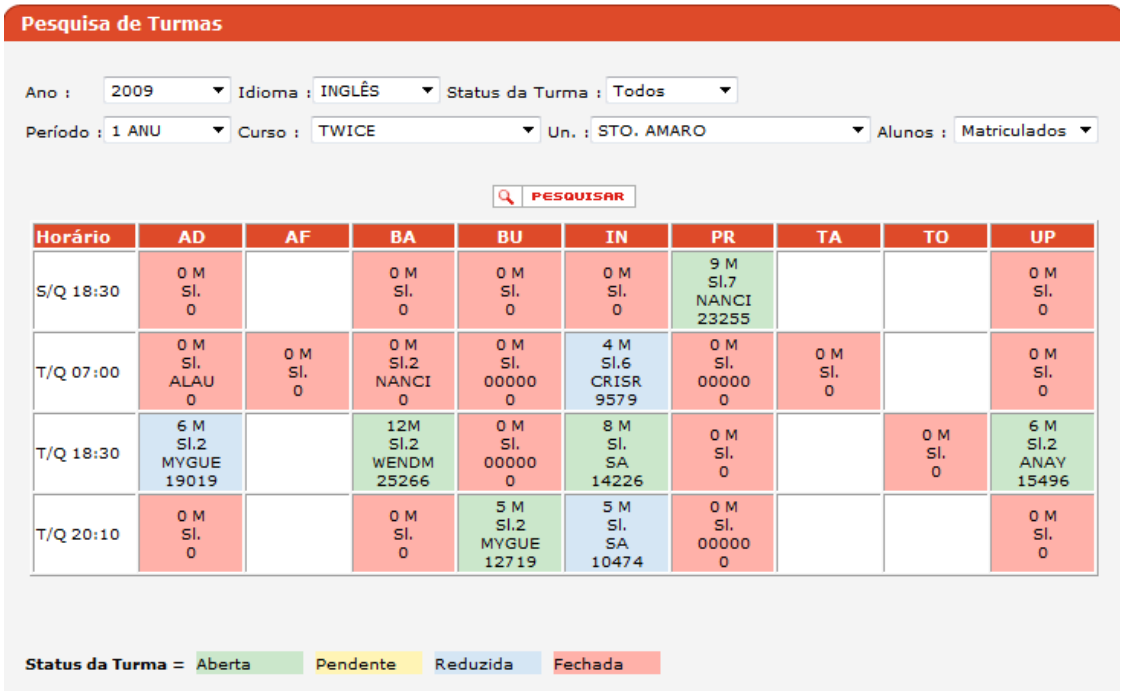
Figura 2.3.1: recorte de tela mostrando a página inical do sistema *online* de acesso restrito da escola de idiomas de onde foram extraídas as redações que compuseram o córpus COBRA-7[58].

Em seguida, acessei o espaço virtual de cada uma das unidades da instituição, como mostra a figura a seguir:

Figura 2.3.2: recorte de tela mostrando a página na qual são escolhidos o ano de produção, o idioma, a situação de turma, o período (anual, semestral, mensal, curso de curta duração), o curso (*Saturday*, *Teen*, *Twice*, *Flexi*, Colégios etc.) e a unidade da rede a ser pesquisada.

[58] Em destaque (linha cinza) o link "turmas", que permitiu o acesso às turmas das diversas unidades que compõem a rede de escolas.

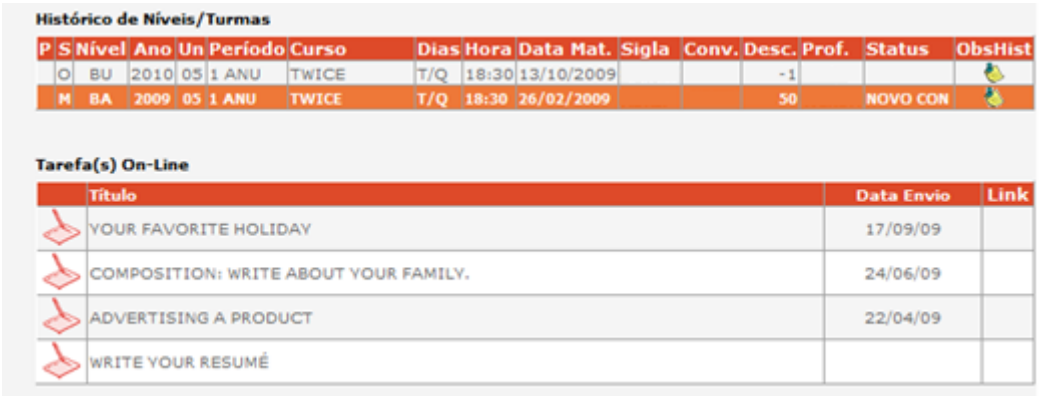
Por fim, busquei uma a uma, no espaço virtual de cada uma das unidades da escola, as redações correspondentes aos níveis básico, básico 2, pré-intermediário, intermediário, intermediário superior, e avançado. A figura abaixo mostra um quadro geral das turmas disponíveis na unidade Santo Amaro em 2009 para as turmas anuais para adultos com aulas duas vezes por semana. Clicando sobre cada quadrinho da tabela o funcionário da rede tem acesso às composições dos aprendizes matriculados nas respectivas turmas.



Horário	AD	AF	BA	BU	IN	PR	TA	TO	UP
S/Q 18:30	0 M Sl. 0		0 M Sl. 0	0 M Sl. 0	0 M Sl. 0	9 M Sl.7 NANCI 23255			0 M Sl. 0
T/Q 07:00	0 M Sl. ALAU 0	0 M Sl. 0	0 M Sl.2 NANCI 0	0 M Sl. 00000 0	4 M Sl.6 CRISR 9579	0 M Sl. 00000 0	0 M Sl. 0		0 M Sl. 0
T/Q 18:30	6 M Sl.2 MYGUE 19019		12M Sl.2 WENDM 25266	0 M Sl. 00000 0	8 M Sl. SA 14226	0 M Sl. 0		0 M Sl. 0	6 M Sl.2 ANAY 15496
T/Q 20:10	0 M Sl. 0		0 M Sl. 0	5 M Sl.2 MYGUE 12719	5 M Sl. SA 10474	0 M Sl. 00000 0			0 M Sl. 0

Figura 2.3.3: recorte de tela mostrando as turmas anuais para adultos, com aulas duas vezes por semana, disponíveis em 2009 na unidade Santo Amaro[59].

[59] Dentro da tabela, a sigla composta por um número seguido da letra "M" (0 M, ou 5 M, por exemplo), indicam quantos aprendizes matriculados há na turma. Turmas com "0 M" não foram abertas.



Histórico de Níveis/Turmas

P	S	Nível	Ano	Un	Período	Curso	Dias	Hora	Data Mat.	Sigla	Conv.	Desc.	Prof.	Status	ObsHist
	O	BU	2010	05	1 ANU	TWICE	T/Q	18:30	13/10/2009			-1			
	M	BA	2009	05	1 ANU	TWICE	T/Q	18:30	26/02/2009			50		NOVO CON	

Tarefa(s) On-Line

	Título	Data Envio	Link
	YOUR FAVORITE HOLIDAY	17/09/09	
	COMPOSITION: WRITE ABOUT YOUR FAMILY.	24/06/09	
	ADVERTISING A PRODUCT	22/04/09	
	WRITE YOUR RESUMÉ		

Figura 2.3.4: recorte de tela mostrando as redações de um aprendiz específico da turma de nível básico com aulas às terças e quintas, das 18h30 às 20h[60].

A seguir, selecionei[61] cada texto e efetuei um comando simples de cópia, como visto na Fundamentação Teórica e ilustrado na figura abaixo:

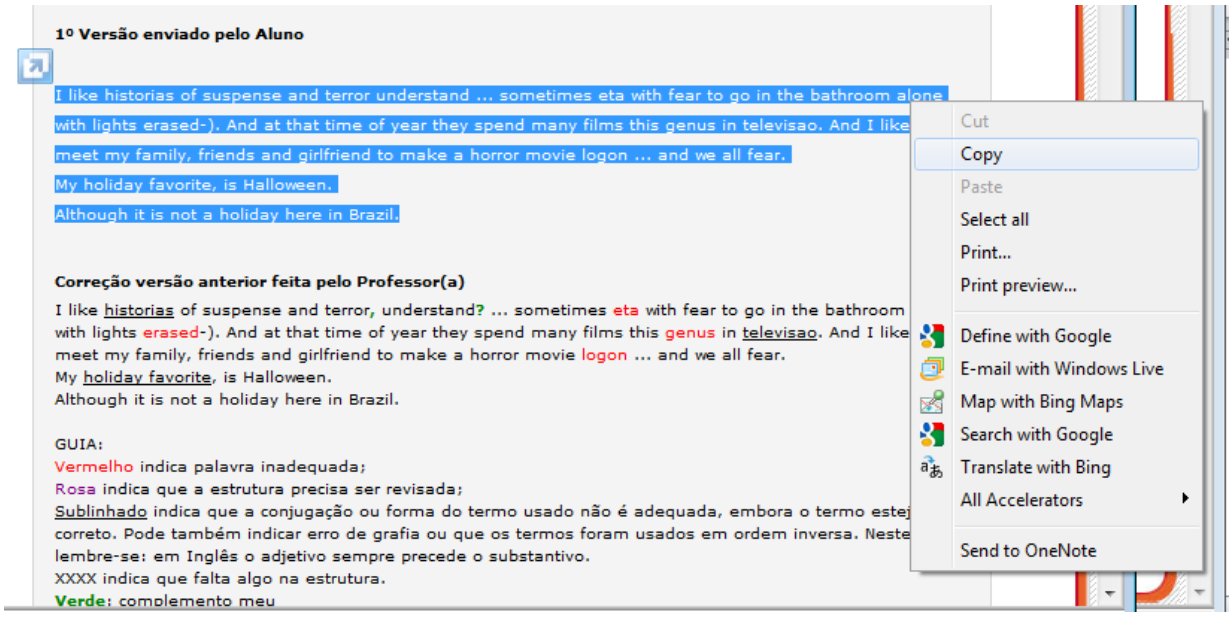


Figura 2.3.5: recorte de tela mostrando uma redação de um aprendiz específico da turma de nível básico com aulas às terças e quintas entre 18h30 e 20h[62].

[60] Clicando sobre os títulos das redações o funcionário da escola de inglês tem acesso às redações e a todas as correções sugeridas e efetuadas pelos aprendizes.

[61] Granger (2002) e O'Keeffe, McCarthy e Carter (2007) prevêm como fontes de dados para córpus de aprendizes o download eletrônico, o escaneamento ou a digitação. O método de coleta das redações para o córpus COBRA-7 não foi contemplado por esses autores, mas foi citado por Sinclair (1991), que prevê também a "adaptação de material que já esteja em formato eletrônico" (p. 14, tradução minha para "*adaptation of material already in electronic form*").

[62] Em destaque vê-se a versão 1 da redação já selecionada e sendo copiada para ser posteriormente colada em um arquivo .TXT. Logo abaixo vê-se a mesma redação já com os comentários do professor, a serem submetidos para correção pelo próprio aprendiz.

É importante dizer que esse procedimento de busca, seleção e cópia de cada redação foi necessário pelo fato de as redações compostas pelos aprendizes da escola de idiomas na qual trabalho estarem armazenadas em um banco de dados. Cabe aqui uma distinção entre banco de dados e córpus: um banco de dados é um conceito de informática. Trata-se de um sistema de gerenciamento de informação. Nele, o conteúdo é acessado por meio de um programa específico apenas, e o conteúdo em geral está distribuído separadamente, de modo que não se tem acesso aos dados diretamente. Um córpus, por outro lado, é uma coletânea de textos com base em critérios específicos para ser usada em pesquisa linguística. O conteúdo pode ser acessado por meio de vários programas, diretamente, ou então o córpus pode ser disponibilizado em pastas ou arquivos compactados. Por isso, foi a partir das redações, coletadas a partir de um banco de dados, que criei o córpus COBRA-7.

Ao mesmo tempo em que selecionei e copiei cada texto, utilizando uma planilha eletrônica do programa computacional *Microsoft Excel*, do pacote *Microsoft Office 2010*, atribuí a cada aprendiz pesquisado um número único com cinco dígitos numéricos obrigatórios e uma letra opcional, que correspondiam, cronologicamente, à ordem em que suas redações foram postadas no servidor *online* (figura a seguir), visto que um mesmo aprendiz pode ter entre três e quatro redações por nível de curso, como visto. Desse modo, enquanto a primeira redação do aprendiz 00997 não recebeu como representação letra alguma, a segunda redação desse mesmo autor recebeu como representação a letra "b", a terceira redação a letra "c", e assim sucessivamente, de modo que os nomes dos arquivos correspondentes às redações desse usuário foram, respectivamente, 00997.txt, 00997b.txt e 00997c.txt[63] (esta é a redação em destaque na figura anterior). Esse procedimento de uso de números e letras como nomes dos arquivos foi usado para facilitar a localização das composições e possibilitar a estudos futuros que se verifique o progresso de um determinado aprendiz diacronicamente, isto é, conforme avança entre os níveis.

[63] As letras que acompanham os números que definem cada aprendiz não se referem às versões das redações, mas sim a cada uma das provas feitas dentro de um nível de curso específico. Segundas e terceiras versões foram arquivadas dentro de pastas internas específicas chamadas "v2" e "v3", respectivamente, e receberam os mesmos nomes dos arquivos da versão 1. Usando novamente como exemplo a redação 00997c.txt, suas segunda e terceira versões, se houvesse, seriam também chamadas 00997c.txt, mas, por estarem inseridas dentro de uma outra pasta, não gerariam erro de repetição de nome de arquivo.

Numero	Aluno	Unidade
00001	NONONONONONO	Aclimação
00002	NONONONONONO	Aclimação
00003	NONONONONONO	Aclimação
00004	NONONONONONO	Aclimação
00005	NONONONONONO	Aclimação
00006	NONONONONONO	Aclimação
00007	NONONONONONO	Aclimação
00008	NONONONONONO	Aclimação
00009	NONONONONONO	Aclimação
00010	NONONONONONO	Aclimação
00011	NONONONONONO	Aclimação
00012	NONONONONONO	Aclimação
00013	NONONONONONO	Aclimação
00014	NONONONONONO	Aclimação
00015	NONONONONONO	Aclimação
00016	NONONONONONO	Aclimação

Figura 2.3.6: atribuição de números a cada aprendiz pesquisado (os nomes dos aprendizes foram substituídos para lhes preservar a identidade)

Para cada texto copiado do sistema *online* da escola de inglês na qual trabalho e colado em um arquivo *.TXT*[64] nomeado segundo os critérios supracitados (código que definiu cada aprendiz seguido ou não de uma letra) foram inseridas informações em etiquetas do tipo *COCOA* para compor o cabeçalho, como visto na Fundamentação Teórica e reproduzido abaixo:

```
<Sexo M>
<Idade Adulto>
<Nível Básico>
<Grupo Saturday>
<Tarefa WRITING QUIZ 2>
<Fonte Seven Aclimação>
<Ano prod 2008>
<Ano coleta 2010>
<Fornecedor http://seven.digiweb.com.br>
<Versão da composição 1>
<Versão do Córpus 2>
<Língua Inglesa>
<Tipo de texto: Corpora de Aprendizes>
<Córpus COBRA-7 - By Wendel Mendes Dantas - LAEL - PUC-SP - 2010-2012 - Mestrado>
```

Figura 2.3.7: exemplo de cabeçalho em formato *COCOA*, usado usado no córpus COBRA-7.

[64] Como visto na Fundamentação Teórica, arquivos em formato TXT são mais fáceis de se trabalhar porque perdem a formatação inicial de espaço entre parágrafos, tamanho de texto, figuras inseridas e outras, mantendo simplesmente o texto.

Esse tipo de organização, além de trazer as informações básicas do arquivo, como fonte, tipologia, pesquisador responsável e data de compilação (cf. Sinclair, 1996), permite localizar facilmente, por meio do uso de programas computacionais como o *Wordsmith Tools 5.0* (Scott, 2008), as informações disponíveis dentro dos sinais "<>" , como o ano de produção, o gênero (sexo) do autor, e o grupo ao qual o aprendiz pertencia.

Em seguida, cada arquivo com extensão *.TXT* foi guardado dentro de pastas cujo caminho no sistema para a plataforma *Windows XP* foi: C:\corpora\COBRA-7\2\txt\adulto_por_niveis. Essas pastas representavam o nível de curso ao qual pertenciam as redações, como mostra a figura a seguir:

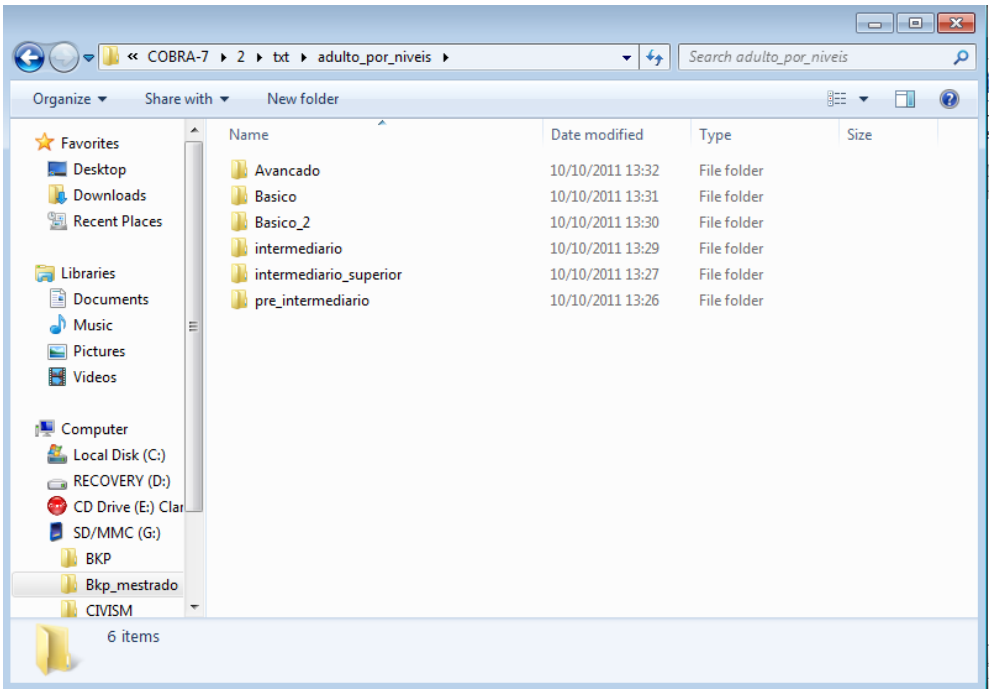


Figura 2.3.8: recorte de tela mostrando, na pataforma *Windows*, as pastas dentro das quais estão disponíveis, por níveis, os arquivos que compõem o córpus COBRA-7[65].

Algumas redações continham informações pessoais como nomes completos, telefones, endereços, locais de trabalho etc. Nesses casos, essas informações foram substituídas pela etiqueta <*personal-info-deleted*>. Isso foi feito para se preservar as informações pessoais, por questões éticas.

Algumas redações tinham versões 2 e 3. Nesses casos, foi executado o mesmo procedimento já citado de acesso ao sistema, localização, cópia, colagem do texto em um arquivo *.TXT* e

[65] O caminho no sistema para as pastas em destaque são: C:\corpora\COBRA-7\2\txt\adulto_por_niveis.

nomeação de arquivo, mas com o diferencial de que esses arquivos receberam os mesmos nomes do arquivo da versão 1, mas foram armazenados dentro de uma pasta interna criada dentro do diretório correspondente ao nível de curso, como mostra a figura a seguir:

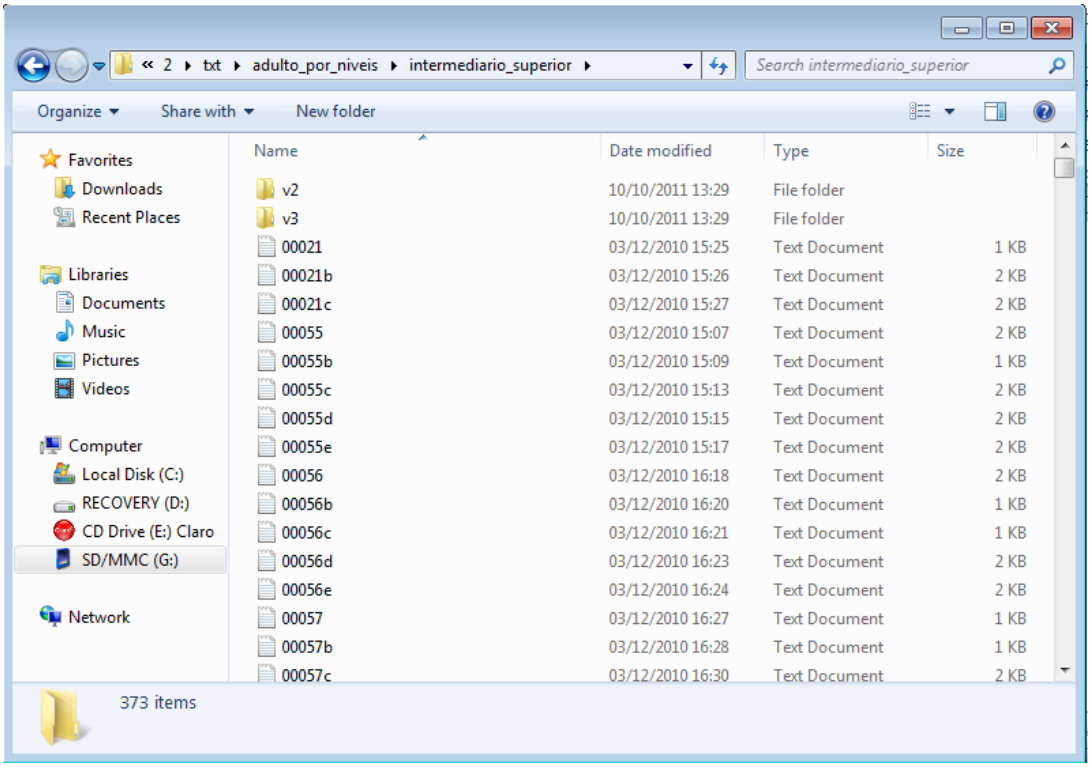


Figura 2.3.9: recorte de tela mostrando as pastas internas que contêm as versões 2 (v2) e 3 (v3) das redações do nível intermediário superior[66].

Nesses casos, quando essas redações apresentavam informações pessoais, foi realizada em cada uma das versões a substituição dessas informações pela etiqueta *<personal-info-deleted>*, como explicado anteriormente.

Com esses procedimentos foi compilado o córpus COBRA-7, que possui 2660 arquivos, 571.564 palavras, o que, dividindo-se pelo número total de arquivos (2660) totaliza uma média aproximada de 215 palavras por redação.

A tabela a seguir ilustra mais claramente a composição desse córpus:

[66] Esse mesmo padrão foi seguido para os outros níveis cujas segundas e terceiras versões foram coletadas, a saber: níveis pré-intermediário e intermediário.

Nível:	**Nº de arquivos**	**Nº de palavras**	**Versões:**
Avançado:	328 arquivos	96.186	Apenas versão 1
Intermediário superior:	371 (v1), 23 (v2) e 06 (v3)	106.709 (versões 1, 2 e 3) 98.664 (apenas versão 1)	Versões 1, 2 e 3
Intermediário:	474 (v1), 35 (v2), 07 (v3)	126.437 (versões 1, 2 e 3) 116.341 (apenas versão 1)	Versões 1, 2 e 3
Pré-intermediário:	363 (v1), 44 (v2), 12 (v3)	90.338 (versões 1, 2 e 3) 77.065 (apenas versão 1)	Versões 1, 2 e 3
Básico 2:	529 arquivos	91.246	Versões 1, 2 e 3
Básico:	468 arquivos	60.570	Apenas versão 1
TOTAL:	2660	571.486[67]	

Tabela 2.3.1: quantidade de arquivos e versões das composições no córpus COBRA-7 por nível[68].

Considerando os critérios tipológicos de Berber Sardinha (2004) e Granger (1998, 2002, 2008), vistos na fundamentação teórica, pode-se definir o córpus COBRA-7 como:

Um córpus médio, monolíngue, sincrônico, quasilongitudinal , estático e contemporâneo, que serve de amostragem de composições escritas de aprendizes de inglês como língua estrangeira e possui a finalidade de estudo.

2.4 Elaboração de uma metodologia de classificação e identificação de erros

Para mostrar os erros mais comuns aos aprendizes brasileiros em cada nível de curso foi preciso desenvolver uma metodologia de classificação e identificação de erros que possa também ser útil ao professor para o "diagnóstico" das dificuldades dos seus aprendizes e o desenvolvimento de técnicas para a redução dos erros.

[67] Este total foi feito somando-se o total de palavras de todas as redações. Os valores 98.664, 116.341 e 77.065 foram removidos, pois correspondiam somente ao número de palavras das versões 1 das redações dos níveis intermediário superior, intermediário e pré-intermediário, respectivamente. Tais valores já constam no número das respectivas linhas anteriores.

[68] As abreviações "v1", "v2" e "v3" indicam as versões das redações no sistema *process writing*: versões 1, 2 e 3, respectivamente.

Para que isso fosse feito seria necessária a análise de textos dos quais poderiam ser retirados e classificados os erros mais comuns dos aprendizes. Pelo caráter científico desta pesquisa, esses textos deveriam provir de um córpus, isto é, uma coletânea de textos em formato eletrônico coletados e armazenados segundo critérios discutidos na Fundamentação Teórica deste trabalho. Por isso, e por esta pesquisa enfocar a produção de aprendizes, o córpus COBRA-7 seria a melhor escolha.

2.4.1 Recorte do córpus COBRA-7

Inicialmente a intenção era que o córpus COBRA-7 fosse utilizado em sua totalidade para essa finalidade. Todavia, uma análise piloto de uma amostra desse córpus demonstrou que havia muitas ocorrências de erro, o que levou a um recorte do córpus, definido em 300 redações, sendo 50 de cada um dos níveis de curso que o compõem, a saber: básico 1, básico 2, pré-intermediário, intermediário, intermediário superior e avançado. Tal recorte pode parecer pequeno, porém, a análise das 300 redações propostas (aproximadamente 10% do córpus) mostrou um total de 3854 ocorrências de erros. Se a proporção se mantivesse no restante do córpus, seriam mais de 38.000 ocorrências a serem analisadas e classificadas manualmente, sem o auxílio de programas computacionais, o que não seria possível dado o tempo reduzido do mestrado. Por isso, os quase quatro mil erros encontrados já parecem suficientes, dentro das restrições apresentadas, para traçar um perfil dos erros cometidos pelos aprendizes do córpus COBRA-7 e servir de modelo para a criação da metodologia aqui proposta.

O recorte foi feito da seguinte maneira:

a) acessei, na plataforma *Windows XP*, o caminho C: \corpora\COBRA-7\2\txt\ e criei uma pasta chamada "recorte_para_metodologia", dentro da qual criei pastas internas com os seguintes nomes: basico, basico_2, pre_intermediario, intermediario, intermediario_superior, avancado[69];

b) dentro do caminho no sistema mencionado no item anterior acessei o conteúdo da pasta "adulto_por_niveis", que já existia. Essa pasta é composta de pastas internas com os nomes de cada nível de curso cujas redações compõem o córpus COBRA-7. Dentro de

[69] Como visto anteriormente, convencionou-se não usar acentuação e separação de palavras na criação de pastas na Linguística de Córpus. Esse procedimento foi emprestado da área da computação e permite, entre outras coisas, que os dados sejam acessíveis a qualquer sistema operacional.

cada pasta interna havia as redações correspondentes, dispostas por ordem crescente conforme o nome do arquivo;

c) acessei a pasta interna correspondente ao nível de curso básico 1;

d) selecionei as 50 primeiras redações e efetuei um comando simples de cópia de arquivos utilizando o botão direito do *mouse*;

e) acessei novamente o caminho base C: \corpora\COBRA-7\2\txt\ e acessei desta vez a pasta "recorte_para_metodologia";

f) acessei a pasta interna correspondente ao nível de curso correspondente às 50 redações que haviam sido copiadas;

g) fiz a colagem dos arquivos por meio de um comando simples com o botão direito do *mouse* do computador;

h) acessei no caminho C: \corpora\COBRA-7\2\txt\adulto_por_niveis a pasta interna do nível seguinte (neste momento basico_2) e repeti os procedimentos dos itens "d" a "g" (acima);

i) repeti o item "h" acima para cada um dos níveis de curso cujas redações compõem o córpus COBRA-7.

Com esse procedimento, criei um recorte do córpus COBRA-7, um mini-córpus para a elaboração da metodologia de classificação e identificação de erros objetivada neste trabalho. Para evitar confusões, optei por dar a essa amostra o nome de COBRA-7_recorte, termo que será utilizado doravante sempre que me referir a ele.

O córpus COBRA-7_recorte possui 61.240 palavras[70]. Dividindo-se esse número pela quantidade de redações (300) pode-se dizer que se trata de um córpus com uma média aproximada de 204 palavras por redação.

A tabela adiante resume as características do córpus COBRA-7_recorte:

[70] Os números de palavras do córpus e de cada nível de curso foram conseguidos utilizando-se o programa computacional *Concord*, da suíte do *Wordsmith Tools 5.0* (Scott, 2008).

Nível:	**Nº de arquivos**	**Nº de palavras**	**Versões:**	**Erros:**
Avançado:	50 arquivos	15.197	Apenas versão 1	497
Intermediário superior:	50 arquivos	11.672	Apenas versão 1	635
Intermediário:	50 arquivos	11.770	Apenas versão 1	918
Pré-intermediário:	50 arquivos	10.108	Apenas versão 1	760
Básico 2:	50 arquivos	6.364	Apenas versão 1	627
Básico:	50 arquivos	6.129	Apenas versão 1	417
TOTAL:	300	61.240		3854

Quadro 2.4.1: quadro-resumo do córpus COBRA-7_recorte.

O quadro abaixo traça uma comparação entre o córpus COBRA-7, córpus fonte deste trabalho, e o córpus COBRA-7_recorte, córpus de análise.

COBRA-7	COBRA-7_recorte
Nº de arquivos: 2660	Nº de arquivos: 300
Nº de palavras: 571.486	Nº de palavras: 61.240
Versões 1, 2 e 3 de parte do córpus	Apenas versão 1 do *process writing*

Quadro 2.4.2: comparação entre o COBRA-7 e o COBRA-7_recorte.

2.4.2 Córpus de consulta

Além do córpus de estudo, o objetivo agora descrito requeriu um córpus de consulta (ou de referência), o COCA (*Corpus of Contemporary American English*, como visto), que continha na época em que esta pesquisa foi realizada aproximadamente 414 milhões de palavras e é composto

de produções orais e escritas da variação norte-americana do inglês compiladas desde 1990 nos Estados Unidos a partir de textos ficcionais, revistas, jornais, periódicos acadêmicos, e transcrições de entrevistas de televisão e programas de rádio. Córpus de consulta é aquele utilizado para se checar determinado padrão léxico-gramatical encontrado em um córpus de estudo, neste caso o COBRA-7_recorte.

O COCA não pode ser baixado pelos usuários do córpus, por isso foi consultado *online* como parte da metodologia de classificação e identificação de erros desenvolvida nesta pesquisa.

Optei por utilizar o COCA como córpus de consulta desta pesquisa porque:

1. o COCA está disponível na Internet gratuitamente, o que facilita o acesso a ele;

2. o COCA é constantemente atualizado, acompanhando, assim, a evolução da língua em uso;

3. o COBRA-7 contém composições que não são acadêmicas e que abordam assuntos diversos. Por isso, o COCA, por englobar um número amplo de gêneros, pode prover um melhor parâmetro com relação ao que se espera dos aprendizes uma vez que exige-se deles também a produção de textos de diversos gêneros;

4. trata-se de um córpus de consulta já utilizado pelos pesquisadores do Grupo de Estudos de Linguística de Córpus (GELC), da PUC-SP, dentre os trabalhos é possível citar Berber Sardinha e Shepherd (2011).

A seguir detalharei cada passo do processo da metodologia de classificação e identificação de erros utilizada nesta pesquisa.

2.4.3 Desenvolvimento da metodologia

A metodologia aqui proposta será apresentada em duas partes. Primeiramente falarei sobre a identificação dos erros e, mais adiante, sobre sua análise a classificação.

2.4.3.1 Identificação dos erros

Os erros foram identificados da seguinte maneira:

Primeiramente, usando a plataforma *Windows XP*, acessei o caminho no sistema C:\corpora\COBRA-7\2\txt\recorte_para_metodologia com o objetivo de localizar as pastas correspondentes aos níveis de curso que compõem o córpus COBRA-7_recorte: básico 1, básico 2, pré-intermediário, intermediário, intermediário superior, avançado.

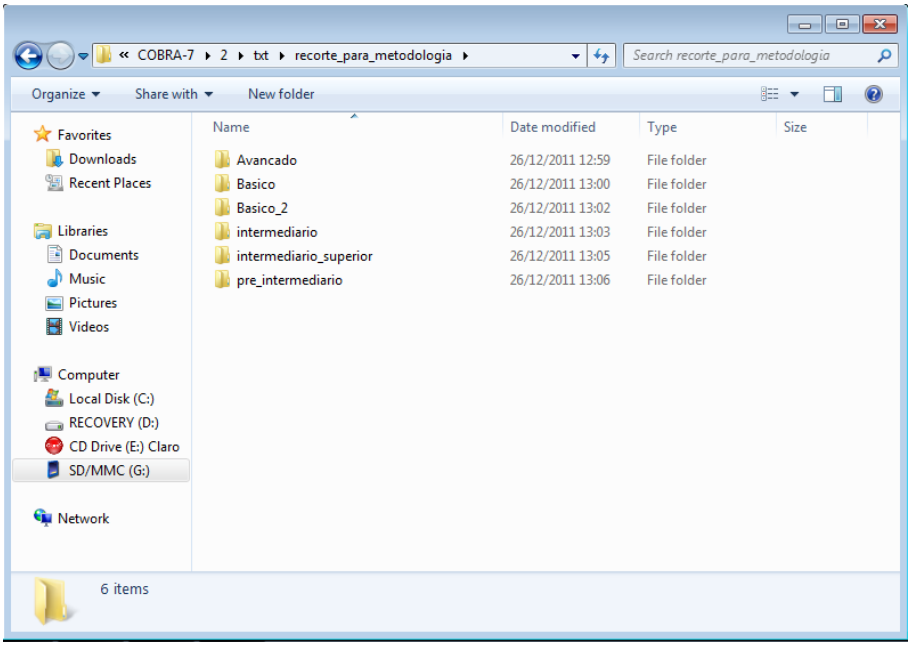


Tabela 2.4.1: impressão de tela mostrando as pastas correspondents aos níveis de curso analisados[71].

Em seguida, acessei cada uma das pastas correspondentes aos níveis de curso e, em cada uma, abri e li cada uma das 50 redações que as compunham.

[71] As pastas foram acessadas por meio do caminho virtual C:\corpora\COBRA-7\2\txt\recorte_para_metodologia, na plataforma (sistema operacional) Windows XP

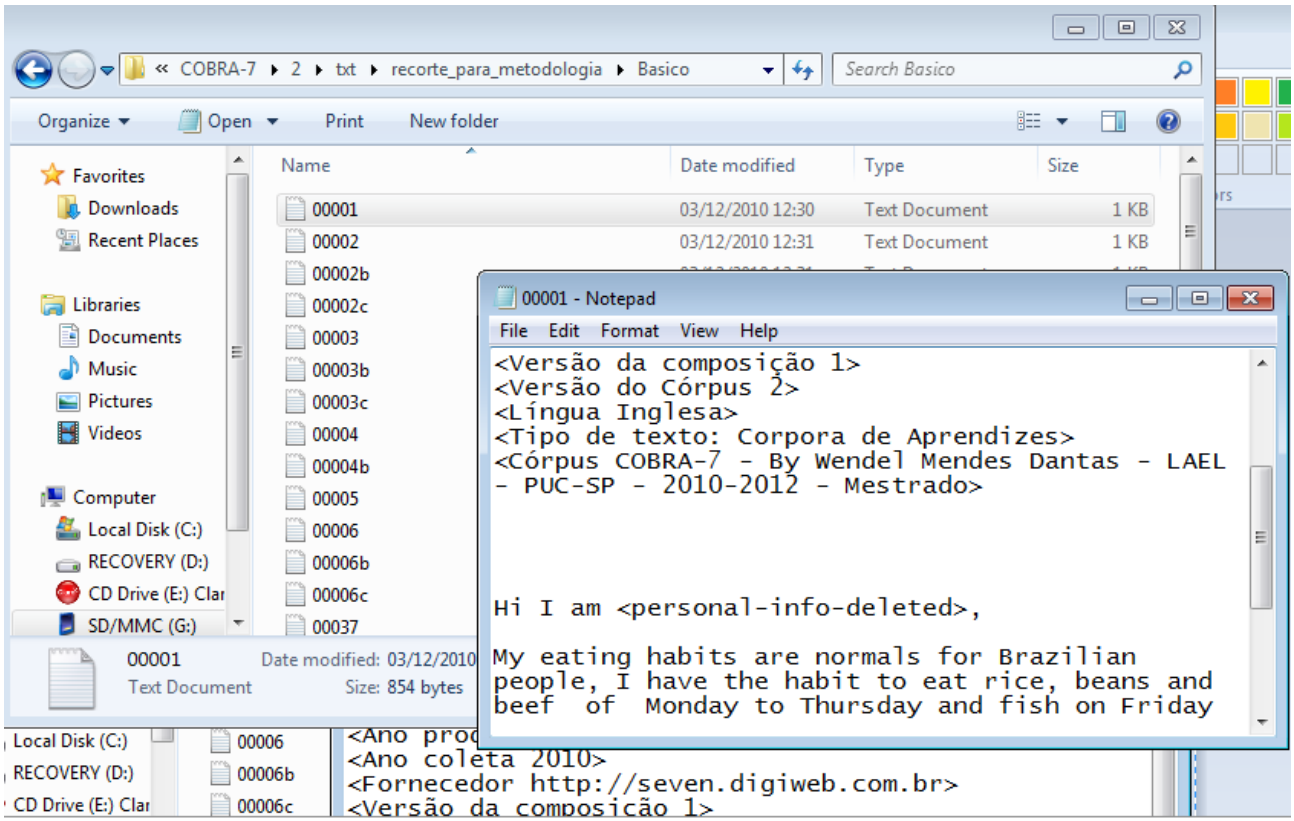


Tabela 2.4.2: recorte de impressão de tela mostrando duas janelas de trabalho[72].

Na figura acima é possível ver, no parágrafo iniciado por "*My eating habits*", o primeiro erro do córpus, o uso pluralizado do adjetivo *normal*. Esses erros foram levantados introspectivamente, por meio de suspeita enquanto professor de idiomas. O que chamo de suspeita é a percepção de que alguma palavra dentro de uma colocação não se encaixa, tanto sob o ponto de vista lexical quanto sob o gramatical. Essas suspeitas foram levantadas à luz da gramática padrão, uma vez que as redações analisadas foram compostas para finalidade de aprendizagem, isto é, escolar. Por isso, a experiência docente com relação à estrutura de determinados tempos e aspectos gramaticais e ao uso colocacional é que permite identificar usos em tese "suspeitos".

Quando houve uma suspeita de erro o primeiro passo foi verificá-la no COCA para observar seu uso (ou não uso) pelos falantes nativos. Para isso foi necessário depreender o padrão léxico-gramatical desses erros de modo a se verificar se tal suspeita se concretizava ou não. Por padrão léxico-gramatical, como visto na Fundamentação Teórica, refiro-me à associação entre o princípio de escolha livre e o colocacional, ou seja, o uso colocacional e o uso gramatical. Na metodologia

[72] À esquerda a pasta correspondente ao nível básico 1 e as redações respectivas; à direita o programa computacional Wordpad, do Windows, com a redação 00001.txt aberta.

aqui desenvolvida, o padrão léxico-gramatical deve ser depreendido por meio das coligações[73]. No caso do uso de *normals* (cf. figura anterior), a coligação foi: substantivo + verbo + *normals* + *for* + adjetivo. Tem-se, portanto, o nódulo (a palavra que levanta suspeita) e um recorte de duas palavras à esquerda e duas à direita dele (caso haja)[74]. Esse recorte foi feito para permitir que a busca no COCA retornasse resultados mais satisfatórios e próximos àquilo que os aprendizes pretendiam dizer[75]. Note-se que a busca poderia ter sido feita sem a substituição das palavras *habits, are* e *Brazilian*. Porém, o uso dessas palavras restringiria demais o escopo e a possibilidade de a busca retornar vazia aumentaria. Por isso recomendo que sejam mantidas apenas a palavra suspeita e as preposições do modo como aparecem na produção dos aprendizes, e as demais palavras do recorte sejam substituídas por coligações.

Após o levantamento das coligações com relação aos erros encontrados, fui ao sítio do COCA na Internet (www.americancorpus.org) e, no campo de busca, utilizei como termo de busca a coligação observada. Todavia, é importante mencionar que o COCA possui uma sintaxe específica para a busca em seu sistema. O quadro a seguir a resume:

[73] Coligação é "associação entre itens lexicais e gramaticais" (Berber Sardinha, 2004, p. 40).
[74] Na ocorrência "*pasta and fruit's juice*", por exemplo, a palavra suspeita é a marca de posse *fruit's*. O termo de busca usado neste caso poderia ser [n*] [c*] *fruit 's* [n*], composto de duas palavras à esquerda e uma à direita.
[75] O que quero dizer é que, nesta metodologia, a busca deve se assemelhar ao máximo do que o aprendiz pretendia dizer.

Classe gramatical:	**Sintaxe no sítio do COCA:**
Adjetivo	[j*]
Advérbio	[r*]
Artigo	[at*]
Conjunção	[c*]
Determinante	[d*]
Números cardinais	[mc*]
Números ordinais	[md*]
Possessivos	[app*]
Preposições	[i*]
Pronomes	[p*]
Verbo *to be*	[vb*]
Verbos	[v*]

Tabela 2.4.3: resumo de como realizar buscas no sítio do COCA usando coligações[76].

Retomando a coligação com relação ao uso de *normals* (substantivo + verbo + *normals* + *for* + adjetivo) e observando a tabela acima, o termo escrito no campo de busca no sítio do COCA foi: [n*] [vb*] *normals for* [j*], como mostra a figura a seguir:

[76] É importante observar também que, na sintaxe do COCA, apóstrofes são separados de palavras como can't, de modo que se digitaria no campo de busca can 't. Os dados foram retirados do próprio sítio.

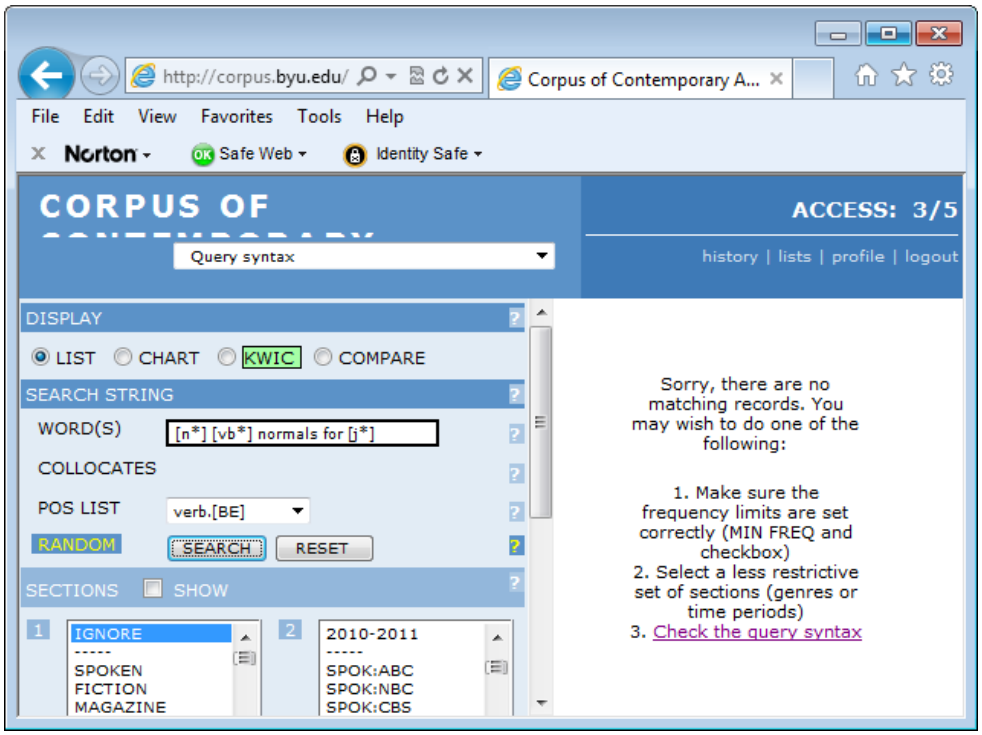


Tabela 2.4.4: recorte de tela mostrando o sítio do COCA na Internet[77].

A figura acima mostra que a busca não retornou resultados para a coligação utilizada, o que indica, dentro desta metodologia, a confirmação de que a suspeita consiste realmente em um uso inadequado, ou seja, um erro. O fluxograma a seguir exemplifica melhor a metodologia de identificação de erros desenvolvida nesta pesquisa:

[77] No campo de busca WORD(S) pode-se ver o termo de busca [n*] [vb*] normals for [j*] e à direita a mensagem indicando que nada foi encontrado no sistema com esse padrão

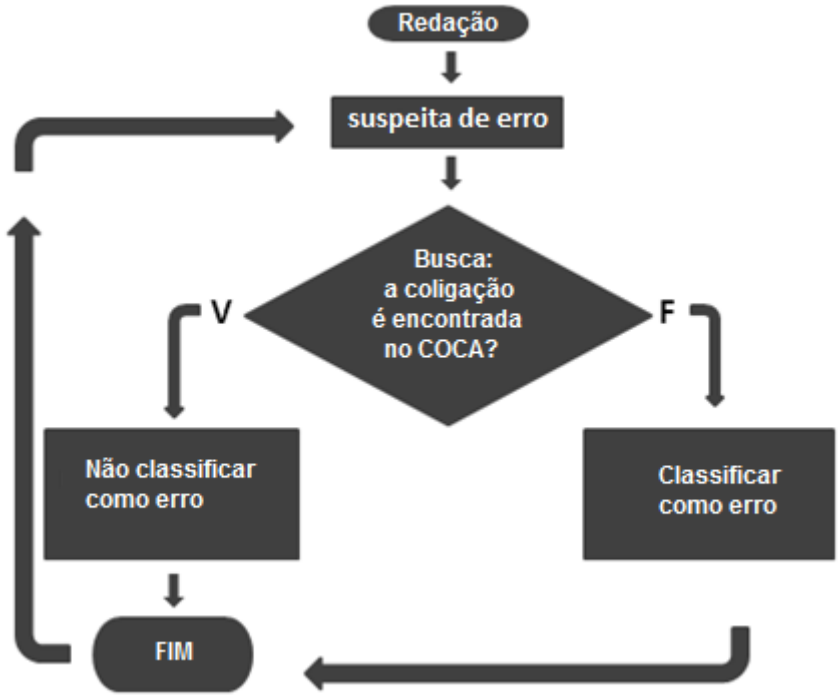


Tabela 2.4.5: fluxograma de decisão sobre erros[78].

Com relação a esse mesmo erro levantei uma nova suspeita: será que a palavra *normal* pode ser usada como sinônima de *common* ou o aprendiz trocou uma pela outra? Para ter essa suspeita solucionada, foi preciso buscar no COCA dois termos de busca, cada um com uma coligação: a) [n*] [vb*] *normal for* [j*]; e b) [n*] [vb*] *common for* [j*].

[78] As letras "V" e "F" em maiúscula correspondem, respectivamente, a "verdadeiro" e "falso" e foram incorporadas a partir da área da lógica da computação. Indicam que se a proposição dentro do losango for verdadeira, acontecerá o desdobramento previsto dentro do retângulo correspondente. A lógica se repete no caso da proposição ser falsa.

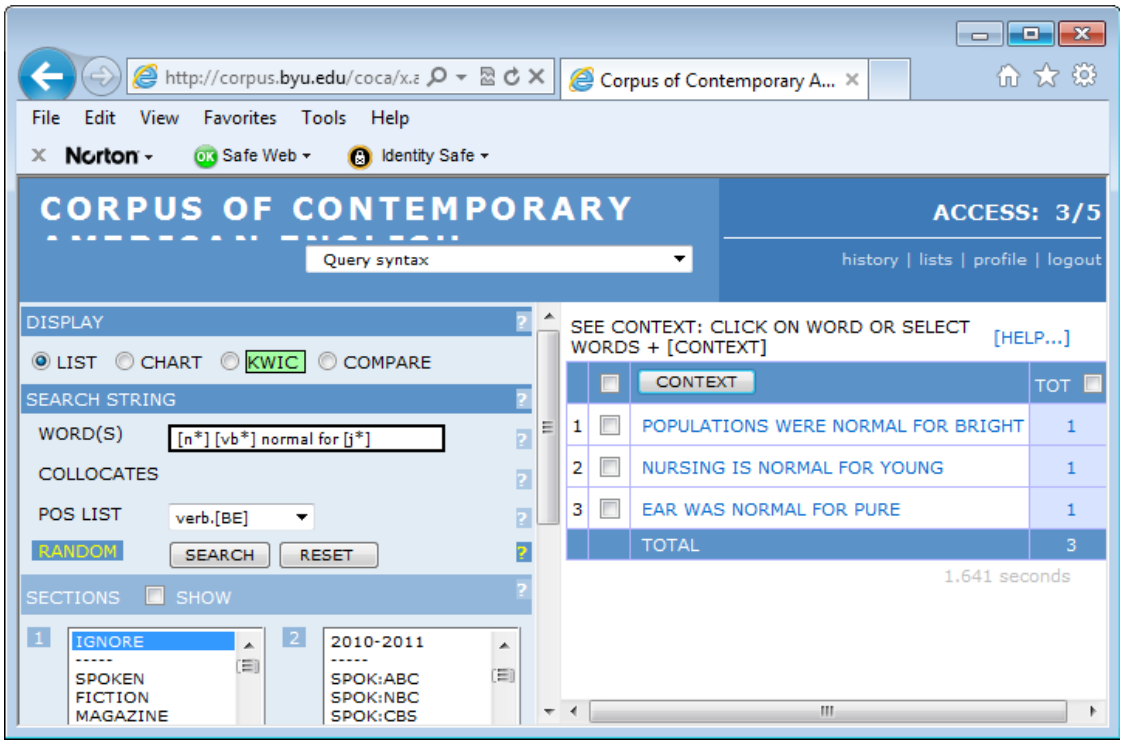


Tabela 2.4.6: impressão de tela do sítio do COCA mostrando os resultados para o termo de busca [n*] [vb*] *normal for* [j*][79].

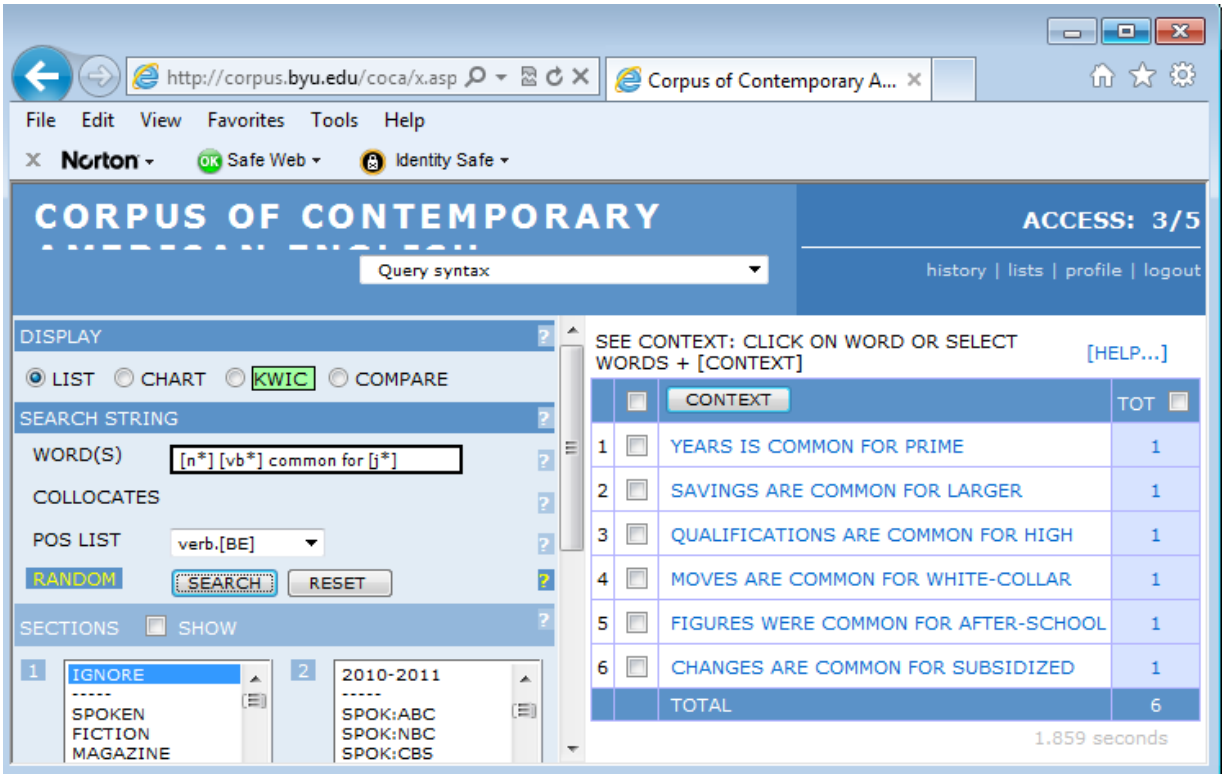


Tabela 2.4.7: impressão de tela do sítio do COCA mostrando resultados para o termo de busca [n*] [vb*] *common for* [j*][80].

[79] Há apenas três ocorrências (parte inferior direita) para essa coligação.
[80] Houve 6 ocorrências (parte inferior direita).

As figuras acima mostram que houve resultados para ambas as coligações. Desse modo, não se pode dizer que o aprendiz tenha cometido um erro com relação à escolha do adjetivo *normal*. Na verdade, o erro consiste na pluralização desse adjetivo, uma vez que, em inglês, os adjetivos não são pluralizados. A suspeita que procurei identificar é: será que as palavras *normal* e *common* podem ser usadas nas mesmas situações e portar o mesmo significado? O resultado acima indica que sim.

Quando há suspeitas que se desdobram em duas opções, como a supracitada, recomendo que sejam buscadas as duas e os resultados sejam comparados. Isso é importante porque como o COCA tem armazenado em seu sistema material que representa a língua em uso, quanto maior o número de ocorrências, maior a escolha pelo padrão buscado. Naturalmente, nos casos expressos pelas duas figuras anterior a diferença entre as ocorrências não é representativa, mas poderia ter sido caso houvesse diferença de uso muito díspar.

O quadro abaixo resume a metodologia de identificação de erros proposta nesta pesquisa:

1. leitura da redação;
2. identificação da suspeita de erro;
3. depreensão da coligação na qual o erro está inserido. Neste passo, manter a própria palavra central do erro (nódulo) e a(s) duas partes do discurso que o acompanha(m) ou antecede(m);
4. busca da coligação no sítio do COCA para verificar se é usada pelos falantes nativos;
5. confirmação ou não de que se trata de um erro.

Quadro 2.4.3: resumo da metodologia de identificação de erros proposta nesta pequisa.

2.4.3.2 Classificação dos erros

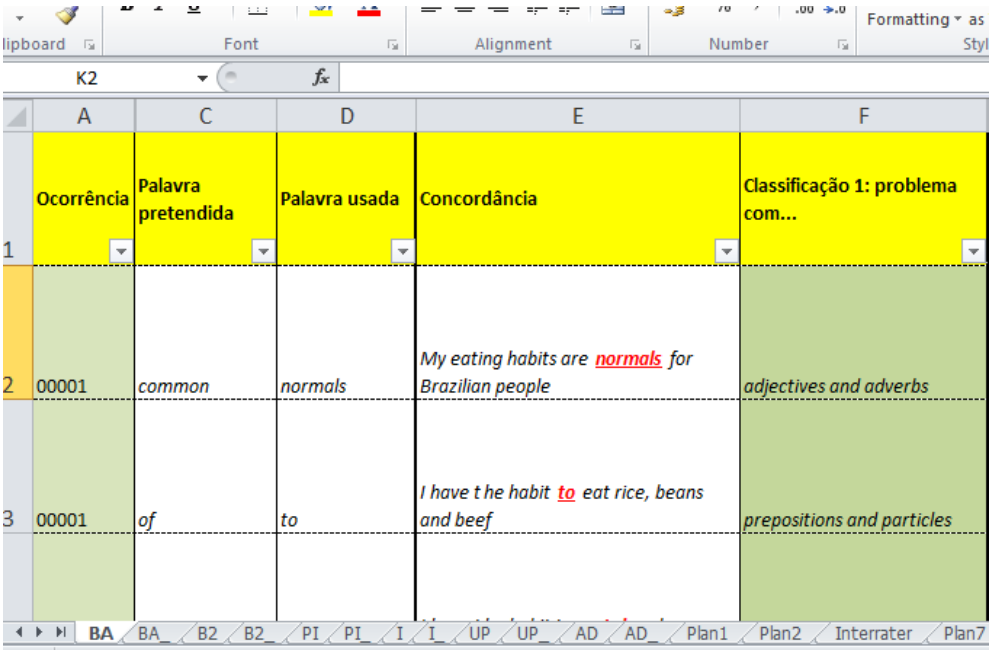
O segundo passo desta metodologia, como visto, é a análise e a classificação dos erros encontrados. Para isso foi preciso fazer inicialmente uma tabulação dos erros em uma planilha do programa computacional *Microsoft Excel 2010*, da suíte do *Microsoft Office 2010*. Tal tabulação foi importante para que pudesse quantificar os erros e responder às questões desta pesquisa (1- Quais

os erros mais comuns no córpus COBRA-7_recorte?; 2- Qual a variação de erro entre os níveis de curso dos aprendizes no córpus COBRA-7_recorte?; 3- Qual nível de curso apresenta maior diversidade de erros no córpus COBRA-7_recorte?). Esse passo foi realizado após o passo anterior[81] para todas as ocorrências de erro encontradas no córpus COBRA-7_recorte.

Após a confirmação do erro no passo anterior da metodologia, tabulei os resultados conforme o procedimento a seguir:

a) copiei, para cada erro encontrado, o nome do arquivo, a palavra usada erroneamente, a palavra pretendida (a partir da minha percepção enquanto professor), e a oração (linha de concordância) que continha o erro;

b) em uma outra coluna da planilha, informei a que correspondia o erro no sistema de classificação de erros baseado em Shepherd (2001), como mostra a figura abaixo:



	A	C	D	E	F
1	Ocorrência	Palavra pretendida	Palavra usada	Concordância	Classificação 1: problema com...
2	00001	common	normals	My eating habits are normals for Brazilian people	adjectives and adverbs
3	00001	of	to	I have t he habit to eat rice, beans and beef	prepositions and particles

Tabela 2.4.8: recorte de tela de planilha do programa computacional *Microsoft Excel 2010* contendo os erros encontrados nas redações[82].

[81] Identificação da suspeita de erros e comprovação da suspeita por meio de busca da coligação correspondente no sítio do COCA.
[82] O título "Classificação 1" se refere ao sistema baseado em Shepherd (2001).

A figura mostra que o erro com o uso de *normals* corresponde, no sistema de classificação baseado na proposta de Shepherd (2001), a um erro no uso de adjetivos ou advérbios (*adjectives and adverbs*[83]).

Na mesma figura, o erro da linha 3 corresponde à escolha incorreta de preposição (*to* ao invés de *of*), o que, no sistema de classificação baseado em Shepherd (2001), corresponde a um erro no uso de preposições ou partículas (*prepositions and particles*).

A seguir mostrarei exemplos de erros encontrados que foram classificados segundo cada um dos itens que compõem sistema de classificação baseado no proposto por esse pesquisador.

A) Identificação do erro segundo o sistema de classificação baseado em Shepherd (2001)

Ao tratar das pesquisas com erros, o sistema de classificação léxico-gramatical proposto por Shepherd (2001)[84] pode ser resumido em 16 categorias, cada uma das quais representa erros no uso de: *adjectives and adverbs*; *articles and countability*; *conditionals*; *determiners*; *modal verbs*; *non-finite forms*; *orthography*; *passive voice*; *prepositions and particles*; *pronouns*; *questions, negatives, auxiliaries*; *relatives*; *there is*; *time, tense, aspect*; *vocabulary*; e *word order* (respectivamente: adjetivos ou advérbios; artigos ou substantivos contáveis/incontáveis; condicionais; determinantes; verbos modais; formas não-finitas; ortografia; voz passiva; conjunções ou preposições; pronomes; questões, negações ou auxiliares; pronomes relativos; verbo there is; tempo ou Aspecto verbal; vocabulário; ordem de palavras). Como visto anteriomente, cada erro encontrado foi buscado no sítio do COCA usando a metodologia de identificação de erros citada no início desta seção e somente quando foi confirmado foi inserido na planilha do *Microsoft Excel 2010* na qual cada erro foi identificado conforme as categorias léxico-gramaticais do sistema proposto por Shepherd (2001).

A base do estudo de Shepherd é a gramatica tradicional, isto é, categorias morfossintáticas como substantivo, preposição, adjetivo, etc. Não há no texto nformação sobre se tais categorias foram baseadas em pesquisa com córpus e, portanto, não há como saber a frequência de suas

[83] Optei por traduzir a conjunção *and* que compõe o termo *adjectives and adverbs* como "ou" por influência da lógica da computação, a qual diz que a conjunção aditiva pressupõe, necessariamente, a coexistência das duas palavras por ela ligadas, enquanto que a conjunção alternativa ("ou") pressupõe, como o nome diz, a alternância entre uma palavra e outra, permitindo, assim, que a presença de uma seja verdadeira enquanto a outra seja falsa, relação impossível no caso anterior. Em outras palavras, uma palavra não pode ser ao mesmo tempo um adjetivo e um advérbio. Em todas as traduções dos itens que compõem o sistema de classificação baseado em Shepherd (2001) essa relação foi mantida.

[84] Utilizo nesta pesquisa somente as classificações de Shepherd (2001) que tratam da léxico-gramática e não de aspectos fonológicos. Tal recorte justificou, como visto, referir-me à aplicação desse sistema de classificação como "baseado em Shepherd (2001)".

ocorrências. Todavia, tais categorias são úteis para a pesquisa com córpus, pois podem ser aplicadas na anotação de córpora de aprendizes, como feito nesta pesquisa.

Apresento a seguir exemplos encontrados no córpus COBRA-7_recorte para cada uma dessas categorias[85].

a) adjetivos ou advérbios (*adjectives and adverbs*): esta classificação foi utilizada quando encontrei no córpus inserção desnecessária ou omissão de um adjetivo ou um advérbio na produção do aprendiz.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00003	00003	*whole*	*integral*	***integral*** *bread and granola with milk.*	*adjectives and adverbs*	(B1)
4	00003g	00003g		*how*	*The scene when yet* ***how*** *a child she was singing La Marseillaise causing emotion.*	*adjectives and adverbs*	(B2)
5	00006i	00006i	*As*	*How*	***How*** *I said, there are restaurants,*	*adjectives and adverbs*	(PI)
6	00070e	00070e	*calmness*	*calm*	*is a little oasis of* ***calm****, accessed by a trail*	*adjectives and adverbs*	(IN)
7	00056d	00056d	*ordinary*	*normal*	*while* ***normal*** *people don't pay that*	*adjectives and adverbs*	(IS)
8	00034c	00034c	*best*	*better*	*Parents will always want the* ***better*** *for you*	adjectives and adverbs	(AV)

Tabela 2.4.9: erros no uso de adjetivos ou advérbios[86].

[85] É importante mencionar que as ocorrências são recortes. Por isso as concordâncias apresentadas podem conter outros erros que não foram indicados nas tabelas. Porém, todos os erros foram classificados, como pode ser confirmado no arquivo do *Microsoft Excel 2010* chamado "Controle_corpus_v3", disponível no CD-rom anexo.

[86] As siglas entre parênteses indicam o nível de curso no qual o erro foi encontrado.

b) artigos e substantivos contáveis e incontáveis (*articles and countability*): esta classificação engloba uso inadequado ou omissão do artigo definido *the* ou pluralização de palavra incontável.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00003b	00003b		*the*	*and if do you like sports, in **the** São Paulo city*	*articles and countability*	(B1)
4	00003d	00003d	*the*		*We are planning a trip **to** U.S.A. next holiday*	*articles and countability*	(B2)
5	000047	000047		*the*	*specifically in **the** donwtown, a UFO was saw*	*articles and countability*	(PI)
6	00070e	00070e		*the*	*because you can only arrived there by **the** ocean or the air*	*articles and countability*	(IN)
7	00057e	00057e	*do*	*the*	*So why **the** scientists continue make animals suffer*	*articles and countability*	(IS)
8	00034c	00034c	*advice*	*advices*	*our parents will always be there given us a lot of **advices***	articles and countability	(AV)

Tabela 2.4.10: erros no uso de artigos.

c) condicionais (*conditionals*): esta categoria engloba erros de lógica na construção das orações com *if* (*if clauses*), como no exemplo a seguir: "*If I went to London, I **visited** the French museums*" (Shepherd, 2001, p. 120, grifo meu).

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00111c	00111c		*would*	*the owl said it would help if he **would** promise to behave from now on,*	*conditionals*	(PI)

Tabela 2.4.11: erros na construção de orações com *if*[87].

d) determinantes (*determiners*): de acordo com Eastwood (2002, p. 3), a classe dos determinantes engloba os artigos, os possessivos, os pronomes demonstrativos e os quantificadores. Porém, Shepherd (2001) parece considerar os erros no uso do artigo definido *the* como pertencentes à categoria dos artigos ou substantivos contáveis e incontáveis, exceto quando esse artigo precede um possessivo, como no caso de "*She wanted to borrow **the** my car*" (p. 124, grifo meu). Nesta pesquisa, considero da mesma forma.

[87] Só houve ocorrência deste item no nível de curso pré-intermediário.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00006	00006		*the*	*I live with **the** my mother*	*determiners*	(B1)
4	00003f	00003f	*his*	*your*	*and dedicated employee and he had a nice family but he spend all **your** free time with charity.*	*determiners*	(B2)
5	000050	000050	*There*	*That*	***That** won't school near my home,*	*determiners*	(PI)
6	00281g	00281g	*many*	*much*	*I have ever studied too **much** things in administration*	*determiners*	(IN)
7	00308	00308	*their*	*its*	*yet preserving **its** own life*	*determiners*	(IS)
8	00191c	00191c	*an*	*a*	*was included as **a** official cocktail by the International Bartender's Association*	determiners	(AV)

Tabela 2.4.12: erros no uso de determinantes.

e) verbos modais (*modal verbs*): para Greenbaum (1996, p. 80), os *modal verbs* são *can*, *could*, *may*, *might*, *shall*, *should*, *will*, *would*, *must*, *ought to*. Shepherd (2001) considera um erro no uso de modais quando são omitidos ou usados inadequadamente.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00006b	00006b	*will*		*I **go to write** about my weekend*	*modal verbs*	(B1)
4	00018d	00018d	*should*		*Christy Brown and when was adult, he study art. Is a beautiful movie who **be** see*	*modal verbs*	(B2)
5	00063b	00063b	*would*	*will*	*On instead there **won**'t be so many technologies or facilities nowadays.*	*modal verbs*	(PI)
6	00281f	00281f		*will*	*but after I´ll realized it´s not a reality*	*modal verbs*	(IN)
7	00358	00358	*can*	*could*	*I **could**n´t do that*	*modal verbs*	(IS)
8	00439	00439	*had*	*would*	*when things don't happen how they **would** like*	modal verbs	(AV)

Tabela 2.4.13: erros no uso de verbos modais.

f) formas não finitas (*non-finite forms*): em sua categorização, Shepherd (2001) prevê três casos para um erro ser considerado do tipo "formas não-finitas":

I. uso de verbo no infinitivo ao invés do *gerund* (como em "*I'm tired to listen to her complaints*", p. 121);

II. uso de *bare infinitive* em combinação de dois verbos, como no caso de "*I tried telephone you*" (p. 121);

III. montagem inadequada da construção "objeto + infinitivo", como em "*She wants that you phone her*" (p. 121).

Porém, como tais casos conflitam com outros critérios já estabelecidos pelo autor, como omissão de preposição (caso II acima) e uso inadequado de pronome relativo (caso III

acima), optei por considerar, nesta pesquisa, apenas o caso I supracitado como erro no uso de formas não-finitas.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00001	00001	*eating rice*	*eat rice*	*I have the habit to **eat rice**, beans and beef*	*non-finite forms*	(B1)
4	00018c	00018c	*reading*	*read*	*Thank you very much for **read** this letter.*	*non-finite forms*	(B2)
5	00006i	00006i	*being*	*be*	*to **be** or become a guest to Paulista Avenue*	*non-finite forms*	(PI)
6	00079	00079	*running*	*run*	*Stop to **run** against time;*	*non-finite forms*	(IN)
7	00057c	00057c	*studying*	*study*	*I agree with you that **study** is very important but*	*non-finite forms*	(IS)
8	00091b	00091b	*help*	*helping*	*for example, to collect and to recycle the garbage and **helping** elderly people*	non-finite forms	(AV)

Tabela 2.4.14: erros no uso de formas não-finitas.

g) ortografia (*orthography*): esta categoria engloba palavras escritas ou pontuadas de modo inadequado.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3		00006	*has*	*hás*	*My grandfather, Nino, and my grandmother, Ivone,* ***hás*** *four children*	*orthography*	(B1)
4	00003f	00003f	*especially*	*specially*	*After time, he wanted to help this people* ***specially*** *the children who lived out of home.*	*orthography*	(B2)
5	00006i	00006i	*Aclimação*	*Acclimation*	*I live in the neighborhood of* ***Acclimation,***	*orthography*	(PI)
6	00070e	00070e	*like*	*liken*	*restaurants, activities in the beach* ***liken*** *a banana boat or diving.*	*orthography*	(IN)
7	00055d	00055d	*remember*	*remeber*	*Nobody Will* ***remeber*** *that bad look*	*orthography*	(IS)
8	00034c	00034c	*babies*	*babys*	*Since* ***babys*** *we are surrounded by people who live with us*	orthography	(AV)

Tabela 2.4.15: erros com a ortografia de palavras.

h) voz passiva (*passive voice*): esta categoria engloba construção inadequada de uma voz passiva.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00006	00006	*divorced*	*are divorced*	*My parents **are divorced** some years ago*	*passive voice*	(B1)
4	00011c	00011c	*was*		*romantic comedy, **it realized** in 2006.*	*passive voice*	(B2)
5	000047	000047	*seen*	*saw*	*specifically in the donwtown, a UFO was **saw***	*passive voice*	(PI)
6	00281g	00281g	*rejected*	*reject*	*I wasn´t **reject** to the others students*	*passive voice*	(IN)
7	00437	00437	*is*		*After this bad habit **installed***	passive voice	(AV)

Tabela 2.4.16: erros na construção da voz passiva[88].

i) conjunções ou preposições (*prepositions and particles*): esta classificação foi utilizada quando encontrei inserção desnecessária ou omissão de uma conjunção ou uma preposição na produção do aprendiz.

[88] Não houve ocorrência deste tipo de erro no nível intermediário superior.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00001	00001	*of*	*to*	*I have the habit **to** eat rice, beans and beef*	*prepositions and particles*	(B1)
4	00011	00011	*then*	*than*	*ts not a good idea to drive, **than** you shouldn´t rent a car.*	*prepositions and particles*	(B2)
5	00006i	00006i	*of*		*I don`t need to go out **the** neighborhood,*	*prepositions and particles*	(PI)
6	00070e	00070e	*whereas*	*while*	*The Third Beach is plenty of accommodation options, **while** the*	*prepositions and particles*	(IN)
7	00055e	00055e	*at*	*to*	*When I arrived home and looked **to** the shoe*	*prepositions and particles*	(IS)
8	00034b	00034b	*too*	*to*	*and I am sure that mom dont likes **to**.*	prepositions and particles	(AV)

Tabela 2.4.17: erros no uso de preposições ou conjunções.

j) pronomes (*pronouns*): esta categoria engloba omissão ou uso inadequado de pronomes pessoais, reflexivos ou pronomes objeto (*me, him, us* etc.).

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00001	00001	*it*		*but **is expensive***	*pronouns*	(B1)
4	00006g	00006g	*her*	*it*	*then the superior nun indicates **it** to be governanta of the house of a captain*	*pronouns*	(B2)
5	00003i	00003i	*it*		*I am sure that you will **like.***	*pronouns*	(PI)
6	00079	00079	*themselves*	*yourself*	*Reliable in **yourself**;*	*pronouns*	(IN)
7	00055b	00055b	*myself*	*me*	*It's like this, I imagine **me** performing that songs on the stage*	*pronouns*	(IS)
8	00028c	00028c	*her*	*she*	*I could make **she** laugh, I felt very well*	pronouns	(AV)

Tabela 2.4.18: erros no uso de pronomes.

k) questões, negações ou auxiliares (*questions, negatives, auxiliaries*): esta categoria compreende omissão ou inserção inadequada de verbo auxiliar. Para Greenbaum (1996, p. 153) e Eastwood (2002, p. 104) os verbos auxiliares são: *be*, *have*, e *do*. Essa classificação considera *did* o passado de *do*. Portanto, aquele estaria também dentro desta categoria (Eastwood, 2002, p. 17). Dessa forma, nesta pesquisa, considerei auxiliares os seguintes verbos: *be*, *have*, *do*, e *did*. Naturalmente, *have* somente foi considerado verbo auxiliar quando esteve inserido nos tempos verbais perfeitos, como o *present perfect* e o *present perfect continuous*. Este item engloba ainda o uso incorreto dos advérbios de negação *no* ou *not*, e erro de lógica no uso de *tag questions*[89]. A tabela a seguir ilustra melhor este item:

[89] Não houve no COBRA-7_recorte erros no uso de *tag questions*.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00003b	00003b		*do*	*How **do** you know, here is summer.*	*questions, negatives, auxiliaries*	(B1)
4	00006e	00006e	*is*	*does*	*because in the start of the night **does** a little of cold.*	*questions, negatives, auxiliaries*	(B2)
5	00070b	*00070b*	*not*	*no*	*your opnion about the place, if you like or **no**.*	*questions, negatives, auxiliaries*	(PI)
6	00078	00078		*Did*	***Did**n´t exist vaccine for many disease like variola*	*questions, negatives, auxiliaries*	(IN)
7	00057e	00057e	*are*	*is*	*If this tests **is** so important why they don't do that in themselves*	*questions, negatives, auxiliaries*	(IS)
8	00087d	00087d		*are*	*but although this **are** they are wonderful and you have to go*	questions, negatives, auxiliaries	(AV)

Tabela 2.4.19: erros no uso de questões, negações ou auxiliares.

l) pronomes relativos (*relatives*): esta categoria engloba omissão ou uso inadequado de pronomes relativos. Greenbaum (1996, p. 85) afirma que esta categoria é composta pelas palavras *who*, *which* *whom*, *that* e *whose*. Eastwood (2002, p. 233) não lista essa última como item pertencente a esta categoria. Neste trabalho optei, neste item, pela definição de Greenbaum (1996) por ser mais abrangente.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00054b	00054b	*what*	*that*	*I love the most part **that** we eat, foods are very hot.*	*relatives*	(B1)
4	00009b	00009b	*who*		*Throughout my live I lived with **people in** some way contributed*	*relatives*	(B2)
5	00251c	00251c	*which*	*that*	*there are many pasta restaurants, **that** is typically Italian too.*	*relatives*	(PI)
6	00078	00078	*which*	*what*	*So they become pregnancy later, **what** improve the risk of life.*	*relatives*	(IN)
7	00194b	00194b	*of*	*that*	*that he beaten in her face*	*relatives*	(IS)
8	00437c	00437c	*so*	*that*	*and had hard times to deal with, even **that** I have to recognize*	*relatives*	(AV)

Tabela 2.4.20: erros no uso de pronomes relativos.

m) *there is*: esta categoria de Shepherd (2001) compreende: a substituição de *is* por *are* (ou vice-versa), no presente ou no passado, nas conjugações do verbo *there be*; o uso dos verbos *have* ou *exist* com o significado de "existir". A tabela adiante ilustra melhor esta categoria:

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00004	00004	*there are*	*has*	*during the Carnival too **has** many dances of the Carnival*	*there is*	(B1)
4	00017b	00017b	*there are*	*there is*	*Be careful in the airports, there **is** many cases of assaults.*	*there is*	(B2)
5	00050b	00050b	*there are*	*has*	*because normally **has** thieves in the true*	*there is*	(PI)
6	00078	00078	*There was*	*exist*	*Didn´t **exist** vaccine for many disease like variola*	*there is*	(IN)
7							(IS)
8	00499c	00499c	*there is*	*has*	*Brazilian presence is so strong that in New York **has** a street where the Carnival*	*there is*	(AV)

Tabela 2.4.21 erros no uso de *there be*.

n) tempo ou aspecto verbal (*time, tense, aspect*[90]): Segundo Greenbaum (1996), "Tempo é uma categoria gramatical que se refere à localização de uma situação no tempo[91]" (p. 253), enquanto que "aspecto" se refere "à forma como o tempo da situação é considerado[92]". Eastwood (2002), por sua vez, apresenta uma classificação mais clara. Para ele, o tempo verbal pode ser identificado fazendo-se a seguinte pergunta: "Passado ou presente?[93]" (p. 77) – enquanto que o aspecto pode ser descoberto por meio das questões a seguir: "Perfeito ou não", "Contínuo ou não?[94]" (p. 77). Portanto, esta categoria engloba erros ocorridos na concordância verbal das orações produzidas pelos aprendizes do córpus COBRA-7_recorte, isto é, ao uso do tempo inadequado (presente no lugar de passado, passado no lugar de

[90] Shepherd (2001) utiliza nesta categoria o nome *time, tense aspect*. Porém, enquanto que o termo *tense* nomeia uma classificação gramatical referente ao comportamento verbal, o termo *time* se refere à nossa percepção real do que é presente, passado e futuro (cf. Michaelis, 2006) e, por isso, não tem relevância nesta análise.
[91] "*Tense is a grammatical category referring to the location of a situation in time*" (p. 253, tradução minha).
[92] "*...to the way thay the time of the situation is regarded*…" (p. 253, tradução minha).
[93] "*Past or present?*" (p. 77, tradução minha).
[94] "*Perfect or not?*", "*Continuous or not?*", respectivamente (p. 77, tradução minha).

futuro etc.) e/ou aspecto (contínuo ao invés de simples, por exemplo). Por não haver uma categoria específica, englobei nesta categoria também os erros de conjugação como, por exemplo, a omissão ou o acréscimo desnecessário o *–s* que marca de terceira pessoa do singular no tempo presente. Nesta categoria não considero as construções do tipo *gerund*, pois, como vimos, possuem uma categoria específica segundo os critérios de classificação baseados em Shepherd (2001).

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00001	00001	*love*	*loved*	*I **loved** Japanese food,*	*time, tense, aspect*	(B1)
4	00003e	00003e	*park*	*parking*	*There are a lot of traffic and a few places to **parking***	*time, tense, aspect*	(B2)
5	00003i	00003i	*play*	*plays*	*People **plays,** goes to gym, and walks with his pets.*	*time, tense, aspect*	(PI)
6	00070e	00070e	*called*	*call*	*The smallest beach, **call** First Beach,*	*time, tense, aspect*	(IN)
7	00056	00056	*shows*	*show*	*when something bad **show** up in our front*	*time, tense, aspect*	(IS)
8	00028	00028	*had*	*have*	*so I started doing things that I **have** never done*	*time, tense, aspect*	(AV)

Tabela 2.4.22: erros no uso de tempo ou aspecto verbal.

o) vocabulário (léxico) (*vocabulary*): esta categoria traz erros de escolha lexical, isto é, quando os aprendizes substituíram uma palavra por uma outra que pertença à mesma ou a outra classe gramatical e tal palavra não se encaixe nas categorias anteriores.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00002c	00002c	*make*	*do*	*I like to **do** new friends*	*Vocabulary*	(B1)
4	00003d	00003d	*talk*	*conversation*	*and would like to know people to **conversation** in English.*	*Vocabulary*	(B2)
5	00003h	00003h	*stuck*	*arrested*	*A man **arrested** in the chimney. Will he be Santa Claus?*	*Vocabulary*	(PI)
6	00070e	00070e	*take*	*pick*	*When you landing, you have to **pick** up a "taxi",*	*Vocabulary*	(IN)
7	00056	00056	*optimistic*	*optimism*	*Sometimes, being so much **optimism** is annoying*	*Vocabulary*	(IS)
8	00034b	00034b	*collapsed*	*gone*	*In that moment my world has **gone**, I was devastated*	*Vocabulary*	(AV)

Tabela 2.4.23: erros de má escolha lexical.

p) ordem de palavras (*word order*): esta categoria engloba colocações nas quais os termos foram invertidos por razões diversas.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Baseado em Shepherd (2001)**	
3	00004	00004	*most popular celebrations*	*celebration more popular*	*The carnival is one of* ***celebration more popular*** *here in Brazil,*	*word order*	(B1)
4	00003e	00003e	*a lot of cabs*	*cap a lot*	*but the public transportation is very good and there are subway, bus and cap* ***a lot.***	*word order*	(B2)
5	00006h	00006h	*horrible smells*	*smells horrible*	*and produced* ***smells horrible****,*	*word order*	(PI)
6		00078b	*has just*	*just had*	*she* ***just had*** *married and so I'm very happy.*	*word order*	(IN)
7	00055e	00055e	*could I*	*I could*	*How* ***I could*** *buy that?*	*word order*	(IS)
8	00437b	00437b	*is it*	*it is*	*what* ***it is*** *?*	*word order*	(AV)

Tabela 2.4.24: erros relativos inversão na ordem de palavras dentro de uma colocação.

No decorrer da análise constatei que o sistema de classificação baseado em Shepherd (2001) revelou muitas categorias, complexidade de distinção entre elas, e não contemplava algumas ocorrências encontradas no córpus COBRA-7_recorte, como uso inadequado do *genitive case* e regularização de passado irregular. Além disso, como nem todos os cursos de inglês focam em gramática e minha experiência docente mostra que muitos aprendizes não estão familiarizados com essas categorias mesmo em português, optei por criar um sistema mais enxuto e confiavel, isto é, um sistema de classificação que contemplasse todas as ocorrências e fosse o mais simples possível, de modo que pudesse ser usado pelos professores de inglês e entendido pelos aprendizes de forma prática. A seguir falo desse sistema.

B) Identificação do erro segundo o sistema de classificação SO2I, desenvolvido nesta pesquisa

Como mencionado anteriormente, foi necessário o uso de sistemas de classificação para que se pudessem tabular os dados e responder às perguntas desta pesquisa. Como o critério baseado em

Shepherd (2001) permitia que um mesmo item fosse classificado em mais de uma categoria e não dava conta de erros com a omissão ou uso inadequado de caso genitivo (*genitive case*) (‘s), plural, palavra escrita em português, omissão de palavra, e escolha inadequada das palavras que compõem uma colocação, foi necessário o desenvolvimento de um sistema de classificação que não apenas contemplasse as categorias propostas por Shepherd (2001), mas também as supracitadas, resumindo-as, e que, desse modo, fosse de fácil identificação e entendimento para professores e pesquisadores.

Tal observação veio à tona na qualificação, quando os membros da banca apontaram que os erros de aprendizes citados nesta pesquisa poderiam-se resumir em três: inserção, troca e omissão. Esse sistema se baseia na noção de colocação, isto é, na sequência típica de palavras empregada na linguagem.

A observação dos erros transcritos na planilha do programa computacional *Microsoft Excel 2010* (tabulação dos resultados, como mencionado anteriormente) confirmou tal pressuposto, mas mostrou que, em todos os casos, os aprendizes fizeram não uma dentre três, mas uma dentre quatro operações a seguir:

a) substituição (troca);

b) inserção;

c) omissão;

d) inversão.

A seguir mostrarei exemplos de como esse critério foi aplicado aos erros encontrados no córpus COBRA-7_recorte.

a) substituição: o aprendiz substituiu uma palavra por outra, seja ela da mesma ou de outra classe gramatical.

1	A	B	C	D	E	F

2	Ocorrência	Redação	Palavra pretendida	Palavra usada	Concordância	Classifificação desta pesquisa	
3	00001	00001	*of*	*to*	*I have the habit **to** eat rice, beans and beef*	substituição	(B1)
4	00003e	00003e	*on*	*by*	*You should search them **by** the net. You can found very good prices.*	substituição	(B2)
5	00003h	00003h	*stuck*	*arrested*	*A man **arrested** in the chimney. Will he be Santa Claus?*	substituição	(PI)
6	00070e	00070e	*will*	*Could*	***Could** you be enchanted in the beginning of the travel*	substituição	(IN)
7	00055c	00055c	*then*	*than*	*all the pressure of people about me, **than** I ran out to my favorite place*	substituição	(IS)
8	00028c	00028c	*her*	*she*	*I could make **she** laugh, I felt very well*	substituição	(AV)

Tabela 2.4.25: erros de substituição de palavras no córpus COBRA-7_recorte[95].

A tabela mostra que as palavras pretendidos (coluna C) foram substituídas pelas utilizadas pelos aprendizes (coluna D), o que gerou o erro.

b) inserção: o aprendiz inseriu uma palavra, um prefixo ou um sufixo inadequadamente;

[95] As siglas entre parênteses indicam o nível de curso no qual o erro foi encontrado.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Classifificação desta pesquisa**	
3	00001	00001	*common*	*normals*	*My eating habits are **normals** for Brazilian people*	inserção	(B1)
4	00003e	00003e	*people*	*peoples*	*walk in the streets and sit at the cafés to relax and to look at the **peoples.***	inserção	(B2)
5	00006i	00006i		*a*	*because there are **a** haircuts, English school, supermarkets,*	inserção	(PI)
6	00070e	00070e	*land*	*landing*	*When you **landing**, you have to pick up a "taxi",*	inserção	(IN)
7	00056b	00056b	*respect*	*respects*	*I like people who make me laugth, who **respects** me*	inserção	(IS)
8	00034b	00034b	*like*	*likes*	*and I am sure that mom dont **likes** to.*	inserção	(AV)

Tabela 2.4.26: erros de inserção de palavras no córpus COBRA-7_recorte.

A tabela mostra, nas linhas 3 e 4, coluna E, a inserção de marca de plural, respectivamente, nas palavras *normal* e *people*. Na linha 5 observa-se a inserção desnecessária do artigo indefinido *a*[96]. A linha 6, coluna E, vê-se a inserção do sufixo *–ing*, e nas linhas 7 e 8, coluna E, a inserção da marca de terceira pessoa dos verbos no presente, o sufixo *–s*.

c) omissão: o aprendiz omitiu uma palavra, prefixo ou sufixo;

[96] Note-se que, conforme justificado anteriormente, esta linha de concordância traz outros erros que, embora não tenham sido contemplados no recorte da tabela, foram devidamente considerados analisados, como pode ser observado no arquivo "Controle_corpus_v3", disponível no CD-rom anexo à pesquisa.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Classificação desta pesquisa**	
3	00002	00002	*a*		*My husband **is doctor** and his name is Flávio.*	omissão	(B1)
4	00006d	00006d	*news*	*new*	*I like of your **new,** you can visit to me in Brazil!*	omissão	(B2)
5	00006i	00006i	*of*		*I don`t need to go **out the** neighborhood,*	omissão	(PI)
6	00070e	00070e	*an*		*If you decide **for airplane**, you will spend only 20 minutes.*	omissão	(IN)
7	00057	00057	*to*		*because sometime the person **need be** realist*	omissão	(IS)
8	00087c	00087c	*being*	*be*	*For me to **be** successful is to balance the demands*	omissão	(AV)

Tabela 2.4.27: erros de omissão de palavras no córpus COBRA-7_recorte.

A tabela mostra, nas linha 3 e 6, coluna E, a omissão de artigo, como visto na coluna C. A linha 4 mostra a omissão da marca de plural *–s* da palavra *news*, sempre plural enquanto substantivo. As linhas 5 e 7 mostram omissão das preposições *of* e *to*. Por fim, a linha 8, coluna E, mostra a omissão do sufixo *–ing*[97].

d) inversão: o aprendiz inverteu a ordem de palavras que compunham uma colocação[98].

[97] Embora o recorte da tabela não mostre, uma instância de anotação posterior prevê a inserção desnecessária da preposição *to* em "*For me **to** be successful*" (grifo meu).

[98] Nesta pesquisa entendo por colocação qualquer combinação de duas ou mais palavras que formam um unidade lexical composta.

1	A	B	C	D	E	F	
2	**Ocorrência**	**Redação**	**Palavra pretendida**	**Palavra usada**	**Concordância**	**Classificação desta pesquisa**	
3	00004	00004	*most popular celebrations*	*celebration more popular*	*The carnival is one of **celebration more popular** here in Brazil,*	inversão	(B1)
4	00006d	00006d	*me soon*	*soon me*	*You will find **soon me**.*	inversão	(B2)
5	00006h	00006h	*well dressed gentleman*	*gentleman well dress*	*I see one **gentleman well dress**,*	inversão	(PI)
6	00357b	00357b	*the statue also*	*also the statue*	*protest about their right to see **also the statue** for free.*	inversão	(IN)
7	00194b	00194b	*accused person*	*person accused*	*and decide to send to the president before talk with the **person accused***	inversão	(IS)
8	00503d	00503d	*pedophile priest*	*priest pedophile*	*The Church hid **priest pedophile** and spent much money in process*	inversão	(AV)

Tabela 2.4.28: erros de inversão de palavras de colocação no córpus COBRA-7_recorte.

A linha 3, coluna E, mostra a inversão de ordem dos termos que compõem a colocação *well dressed man*. Inversões semelhantes ocorrem nas linhas 4 a 8. A coluna C mostra a colocação pretendida pelos aprendizes.

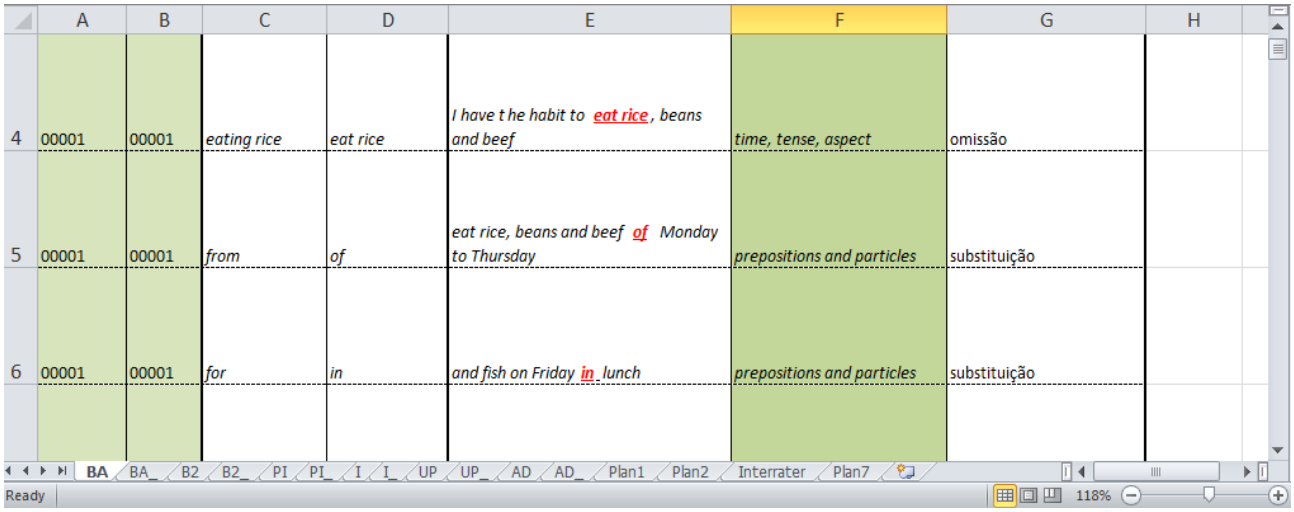
Para fins de referência, empregarei doravante a sigla SO2I (leiam-se as letras como no alfabeto) – sigla que corresponde a substituição, omissão, inserção, inversão – para designar o procedimento de anotação de erros com base colocacional usado nesta pesquisa.

2.4.3.3 Cálculo dos resultados gerados pelos dois critérios

Para que a tabulação dos dados fosse realizada, foram usados alguns critérios. O primeiro deles foi o uso do comando *=CONT.SE(Planilha!Coluna_de_destino:coluna_de_destino;coluna_e_linha_de_origem)* do

próprio programa computacional supracitado (*Microsoft Excel 2010*) para cada critério de análise utilizado. Esse comando permite ao sistema contar quantas vezes uma expressão apareceu. Por isso, foi necessário criar, dentro do mesmo arquivo de computador usado para listar os erros encontrados, planilhas paralelas a cada nível de curso que havia sido analisado.

As próximas figuras ilustram melhor esse procedimento:



	A	B	C	D	E	F	G	H
4	00001	00001	*eating rice*	*eat rice*	*I have t he habit to eat rice , beans and beef*	*time, tense, aspect*	omissão	
5	00001	00001	*from*	*of*	*eat rice, beans and beef of Monday to Thursday*	*prepositions and particles*	substituição	
6	00001	00001	*for*	*in*	*and fish on Friday in lunch*	*prepositions and particles*	substituição	

Tabela 2.4.29: recorte de tela do programa computacional *Microsoft Excel 2010*[99].

Na figura acima, há diversas planilhas, nomeadas segundo o padrão a seguir:

a) siglas que se referem aos níveis de curso analisados nesta pesquisa, a saber: básico 1 (BA), básico 2 (B2), pré-intermediário (PI), intermediário (I), intermediário superior (UP) e avançado (AD). Essas planilhas trazem as ocorrências, como mostra a planilha em destaque na figura acima;

b) as siglas supracitadas seguidas de um traço sobrescrito. Exemplo: BA_. Essas planilhas, doravante chamadas "planilhas de cálculo", trazem a totalização dos resultados encontrados obtida por meio do comando *=CONT.SE()* para cada um dos itens dos sistemas de classificação analisados (baseado em Shepherd, 2001 e SO2I, desenvolvido nesta pesquisa), conforme ilustra a planilha em destaque na figura a seguir:

[99] Em destaque a planilha BA, referente ao nível de curso básico 1. Observe-se ao lado a planilha BA_, que contém a transformação das ocorrências da planilha BA em números.

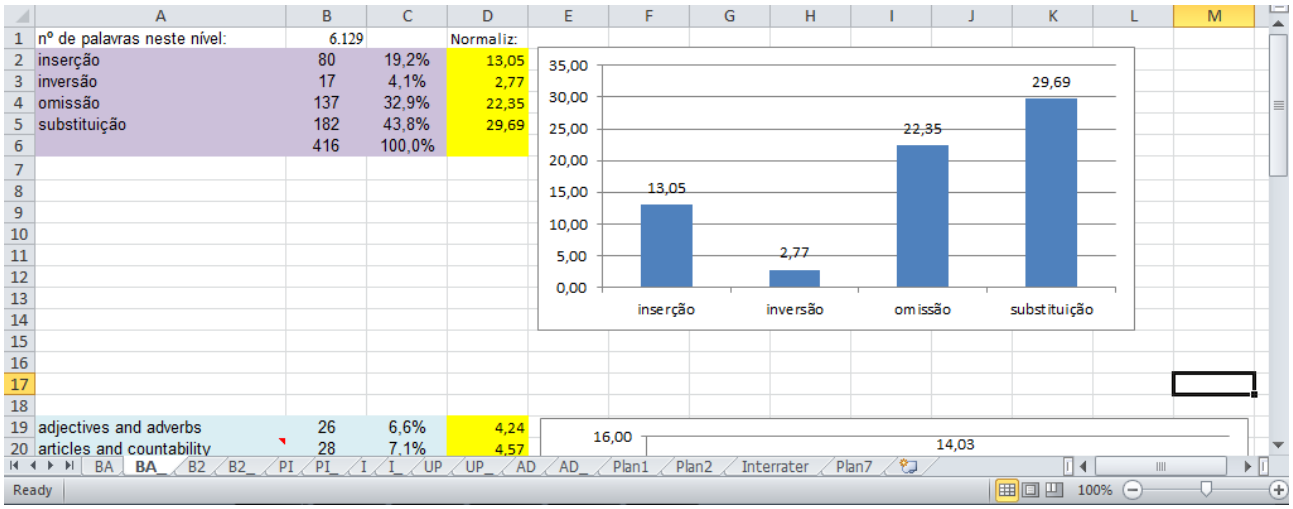
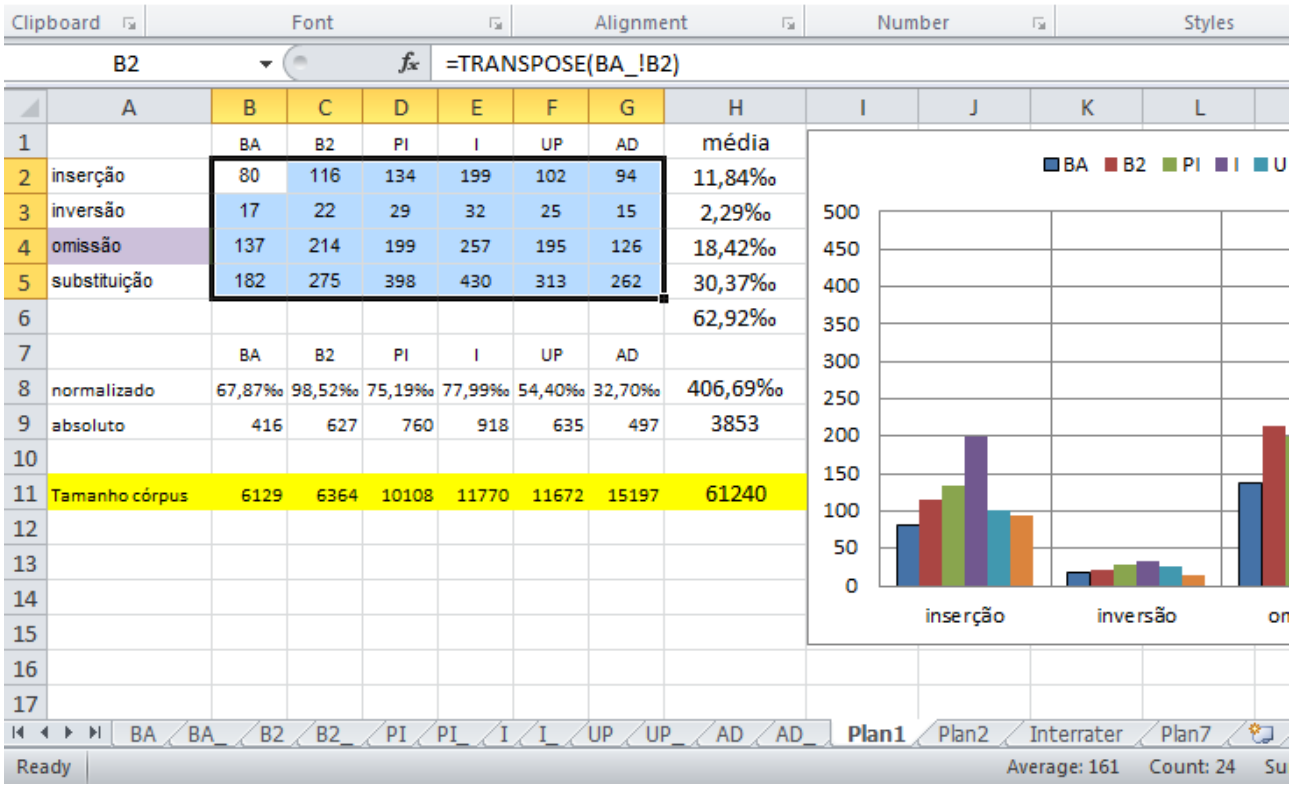


Tabela 2.4.30: recorte de tela do programa computacional *Microsoft Excel 2010*[100].

Como mostra a figura acima (na coluna B acima, a partir da linha 2), o comando *=CONT.SE()* procura, na planilha anterior (neste caso a BA) por ocorrências que tenham por nome as palavras da coluna A (a partir da linha 2 acima) e retorna (na coluna B) quantas vezes a palavra aparece. No caso de "inserção", vê-se que aparece 80 vezes, o que significa que foram encontrados, nesse nível, 80 ocorrências referentes a ele. Isso corresponde a 19,2% dos erros encontrados[101] (coluna C, linha 2). Na coluna D, observa-se a normalização[102] do resultado por mil: (37 / 6129) * 1000 (6129 é o número de palavras encontradas neste nível[103]). Isso indica que, no córpus COBRA-7_recorte, um erro de inserção aparece 13,05 vezes a cada mil palavras.

A seguir, foi necessário coletar os resultados de cada planilha de cálculo cujo nome é seguido de um traço sobrescrito (Exemplo: BA_) e totalizar as ocorrências. Isso foi feito criando-se uma nova planilha (Plan1) e inserindo-se nela colunas correspondentes aos erros encontrados por meio dos dois sistemas classificatórios e aos níveis de curso analisados. Os dados dessas colunas foram transpostos das planilhas de cálculo de cada nível de curso. Para isso, foi usado o comando *=TRANSPOR()* para retornar o número bruto de ocorrências de erro encontrado em cada um dos níveis de curso. A próxima figura mostra esse procedimento:

[100] Em destaque a planilha BA_, referente ao nível de curso básico 1, e o critério de análise desenvolvido nesta pesquisa. À direita está um gráfico demonstrativo gerado a partir dos resultados do nível.
[101] De um total de 416 erros (ver coluna B, linha 6).
[102] A normalização torna os resultados dos níveis de cursos estatisticamente comparáveis, independentemente da diferença entre os números de palavras e de erros.
[103] Como visto anteriormente, esse número foi conseguido utilizando-se o programa computacional Concord, da suíte do Wordsmith Tools 5.0 (Scott, 2008).



	A	B	C	D	E	F	G	H
1		BA	B2	PI	I	UP	AD	média
2	inserção	80	116	134	199	102	94	11,84‰
3	inversão	17	22	29	32	25	15	2,29‰
4	omissão	137	214	199	257	195	126	18,42‰
5	substituição	182	275	398	430	313	262	30,37‰
6								62,92‰
7		BA	B2	PI	I	UP	AD	
8	normalizado	67,87‰	98,52‰	75,19‰	77,99‰	54,40‰	32,70‰	406,69‰
9	absoluto	416	627	760	918	635	497	3853
10								
11	Tamanho córpus	6129	6364	10108	11770	11672	15197	61240

Tabela 2.4.31: recorte de tela do programa computacional *Microsoft Excel 2010*[104].

A partir dos dados transpostos foi possível gerar gráficos diversos que compreendiam ocorrências em todos os níveis de curso analisados, considerando-se tanto os critérios de análise baseados na proposta de Shepherd (2001) quanto os propostos nesta pesquisa.

2.4.3.4 Concordância entre avaliadores

Para verificar até que ponto o critério de classificação desenvolvido nesta pesquisa seria confirmado por uma outra pessoa, convidei uma pesquisadora do GELC (Grupo de Estudos de Linguística de Córpus) para refazer a classificação de 100 ocorrências do córpus COBRA-7. Tal procedimento, conhecido como cálculo da *concordância entre avaliadores*, tem o objetivo de validar a anotação manual de um córpus.

Para isso, selecionei a partir das planilhas do programa computacional *Microsoft Excel 2010* nas quais registrei os erros encontrados (arquivo "Controle_corpus_3", disponível no CD-rom anexo), aleatoriamente, 100 erros encontrados em todos os níveis de curso pesquisados (básico 1,

[104] Em destaque a planilha Plan1, que compila os dados das planilhas de cálculo (nomeadas segundo o padrão XX_). Em destaque (dentro da seleção) os números transpostos da planilha BA_ por meio do uso da função =TRANSPOR(BA_!D2), na qual BA_ se refere à planilha de origem e D2 à coluna e à linha que continham o número de ocorrências do erro.

básico 2, pré-intermediário, intermediário, intermedário superior e avançado). Nessa seleção mantive a configuração original: nome do arquivo, palavra pretendida, palavra usada, concordância, classificação 1 e classificação 2. Posteriormente, colei todas as 100 ocorrências em uma planilha de um novo arquivo do programa computacional supracitado. Não intencionalmente, as 100 ocorrências supracitadas não continham todos os itens prescritos nos critérios de avaliação desta pesquisa. Não havia nas linhas erros no uso de *conditionals*, *modal verbs, passive voice*, e *relatives* (Condicionais, verbos modais, voz passiva e pronomes relativos, respectivamente). Havia, portanto, erros pertencentes às seguintes categorias léxico-gramaticais baseadas na proposta de Shepherd (2001):

1. *adjectives and adverbs*;
2. *articles and countability*;
3. *determiners*;
4. *non-finite forms*;
5. *orthography*;
6. *prepositions and particles*;
7. *pronouns;*
8. *questions, negatives, auxiliaries*;
9. *there is*;
10. *time, tense, aspect*;
11. *vocabulary*;
12. *word order*;
13. (nada)[105];

Com relação ao sistema de classificação proposto nesta pesquisa (SO2I), todos foram encontrados na amostra de 100 ocorrências:

14. substituição
15. inserção
16. omissão

[105] Como no decorrer da análise notei que o sistema de classificação baseado em Shepherd (2001) não contemplava erros como plural e omissão, criei este critério para permitir ao outro pesquisador que validaria a pesquisa identificar erros que não poderiam ser colocados em nenhuma categoria do referido sistema.

17. inversão

Como os critérios totalizavam 17, inseri na nova planilha criada, nos campos correspondentes aos sistemas de classificação baseado em Shepherd (2001) (a) e SO2I, desenvolvido nesta pesquisa (b), em cada linha contendo os erros, uma lista *drop-down*[106] contendo os 17 itens acima, os pertencentes ao sistema de classificação (a) na coluna F e os pertencentes ao sistema (b) na coluna G (ver figura a seguir). Desse modo, o outro pesquisador poderia selecionar os itens de classificação a cada análise sem a necessidade de escrevê-los.

A figura abaixo mostra a lista *drop down* a partir da qual o Anotador B poderia selecionar, a cada linha, a opção de erro que utilizaria:

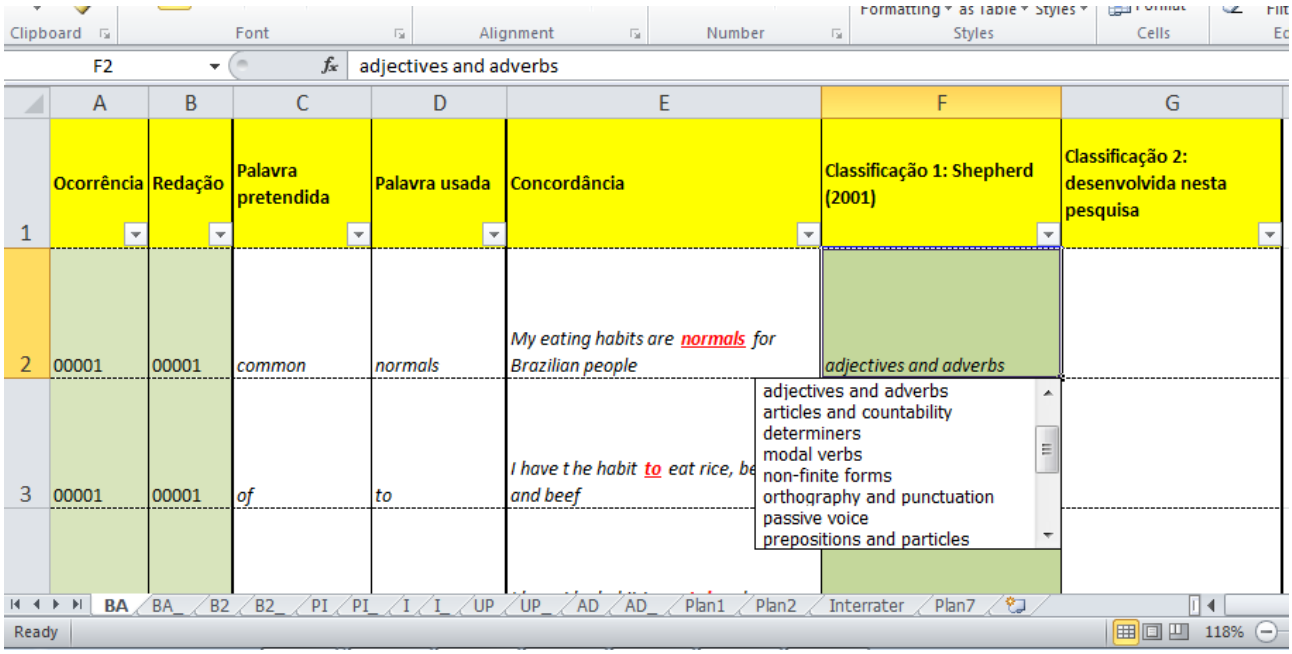


Tabela 2.4.32: impressão de página da planilha do programa computacional *Microsoft Excel 2010*, da suíte do *Microsoft Office 2010*[107].

[106] Em informática, sobretudo na área de *webdesign*, refere-se a uma linha que, quando clicada, se desdobra verticalmente em outras linhas contendo títulos diversos. A Tabela 2.4.32 traz uma visualização desse tipo de lista.

[107] Na coluna F pode-se ver a lista drop down contendo as categorias baseadas em Shepherd (2001). Cada linha dessa coluna e da coluna seguinte possui esse recurso.

Posteriormente, enviei por e-mail o arquivo à pesquisadora do GELC supracitada (doravante Anotador B), pois fui o primeiro anotador (doravante Anotador A), já que havia feito a primeira classificação dos erros.

Quando o Anotador B concluiu seu trabalho e me reenviou o arquivo contendo a planilha, constatei que nossas classificações divergiam em praticamente 85%. Ao confrontar suas classificações com as minhas, observei que as diferenças se concentravam, sobretudo, nos critérios a seguir:

a) baseado em Shepherd (2001): *articles and countability*; *conditionals*; *determiners*; *modal verbs*; *non-finite forms*; *pronouns*; *questions, negatives, auxiliaries*; *relatives*; *there is*;

Notei, então, que os critérios não eram claros o suficiente e, por isso, decidi apagar as classificações do Anotador B e reenviar as 100 linhas com erros para uma nova análise. Porém, desta vez inseri à direita da planilha contendo os erros uma outra planilha chamada “Glossário”, a qual continha explicações sobre as categorias que causaram discordância em nossas avaliações. Isso se fez necessário também porque o Anotador B não havia lido Shepherd (2001) e, por isso, não sabia exatamente o que cada item classificatório englobava. A próxima tabela reproduz a planilha Glossário:

Baseado em Shepherd (2001)	
articles and countability	Emprego ou omissão de artigo DEFINIDO (*THE*) como em: *I bought it in THE Oxford Street*;
conditionals	Erro no paralelismo de *if clauses*: *If I see him, I'll...*;
determiners	palavras seguidas de substantivos: são artigos, pronomes demonstrativos, pronomes possessivos, quantificadores, artigo "*The*" antes de adjetivo de posse; "*The*" antes de adjetivo de posse;
modal verbs	*should, must, can, have to, may, might, can* (omissão ou uso incorreto) ;
non-finite forms	*Gerund*; verbos + infinitivo c/ omissão de preposição: *she tried telephone you* (Ela quer QUE você ligue para ela);
pronouns	Pronomes pessoais, objeto e reflexivo;
questions, negatives, auxiliaries	Omissão ou inserção desnecessária de auxiliar; Uso incorreto de negação com "*not*" ou "*no*" Erro com *tag questions*;
relatives	Pronomes relativos: *that, who, which, whose...* ;
there is	Trocar "*is*" por "*are*" ou vice-versa; Usar "*have*" como existir; Usar "*exist*" como existir;
(nada)	Use isso quando não houver categorias correspondentes em Shepherd (2001).

Tabela 2.4.33: explicação dos critérios mais confusos que compunham o sistema de classificação baseado em Shepherd (2001)[108].

[108] Note-se que as categorias *conditionals*, *modal verbs*, e *relatives*, que na visão do Anotador A não haviam aparecido na amostra de 100 ocorrências, apareceram na classificação do Anotador B, provavelmete porque este, por não ter lido Shepherd (2001), não conhecia o contexto que levava um erro a ser classificado como um desses itens.

Após a nova análise, houve concordância entre o Anotador A e o Anotador B na grande maioria dos casos. Um cálculo simples considerando o número de concordâncias e o número de ocorrências revelou uma convergência de mais de 80%. Porém, a concordância entre avaliadores exige um cálculo estatístico, pois, desse modo, é possível se medir não somente a concordância específica dos dados analisados, mas se traçar uma estimativa com relação ao grau de concordância de dois ou mais avaliadores e se estabelecer um parecer sobre a eficiência de um critério de análise determinado.

Há diversas formas para se calcular estatisticamente a concordância entre avaliadores: Probabilidade de junção, ICC, Alfa e outras. Porém, um dos cálculos mais utilizados para esta finalidade é o chamado Kappa, o qual leva em consideração a concordância dos avaliadores que ocorreu por acaso, removendo-a (cf. Cohen, 1960, p. 40; Grayson e Rust, 2001, p. 72), motivo que justifica sua preferência.

Programas coputacionais como o SPSS (*Statistical Package for the Social Sciences*), da IBM, trazem em sua configuração a fórmula de cálculo para a verificação de concordância entre avaliadores por meio do Kappa. Nesta pesquisa utilizarei esse programa computacional para a geração dos resultados relativos à concordância entre avaliadores.

O cálculo estatístico Kappa é efetuado por meio da fórmula abaixo (cf. Cohen, 1960, p. 39):

$$\mathbf{K = \frac{P_o - P_c}{1 - P_c}}$$

A fórmula diz que o Kappa é calculado subtraindo-se a probabilidade hipotética de concordância ao acaso (P_c) da concordância relativa observada entre os anotadores (P_o) e dividindo-se por 1 menos P_c (cf. Cohen, 1960, p. 39).

O cálculo estatístico Kappa prevê 0,00 como o menor número (o que indica um resultado pobre) e 1,00 como o maior número possível, o que indica uma excelente concordância entre avaliadores (cf. Cohen, 1960, p. 46). Com base nessas possibilidades de resultados, alguns autores passaram a propor uma gradação, como no caso de Cicchetti (1994):

a) < 0,40 = Pobre
b) 0,40 a 0,59 = Razoável
c) 0,60 a 0,74 = Bom

d) > 0,74 = Excelente

Simon (2005/2008) propõe uma gradação diferente. Para ele:

a) concordância pobre = resultado inferior a 0,20;
b) concordância razoável = 0,20 a 0,40;
c) concordância moderada = 0,40 a 0,60;
d) concordância boa = 0,60 a 0,80;
e) concordância muito boa = 0,80 a 1,00.

Como foram usados dois procedimentos distintos de anotação de erro, foi preciso calcular Kappa separadamente para cada um deles.

Segundo o sistema de classificação baseado em Shepherd (2001) temos a seguinte matriz:

Anotador A \ Anotador B	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	
1	6			1										7
2		6											1	7
3		1	5					1					1	8
4				1						1				2
5					4									4
6						21								21
7							6							6
8									1					1
9									1					1
10					1					8				9
11											16		2	18
12												2		2
13	1	1			2						5			9
	7	8	5	2	7	21	6	1	2	9	21	2	4	**76**

Tabela 2.4.34: soma dos resultados da matriz de codificação para análise de concordância entre avaliadores segundo o sistema de classificação baseado em Shepherd (2001).

Na tabela, os números em negrito na vertical e na horizontal correspondem às seguintes categorias de erro:

1. *adjectives and adverbs*;
2. *articles and countability*;
3. *determiners*;
4. *non-finite forms*;
5. *orthography*;
6. *prepositions and particles*;
7. *pronouns;*
8. *questions, negatives, auxiliaries*;
9. *there is*;
10. *time, tense, aspect*;
11. *vocabulary*;
12. *word order*;
13. (nada)[109];

A matriz deve ser lida da seguinte maneira:

As linhas correspondem aos critérios observados pelo Anotador A. As colunas trazem os critérios de avaliação observados pelo Anotador B. Os números dentro da matriz correspondem à somatória dos erros encontrados para cada uma das categorias numericamente em negrito. Idealmente, os números dentro da matriz deveriam estar nas intersecções entre os números negritados, de modo a formar uma linha diagonal que decresce da esquerda para a direita (em destaque na tabela).

Observando a tabela anterior, pode-se notar que tanto o Anotador A (linha 1) quanto o B (coluna 1) classificaram 6 dos erros encontrados como erros do tipo 1 (uso de adjetivos ou advérbios). Porém, 1 erro classificado como tipo 1 pelo Anotador A (linha 1) foi classificado como erro do tipo 4 pelo Anotador B (coluna 4). Tal relação pode ser observada com relação aos outros dados da tabela.

[109] Como mencionado anteriormente, no decorrer da análise notei que o sistema de classificação baseado em Shepherd (2001) não contemplava erros como plural e omissão. Por isso criei este critério para permitir ao outro pesquisador que validará a pesquisa identificar erros que não poderiam ser colocados em nenhuma categoria do referido sistema.

Calculando o Kappa no *SPSS 19 for Mac* obtive o valor de 0,736 (p=0,000[110]).

Como visto anteriormente, o cálculo estatístico Kappa prevê 0,00 como o menor número (o que indica um resultado pobre) e 1,00 como o maior número possível, o que indica uma excelente concordância entre avaliadores (cf. Cohen, 1960, p. 46).

O resultado acima (0,736) indica uma boa concordândia entre avaliadores segundo os sistemas de gradação propostos por Cicchetti (1994) e por Simon (2005/2008).

Segundo o sistema colocacional SO2I, temos a seguinte matriz:

		Anotador B				
		substituição	**inserção**	**omissão**	**inversão**	
Anotador A	**substituição**	50				50
	inserção	4	12			16
	omissão	1		13		14
	inversão				2	2
		55	12	13	2	**77**

Tabela 2.4.35: soma dos resultados da matriz de codificação para análise de concordância entre avaliadores segundo o sistema de classificação SO2I, desenvolvido nesta pesquisa.

O valor de Kappa segundo o *SPSS 19 for Mac* para esse sistema colocacional é de 0,886 (p=0,000).

O resultado acima é indica uma excelente concordândia entre avaliadores segundo o sistema de gradação proposto por Cicchetti (1994), e uma concordância muito boa, segundo o sistema de gradação proposto por Simon (2005/2008). Porém, se confrontarmos o cálculo estatístico Kappa

[110] O "p" indica, no programa computacional SPSS, o grau de significância da prova estatística feita em cima dos resultados do Kappa. O valor 0,000 indica que o resultado do Kappa é significativo, isto é, pode ser confiado.

efetuado segundo os resultados gerados pelos anotadores tendo por base o sistema de classificação baseado por Shepherd (2001) e o sistema SO2I, proposto nesta pesquisa, veremos que este apresenta um grau de concordância maior, o que sugere que o sistema SO2I é mais confiável para a anotação de erros, como mostra o quadro a seguir:

Kappa - sistema baseado em Shepherd (2001): **0,736 de 1.0 (valor máximo)**	Kappa – sistema SO2I: **0,886 de 1.0 (valor máximo)**

Quadro 2.4.4: resultados do cálculo estatístico Kappa gerados a partir dos dados obtidos por meio do uso dos sistemas de classificação usados nesta pesquisa.

Neste capítulo foi apresentada a metodologia de pesquisa empregada neste estudo, incluindo a descrição do córpus, bem como a especificação dos procedimentos de coleta e criação do córpus COBRA-7, além dos procedimentos de recorte e criação do córpus COBRA-7_recorte e a metodologia de identificação e classificação dos erros. No capítulo a seguir serão apresentados e analisados os resultados da pesquisa.

Capítulo 3: Apresentação e Análise dos Resultados

Antes de partir para os resultados, convém reiterar o objetivo desta pesquisa: identificar e classificar os erros na escrita de aprendizes brasileiros de inglês. Um dos desdobramentos possíveis desta pesquisa seria a possibilidade de prover aos professores e pesquisadores um sistema de identificação e classificação de erros, com vistas a auxiliá-los em seu trabalho e informar a produção de materiais didáticos locais, isto é, voltados a aprendizes brasileiros.

Neste capítulo são apresentados e interpretados os resultados obtidos para cada uma das questões de pesquisa que nortearam o trabalho. Cada seção do capítulo trata de uma questão de pesquisa. Em cada seção são apresentados e interpretados os resultados em três passos:

a) com base no sistema de anotação de erros desenvolvido por Shepherd (2001);

b) com base nos critérios de classificação desenvolvidos nesta pesquisa;

c) com base na comparação dos dois critérios supracitados.

3.1 Questão 1: Quais os erros mais comuns no córpus COBRA-7_recorte?

3.1.1 Sistema de classificação baseado em Shepherd (2001)

De acordo com o primeiro sistema de classificação escolhido nesta pesquisa, os erros mais comuns no córpus COBRA-7_recorte são:

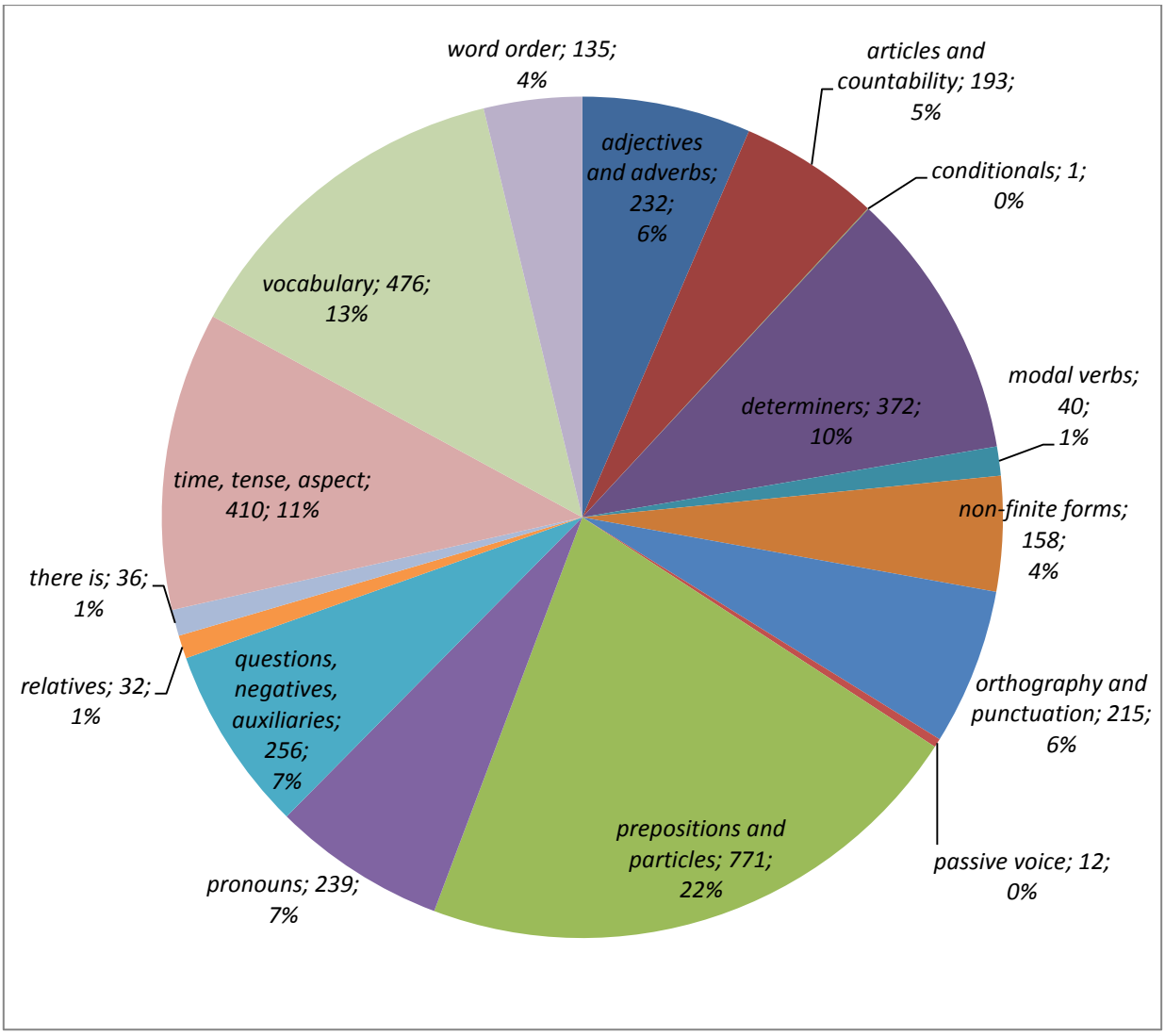


Figura 3.1.1: erros mais cometidos pelos aprendizes segundo critérios baseados em Shepherd (2001).

Os números indicam os valores absolutos das ocorrências, isto é, o número total de erros em cada categoria.

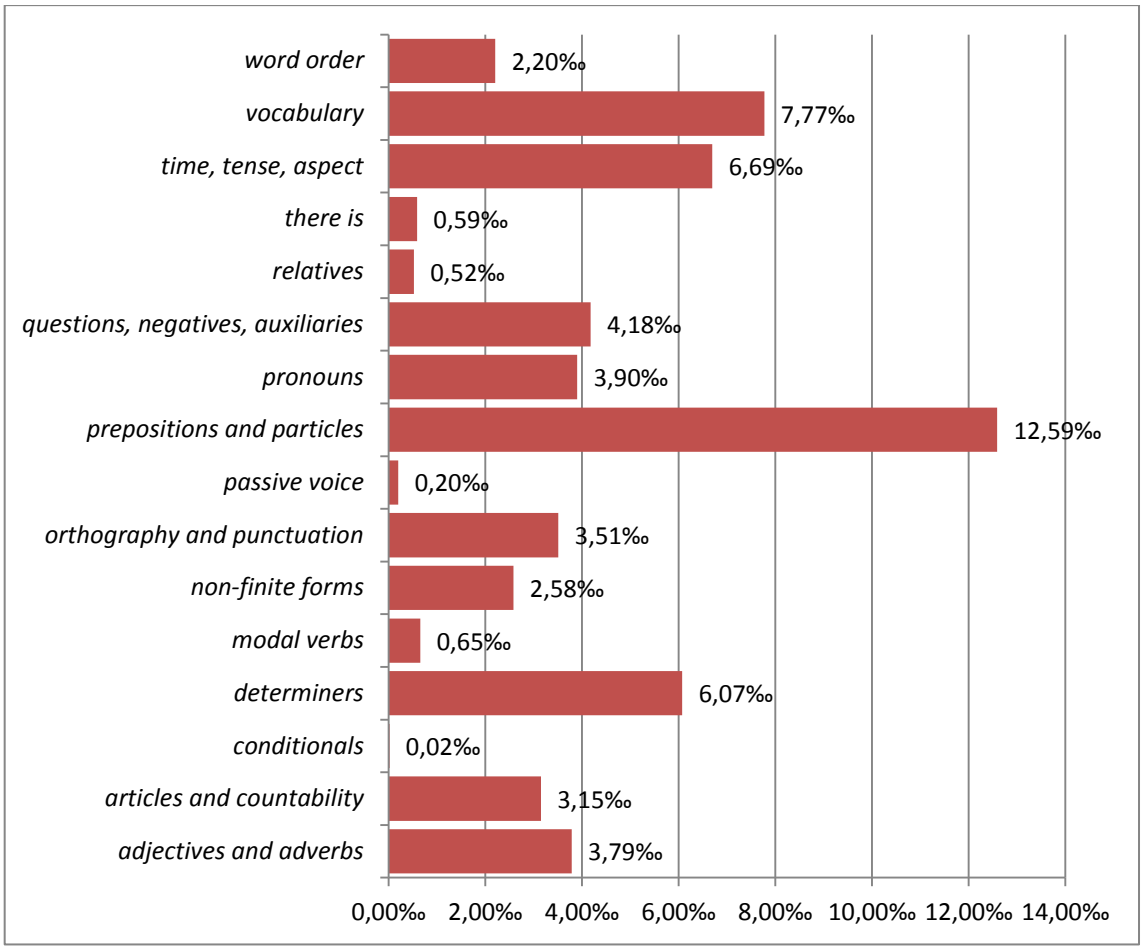


Figura 3.1. 2: erros mais cometidos pelos aprendizes segundo critérios de Shepherd (2001) (valores normalizados).

Os números no gráfico acima, ao lado das barras, indicam a frequência encontrada no córpus de estudo. Os resultados foram normalizados por mil. Isso significa que há 12,59 erros no uso de preposições e partículas a cada mil palavras do córpus (por isso o uso do símbolo "‰" – por mil). Relação semelhante pode ser traçada considerando os outros elementos do gráfico. Porque a normalização torna os valores estatisticamente comparáveis, utilizarei os valores normalizados em detrimento dos números absolutos apresentados no gráfico anterior. Tal escolha será tomada doravante para responder às questões de pesquisa deste trabalho.

O gráfico permite dizer que os quatro erros mais cometidos pelos aprendizes de inglês como língua estrangeira observados a partir do córpus COBRA-7_recorte e que podem refletir características gerais do aprendiz de inglês brasileiro são:

a) erros no uso de preposições e partículas (conjunções): 12,59‰; 771 ocorrências no córpus;

b) más escolhas de palavras (*vocabulary*): 7,77‰; 476 ocorrências no córpus;

c) tempo e aspecto verbal (*time, tense, aspect*): 6,69‰; 410 ocorrências no córpus;

d) erros no uso de determinantes (*determiners*): 6,07‰; 372 ocorrências no córpus;

3.1.2 Segundo o sistema SO2I, desenvolvido nesta pesquisa

De acordo com o sistema de classificação de erros SO2I, os erros mais comuns no córpus COBRA-7_recorte foram:

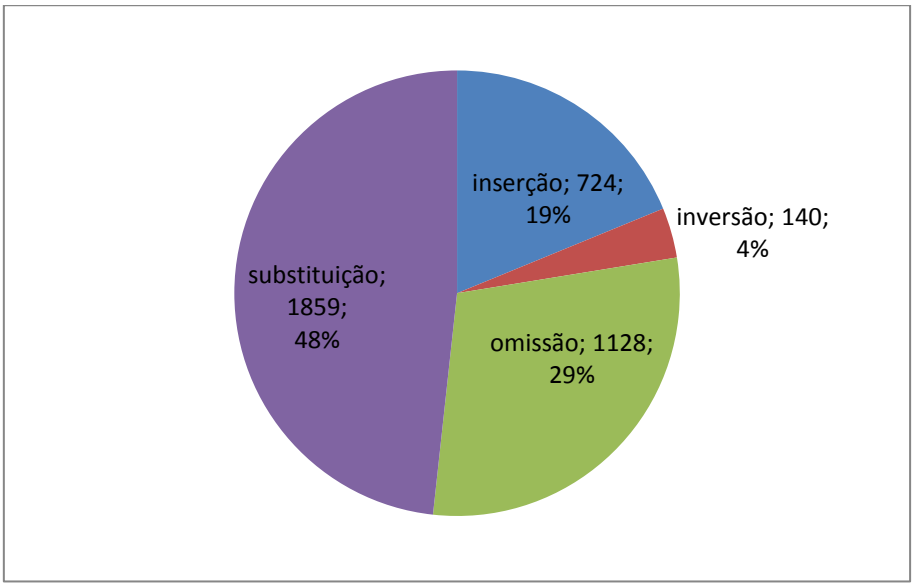


Figura 3.1.3: erros mais cometidos pelos aprendizes segundo o sistema de classificação SO2I.

Os números indicam os valores absolutos das ocorrências, isto é, o número total de erros em cada categoria.

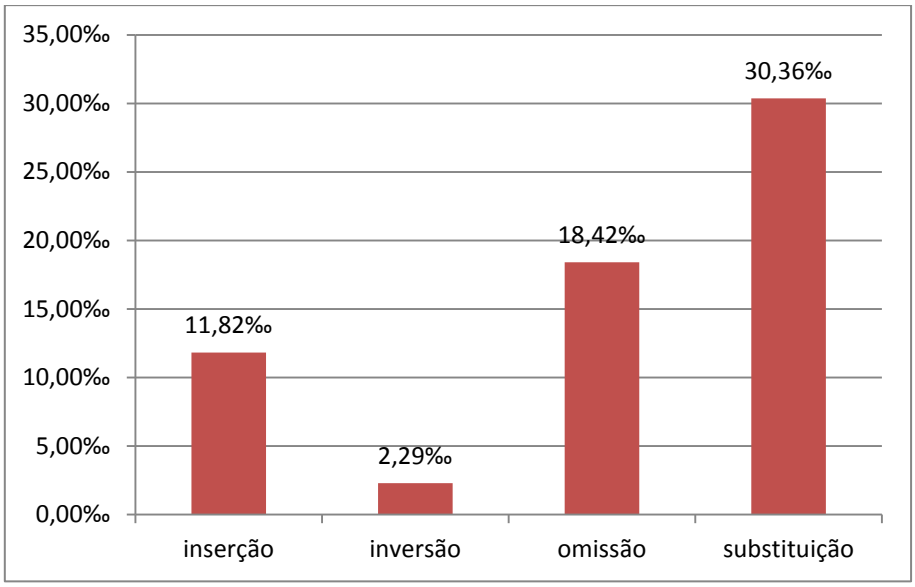


Figura 3.1.4: erros mais cometidos pelos aprendizes segundo o sistema de classificação SO2I (valores normalizados).

Na figura acima, os números indicam a frequência normalizada por mil dos erros encontrados no córpus de estudo. O gráfico mostra, do maior para o menor, os erros encontrados no córpus COBRA-7_recorte com relação a este sistema de classificação de erros:

a) substituição de palavras por uma outra que pertença ou não à mesma classe gramatical: 30,36‰; 1859 ocorrências no córpus;

b) omissão de uma palavra, sufixo ou prefixo: 18,42‰; 1128 ocorrências no córpus;

c) inserção de uma palavra, sufixo ou prefixo desnecessários: 11,82‰; 724 ocorrências no córpus);

d) inversão de palavras que compõem uma colocação: 11,82‰; 724 ocorrências no córpus);

3.1.3 Comparação entre os dois sistemas de classificação

Baseado em Shepherd (2001):		**Sistema SO2I:**	
prepositions and particles;	12,59‰	substituição	30,36‰
vocabulary;	7,77‰	omissão	18,42‰
time, tense, aspect;	6,69‰	inserção	11,82‰
determiners;	6,07‰	inversão	2,29‰

Quadro 3.1.1: erros mais comuns segundo o sistema de classificação baseado em Shepherd (2001) e o SO2I, proposto nesta pesquisa.

É importante lembrar que o córpus COBRA-7_recorte foi analisado utilizando-se, simultaneamente, os dois sistemas de classificação aqui expressos: o baseado em Shepherd (2001) e o SO2I, desenvolvido nesta pesquisa. Por isso, embora os dois sistemas possam ser relacionados, podem também ser considerados separadamente.

Pela característica desta pesquisa, pode-se imaginar que ela teve caráter meramente gramatical. Porém, como visto na Fundamentação Teórica, léxico e gramática são inseparáveis. Portanto, uma escolha gramatical é, ao mesmo tempo, colocacional, como indica a análise dos dados no capítulo anterior.

O quadro acima sugere que os erros agrupados segundo o modelo baseado em Shepherd (2001) (à esquerda na tabela acima) são "materializados" nas redações dos aprendizes por meio de um dos quatro critérios de classificação desenvolvidos nesta pesquisa (à direita na tabela acima). Isso significa, por exemplo, que um erro no uso de uma preposição (primeiro item do sistema de classificação baseado em Shepherd, 2001 – tabela acima) pode ter sido causado pela[111]:

a) substituição da preposição correta por uma incorreta;

b) omissão da preposição;

c) inserção de uma preposição não esperada;

[111] Considero aqui o sistema de classificação proposto nesta pesquisa.

d) inversão da ordem das palavras que compõem uma colocação da qual uma preposição faça parte.

Essa mesma lógica de cruzamento de informação entre os dois critérios de classificação usados nesta pesquisa se aplica às outras categorias gramaticais propostas por Shepherd (2001).

Isso indica que pode ser importante aos professores dar um enfoque maior nesses quesitos ao longo dos níveis de curso.

3.2 Questão 2: Qual a variação de erro entre os níveis no córpus COBRA-7_recorte?

3.2.1 Nível de curso básico 1

3.2.1.1 Segundo o sistema de classificação baseado em Shepherd (2001)

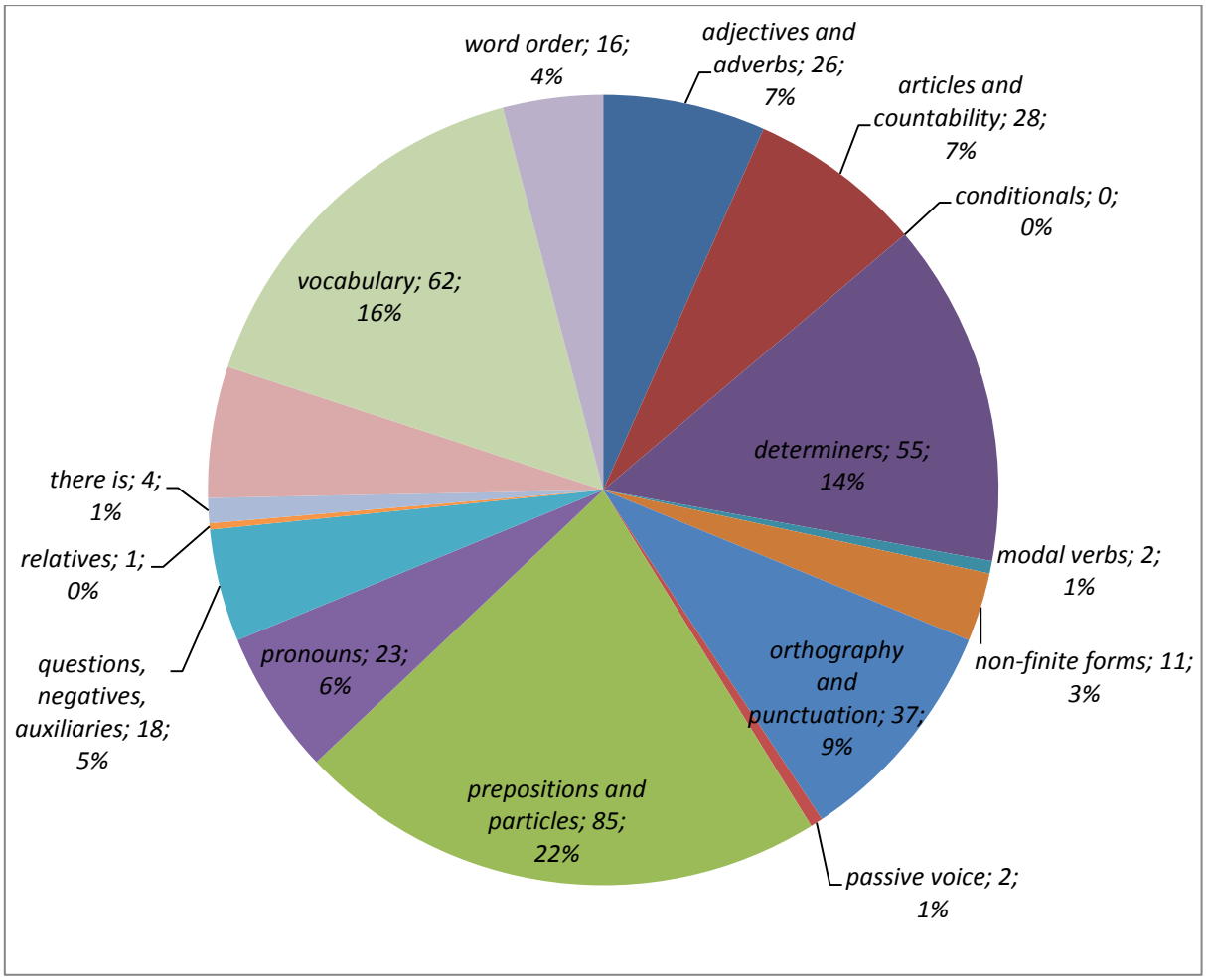


Figura 3.2.1: gráfico com os os erros mais cometidos pelos aprendizes no nível básico 1.

Os valores numéricos indicam a somatória bruta das ocorrências.

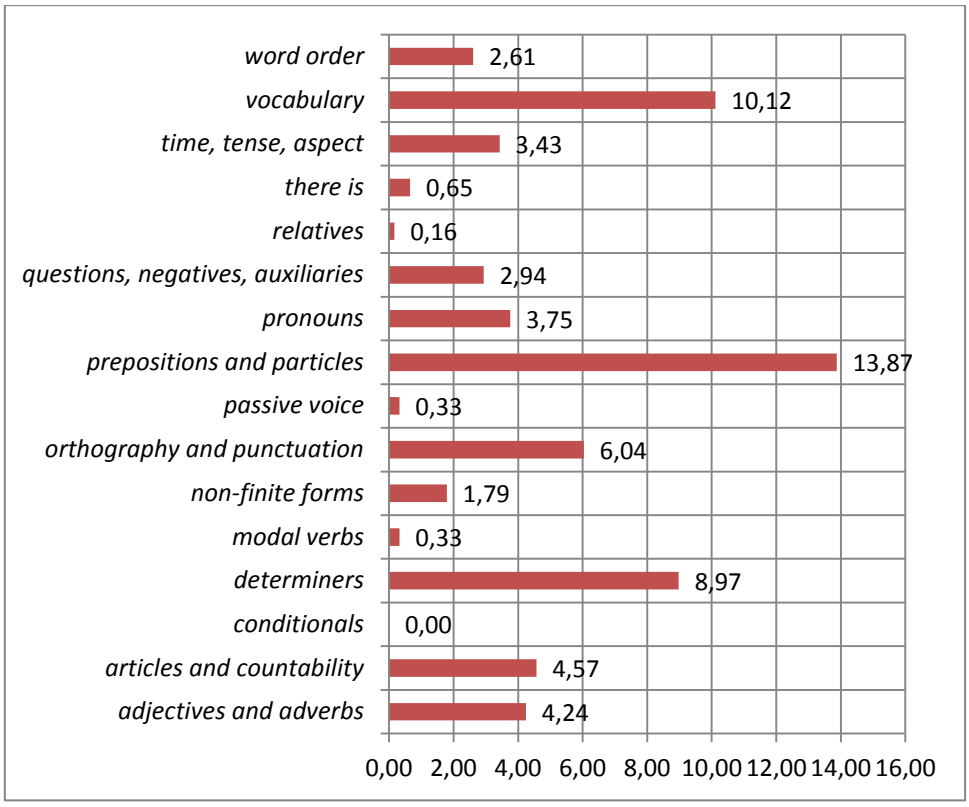


Figura 3.2. 2: gráfico com os os erros mais cometidos pelos aprendizes no nível básico 1 (valores normalizados).

Os valores no gráfico, ao lado das barras, indicam as ocorrências normalizadas por mil. O gráfico mostra que os erros mais comuns no nível básico 1 segundo este sistema de classificação são:

a) uso inadequado de preposições ou partículas (conjunções): 13,87‰ (85 ocorrências);

b) escolha de palavras (*vocabulary*): 10,12‰ (62 ocorrências);

c) uso de determinantes: 8,97‰ (55 ocorrências no córpus);

d) ortografia: 6,04‰ (37 ocorrências).

3.2.1.2 Segundo o sistema de classificação SO2I, desenvolvido neste trabalho

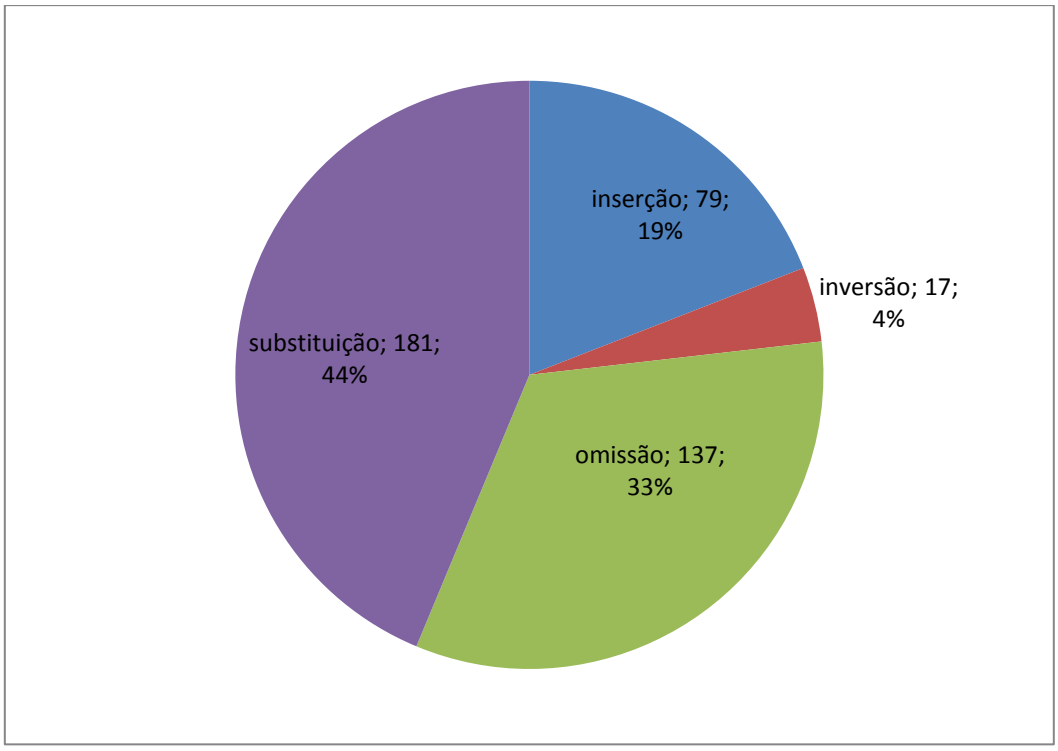


Figura 3.2.3: os erros mais comuns no nível básico 1 segundo o sistema de classificação desenvolvido nesta pesquisa.

Os números no gráfico indicam os valores absolutos das ocorrências, isto é, o número total de erros em cada categoria.

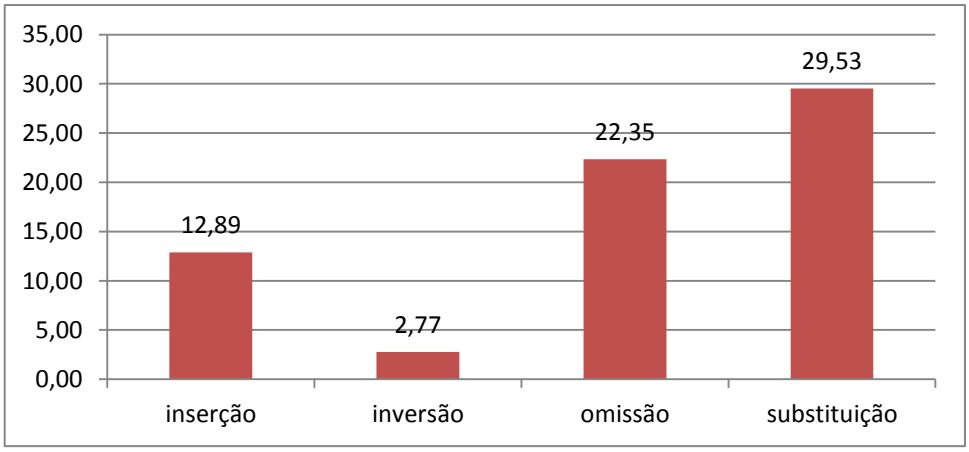


Figura 3.2.4: os erros mais comuns no nível básico 1 segundo o sistema de classificação SO2I (valores normalizados).

Os números no gráfico acima sobre as barras indicam a frequência normalizada por mil. O gráfico indica que os erros mais cometidos pelos aprendizes no nível de curso básico 1 segundo o sistema de classificação SO2I, desenvolvido nesta pesquisa, são:

a) substituição: 29,53‰ (181 ocorrências);

b) omissão: 22,35‰ (137 ocorrências);

c) inserção: 12,89‰ (79 ocorrências);

d) inversão: 2,77‰ (17 ocorrências).

3.2.1.3 Comparação entre os sistemas de classificação usados nesta pesquisa

Baseado em Shepherd (2001):		**Sistema SO2I:**	
prepositions and particles	13,87‰	substituição	29,53‰
vocabulary	10,12‰	omissão	22,35‰
determiners	8,97‰	inserção	12,89‰
orthography and puctuation	6,04‰	inversão	2,77‰

Quadro 3.2.1: erros mais comuns no nível de curso Básico 1 segundo o sistema de classificação baseado em Shepherd (2001) e o SO2I, proposto nesta pesquisa.

O quadro indica que os aprendizes deste nível de curso cometem os erros considerados pelo sistema de classificação baseado em Shepherd (2001) (acima à esquerda) de uma das seguintes maneiras (acima à direita):

a) substituindo uma palavra por outra que pertença ou não à mesma classe gramatical;

b) omitindo palavras, prefixos ou sufixos;

c) inserindo palavras, prefixos ou sufixos;

d) invertendo palavras que compõem a colocação pretendida.

3.2.2 Nível de curso básico 2

3.2.2.1 Segundo o sistema de classificação baseado em Shepherd (2001)

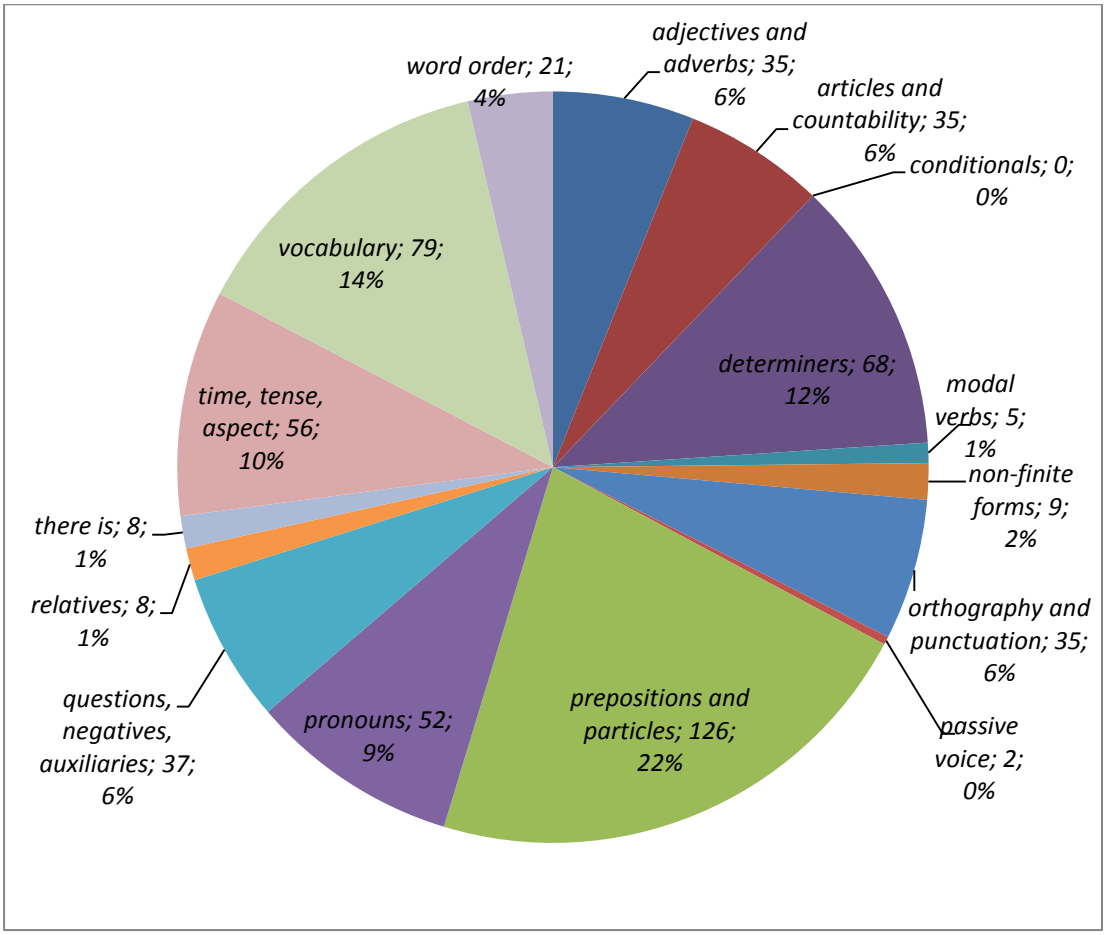


Figura 3.2.5: gráfico com os os erros mais cometidos pelos aprendizes no nível básico 2.

Os valores numéricos dentro do gráfico indicam a somatória bruta das ocorrências.

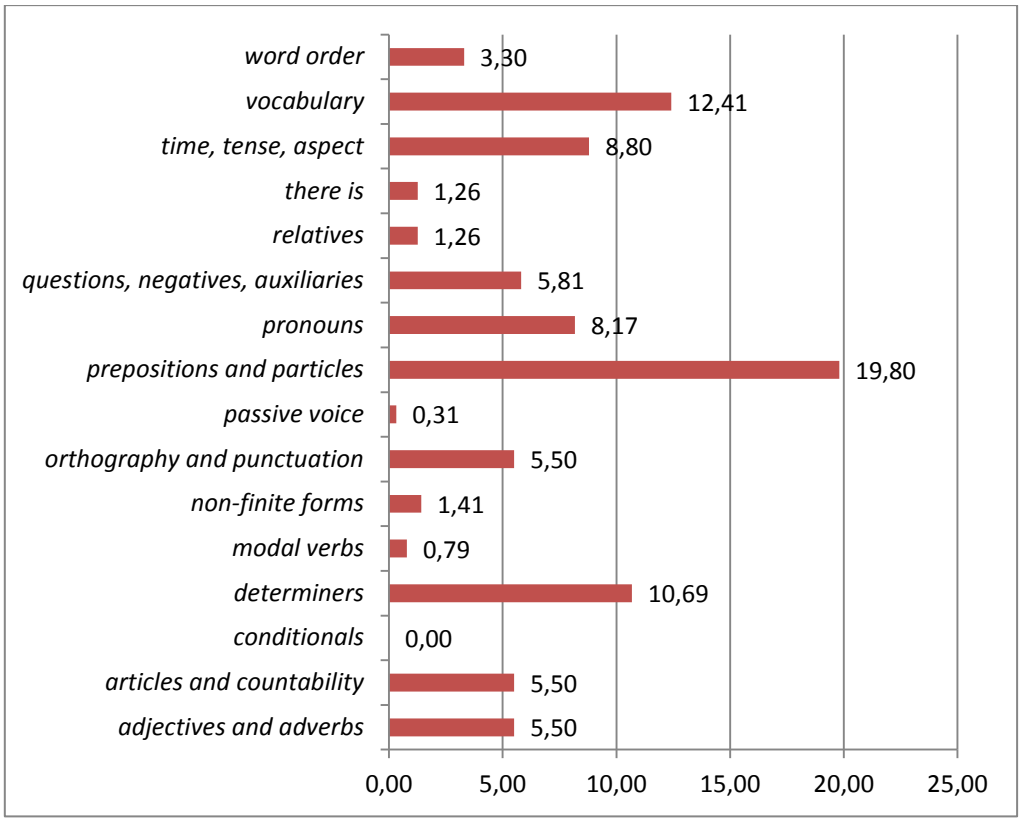


Figura 3.2.6: gráfico com os os erros mais cometidos pelos aprendizes no nível básico 2 (valores normalizados).

Os valores no gráfico acima ao lado das barras indicam as ocorrências normalizadas por mil. O gráfico mostra que os erros mais comuns no nível básico 2 segundo este sistema de classificação são:

a) uso inadequado de preposições ou partículas (conjunções): 19,80‰ (126 ocorrências);

b) escolha de palavras (*vocabulary*): 12,41‰ (79 ocorrências);

c) uso de determinantes: 10,69‰ (68 ocorrências no córpus);

d) tempo e aspecto verbal: 8,80‰ (56 ocorrências).

3.2.2.2 Segundo o sistema de classificação SO2I, desenvolvido neste trabalho

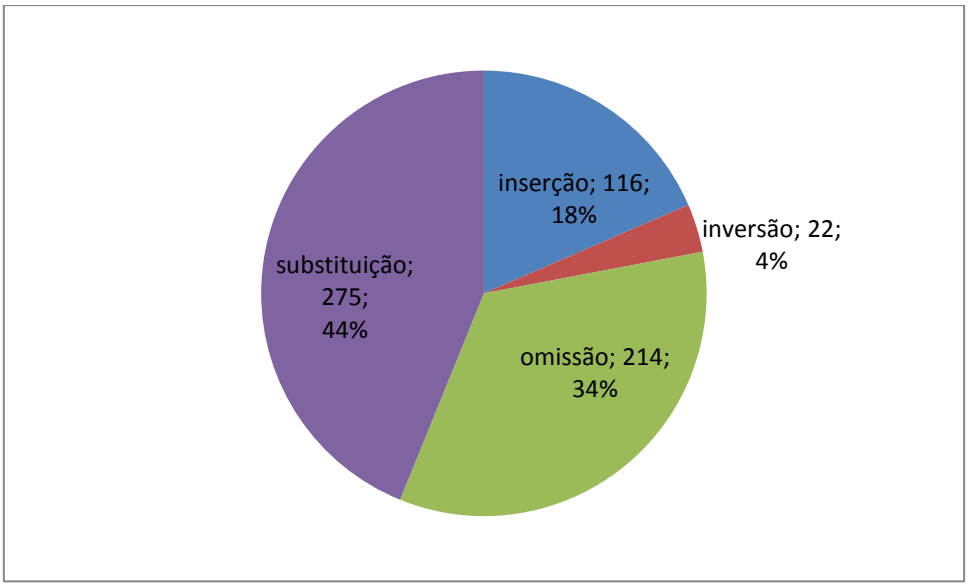


Figura 3.2.7: os erros mais comuns no nível básico 2 segundo o sistema de classificação desenvolvido nesta pesquisa.

Os números indicam os valores absolutos das ocorrências, isto é, o número total de erros em cada categoria.

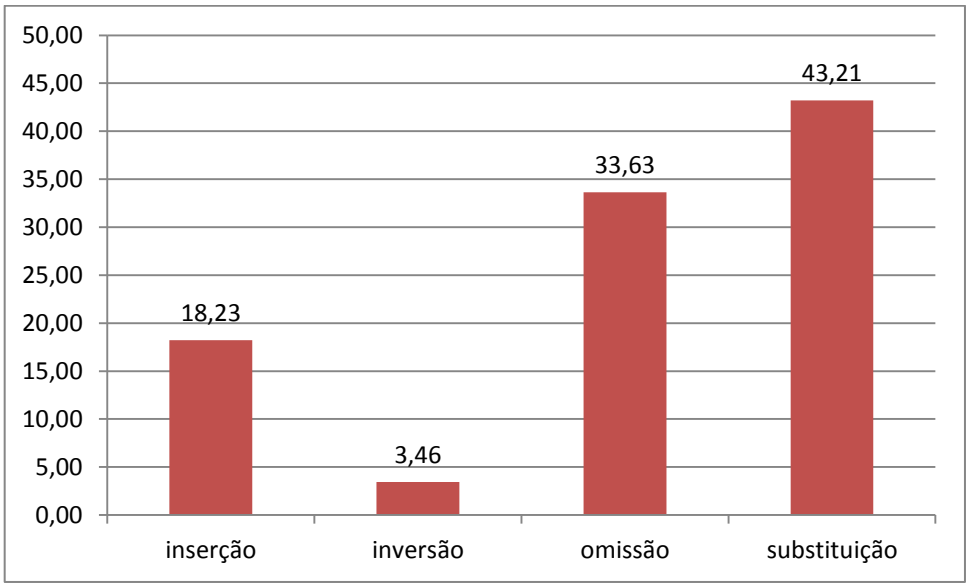


Figura 3.2.8: os erros mais comuns no nível básico 2 segundo o sistema de classificação SO2I (valores normalizados).

Os números no gráfico acima sobre as barras indicam a frequência normalizada por mil. O gráfico indica que os erros mais cometidos pelos aprendizes no nível de curso básico 2 segundo o sistema de classificação desenvolvido nesta pesquisa são:

a) substituição: 43,21‰ (275 ocorrências);

b) omissão: 33,63‰ (214 ocorrências);

c) inserção: 18,23‰ (116 ocorrências);

d) inversão: 3,6‰ (22 ocorrências).

3.2.2.3 Comparação entre os sistemas de classificação usados nesta pesquisa

Baseado em Shepherd (2001):		**Sistema SO2I:**	
prepositions and particles	19,80‰	substituição	43,21‰
vocabulary	12,41‰	omissão	33,63‰
determiners	10,69‰	inserção	18,23‰
time, tense, aspect	8,80‰	inversão	3,60‰

Quadro 3.2.2: resumo das ocorrências de erros mais comuns no nível de curso básico 2segundo o sistema de classificação baseado em Shepherd (2001) e o proposto nesta pesquisa mostrando.

O quadro indica que os aprendizes do nível básico 2 cometem os erros considerados segundo o sistema de classificação baseado em Shepherd (2001) de uma das maneiras a seguir:

a) substituindo uma palavra por outra que pertença ou não à mesma classe gramatical;

b) omitindo palavras, prefixos ou sufixos;

c) inserindo palavras, prefixos ou sufixos;

d) invertendo palavras que compõem a colocação pretendida.

Note-se que, com relação ao nível anterior, os problemas com uso de tempo e aspecto verbal substituíram os problemas de ortografia. Isso provavelmente se deu porque, nesse nível, os aprendizes são expostos a outros tempos verbais como os futuros com *will*, com *be going to*, o presente progressivo e mesmo o passado simples. Porém, ainda assim os dados sugerem que nesse nível os aprendizes têm mais dificuldade com o uso de preposições e a escolha de palavras do que no nível anterior.

3.2.3 Nível de curso pré-intermediário

3.2.3.1 Segundo o sistema de classificação baseado em Shepherd (2001)

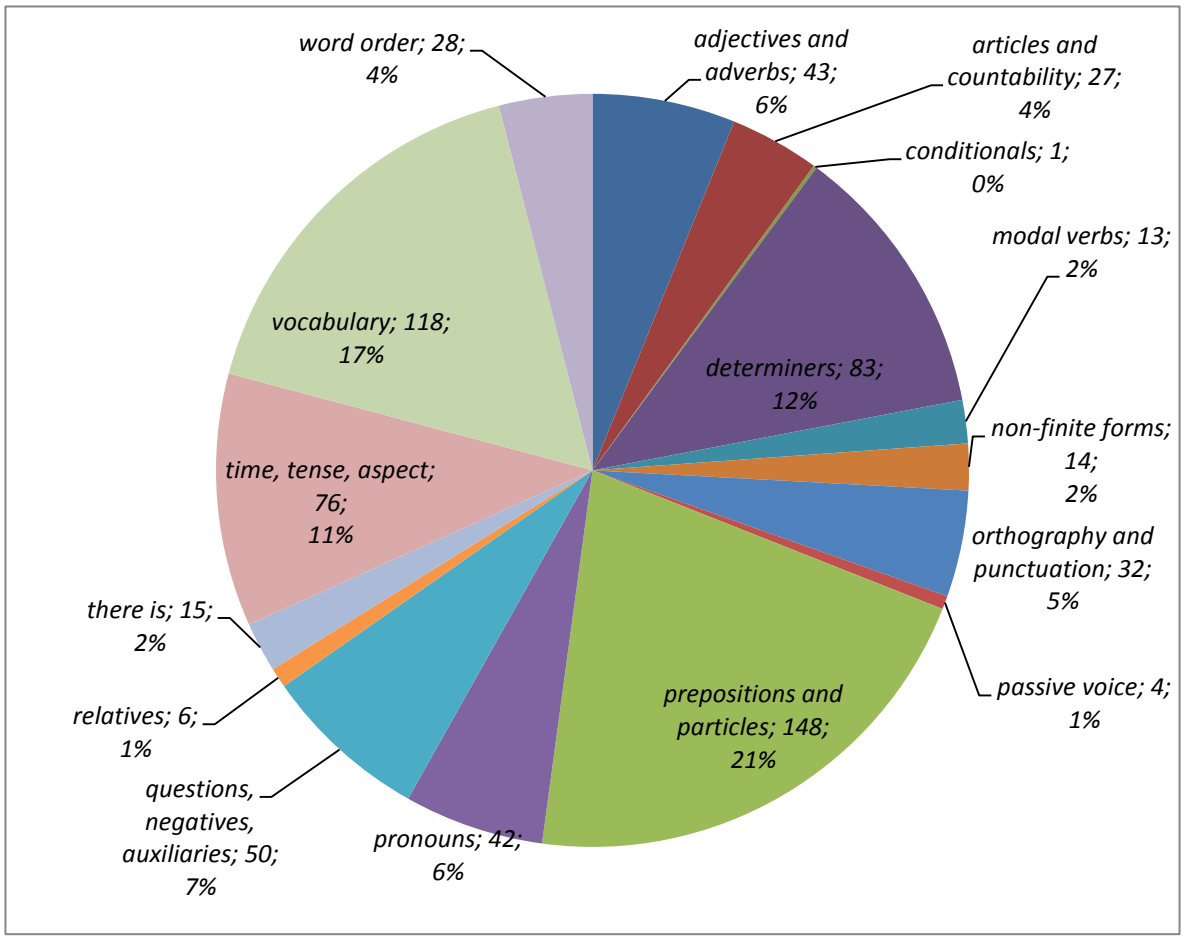


Figura 3.2.9: gráfico com os os erros mais cometidos pelos aprendizes no nível pré-intermediário.

Os valores numéricos indicam a somatória bruta das ocorrências para cada categoria.

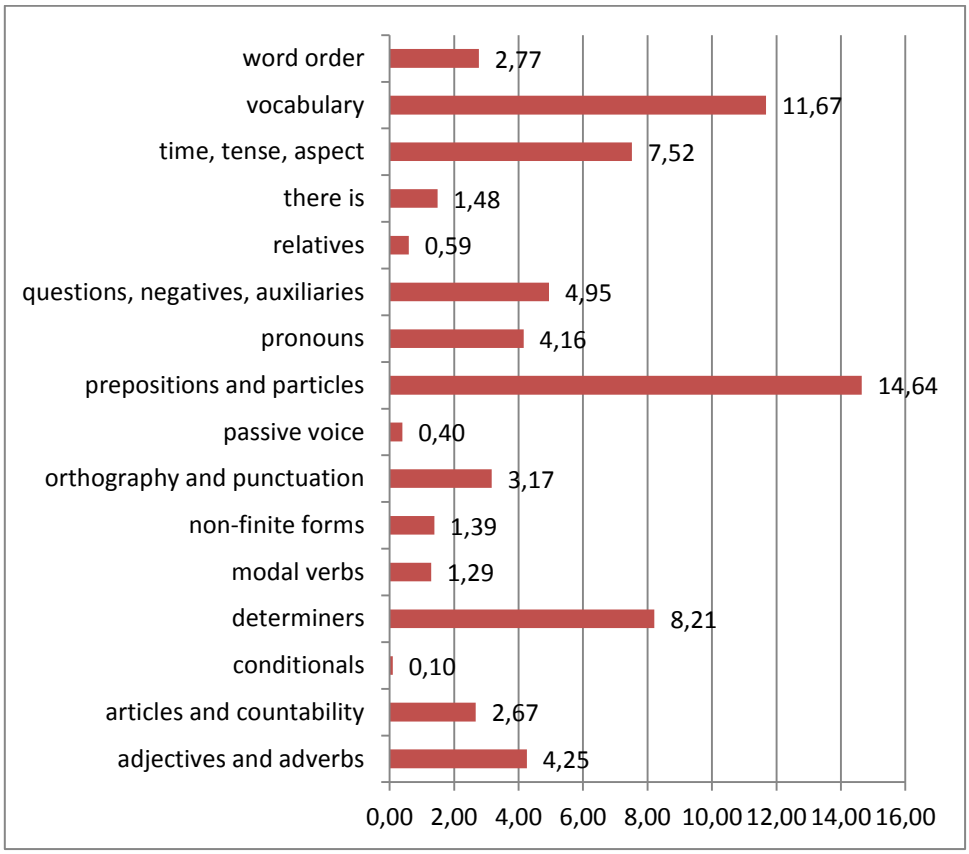


Figura 3.2.10: gráfico com os os erros mais cometidos pelos aprendizes no nível pré-intermediário (valores normalizados).

Acima, os valores no gráfico acima ao lado das barras indicam as ocorrências normalizadas por mil. O gráfico mostra que os erros mais comuns no nível pré-intermediário segundo o sistema de classificação baseado em Shepherd (2001) são:

a) uso inadequado de preposições ou partículas (conjunções): 14,64‰ (148 ocorrências);

b) escolha de palavras (*vocabulary*): 11,67‰ (118 ocorrências);

c) uso de determinantes: 8,21‰ (83 ocorrências no córpus);

d) tempo e aspecto verbal: 7,52‰ (76 ocorrências).

3.2.3.2 Segundo o sistema de classificação SO2I, desenvolvido neste trabalho

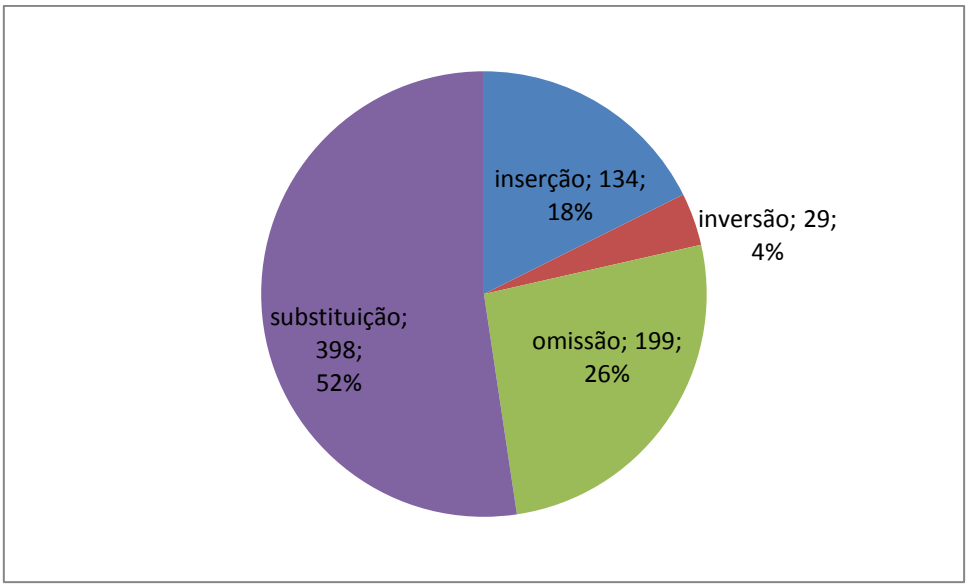


Figura 3.2.11: os erros mais comuns no nível pré-intermediário segundo o sistema de classificação SO2I.

Os números no gráfico indicam os valores absolutos das ocorrências, isto é, o número total de erros em cada categoria.

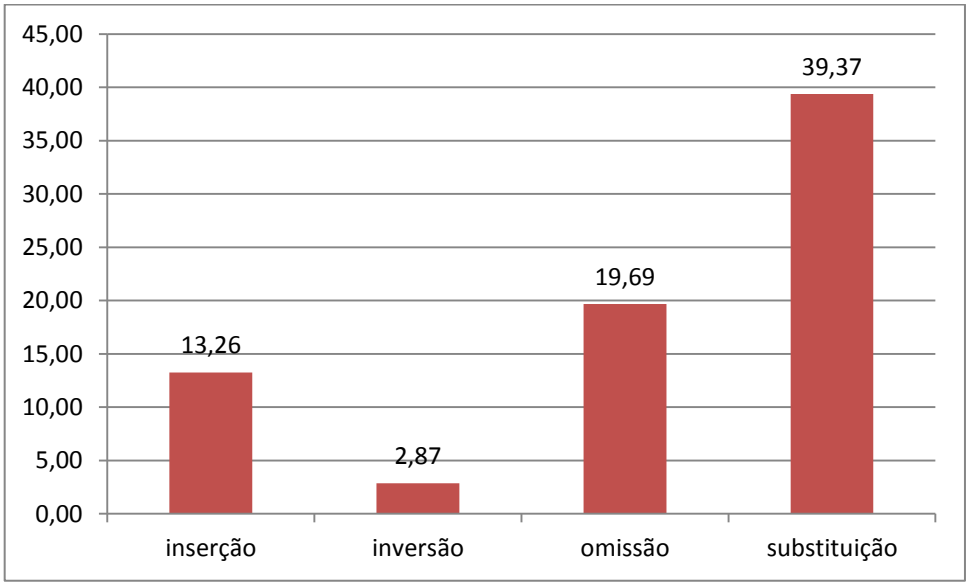


Figura 3.2.12: os erros mais comuns no nível pré-intermediário segundo o sistema de classificação SO2I (valores normalizados).

Os números no gráfico acima sobre as barras indicam a frequência normalizada por mil. O gráfico mostra que os erros mais cometidos pelos aprendizes no nível de curso pré-intermediário segundo o sistema de classificação desenvolvido nesta pesquisa são:

a) substituição: 39,37‰ (398 ocorrências);

b) omissão: 19,69‰ (199 ocorrências);

c) inserção: 13,26‰ (134 ocorrências);

d) inversão: 2,87‰ (29 ocorrências).

3.2.3.3 Comparação entre os sistemas de classificação usados nesta pesquisa

Baseado em Shepherd (2001):		**Sistema desenvolvido nesta pesquisa:**	
prepositions and particles	14,64‰	substituição	39,37‰
vocabulary	11,67‰	omissão	19,69‰
determiners	8,21‰	inserção	13,26‰
time, tense, aspect	7,52‰	inversão	2,87‰

Quadro 3.2.3: erros mais comuns no nível de curso pré-intermediário segundo o sistema de classificação baseado em Shepherd (2001) e o SO2I, proposto nesta pesquisa.

O quadro indica que, neste nível, os erros mais cometidos pelos aprendizes considerando o sistema de classificação de Shepherd (2001) são com relação ao uso de preposições ou partículas (14,64‰), vocabulário (11,67‰), determinantes (8,21‰), e tempos e aspecto verbal (7,52‰), e ocorrem devido a um dos aspectos abaixo:

a) substituição de uma palavra por outra que pertença ou não à mesma classe gramatical;

b) omissão de palavras, prefixos ou sufixos;

c) inserção de palavras, prefixos ou sufixos;

d) inversão de palavras que compõem a colocação pretendida.

3.2.4 Nível de curso intermediário

3.2.4.1 Segundo o sistema de classificação baseado em Shepherd (2001)

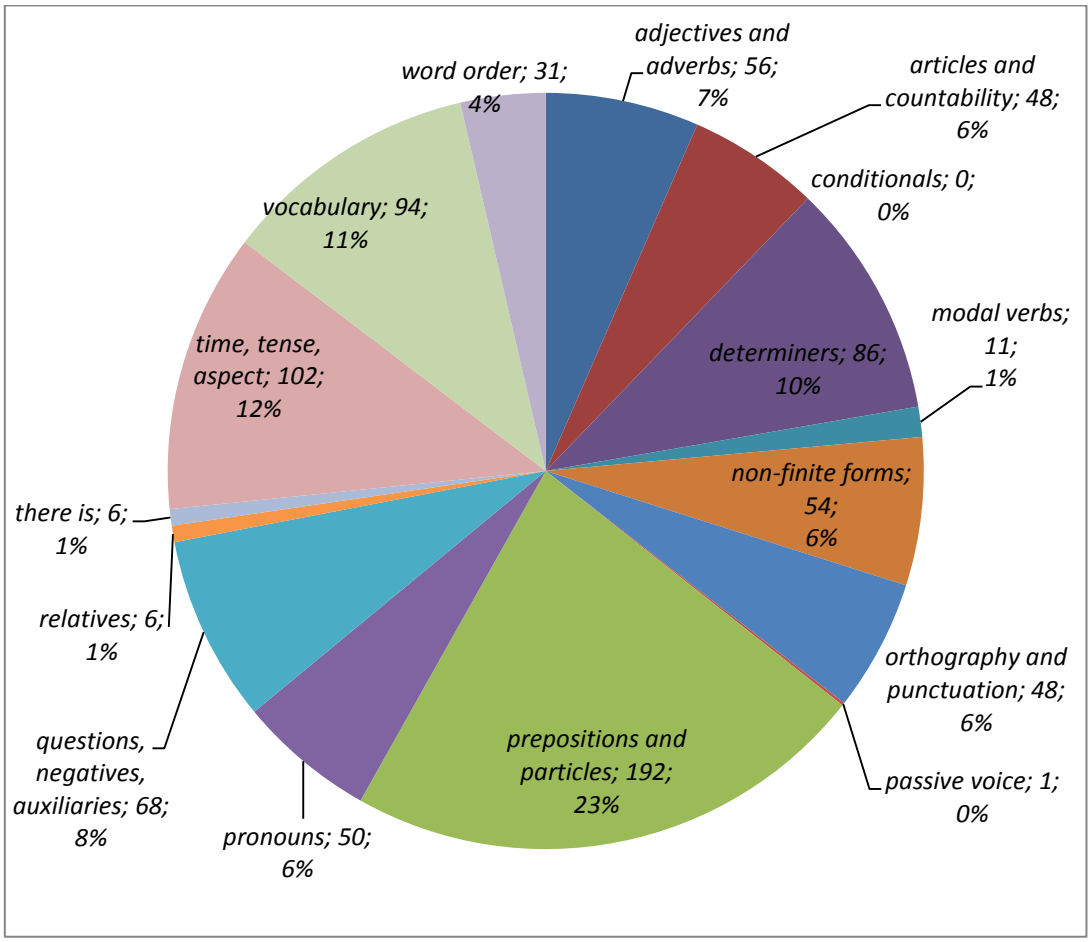


Figura 3.2.13: gráfico com os os erros mais cometidos pelos aprendizes no nível intermediário.

Os valores numéricos indicam a somatória bruta das ocorrências para cada categoria.

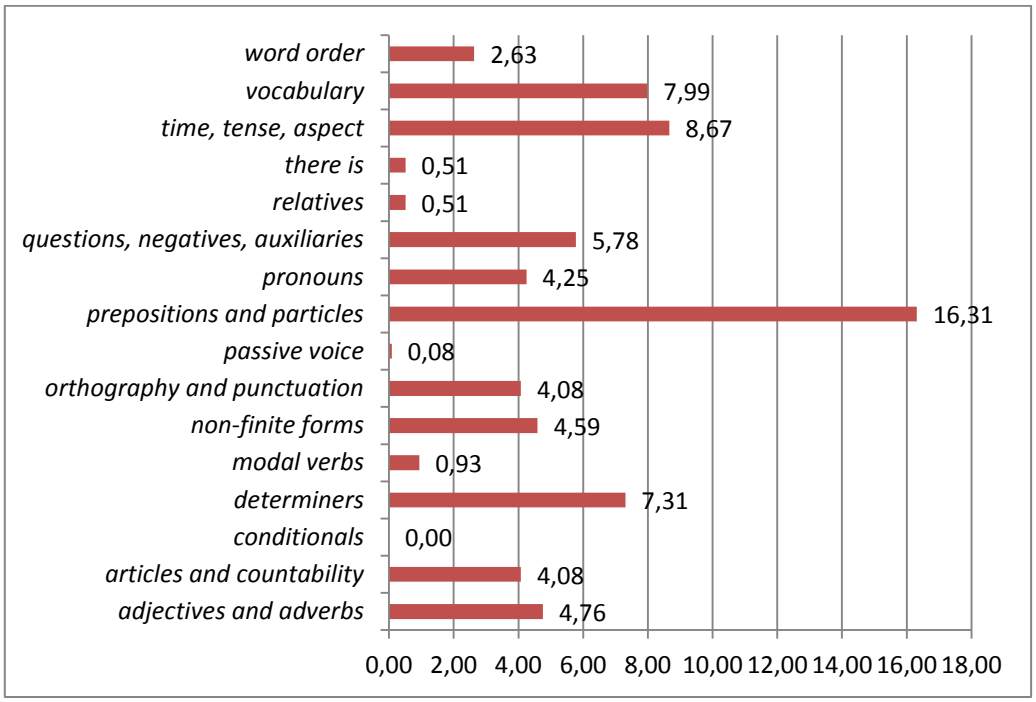


Figura 3.2.14: gráfico com os os erros mais cometidos pelos aprendizes no nível intermediário (valores normalizados).

Os valores no gráfico acima ao lado das barras no gráfico acima indicam as ocorrências normalizadas por mil. O gráfico mostra que os erros mais comuns no nível intermediário segundo o sistema de classificação baseado em Shepherd (2001) são:

a) uso inadequado de preposições ou partículas (conjunções): 16,31‰ (192 ocorrências);

b) tempo e aspecto verbal: 8,67‰ (102 ocorrências);

c) escolha de palavras (*vocabulary*): 7,99‰ (94 ocorrências no córpus);

d) formação de questões e o uso de negações e auxiliares: 5,78‰ (68 ocorrências).

3.2.4.2 Segundo o sistema de classificação SO2I, desenvolvido neste trabalho

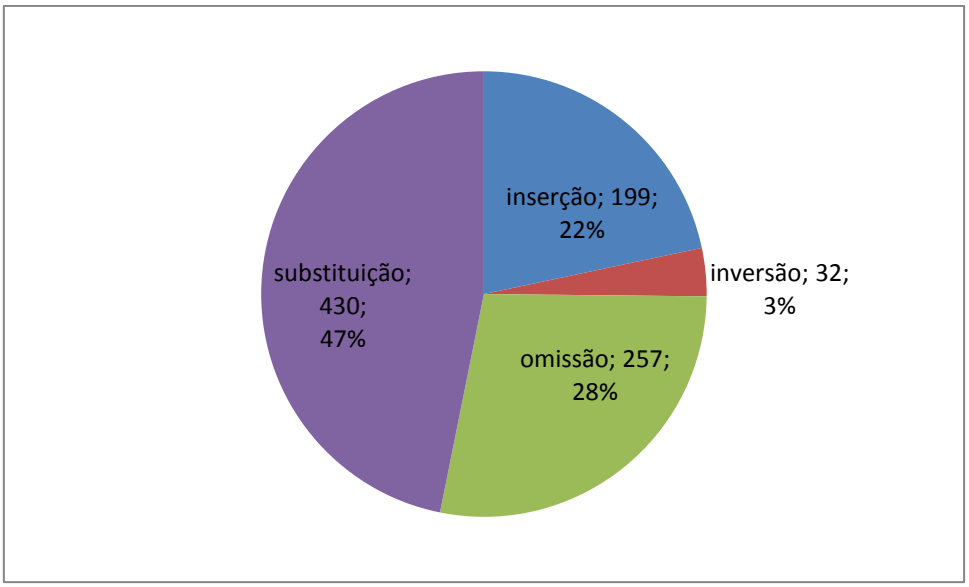


Figura 3.2.15: os erros mais comuns no nível intermediário segundo o sistema de classificação desenvolvido nesta pesquisa.

Os números indicam os valores absolutos das ocorrências, isto é, o número total de erros em cada categoria.

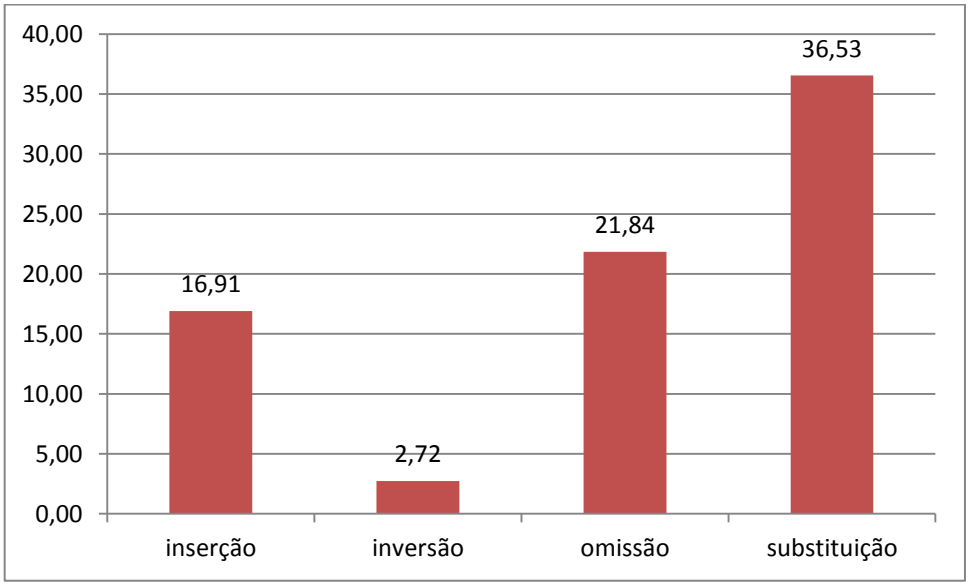


Figura 3.2.16: os erros mais comuns no nível intermediário segundo o sistema de classificação SO2I, desenvolvido nesta pesquisa (valores normalizados).

Na figura, os números no gráfico anterior sobre as barras indicam a frequência normalizada por mil. O gráfico indica que os erros mais cometidos pelos aprendizes no nível de curso intermediário segundo o sistema de classificação desenvolvido nesta pesquisa são:

a) substituição: 36,53‰ (430 ocorrências);

b) omissão: 21,84‰ (257 ocorrências);

c) inserção: 16,91‰ (199 ocorrências);

d) inversão: 2,72‰ (32 ocorrências).

3.2.4.3 Comparação entre os sistemas de classificação usados nesta pesquisa

Baseado em Shepherd (2001):		**Sistema SO2I:**	
prepositions and particles	16,31‰	substituição	36,53‰
time, tense, aspect	8,67‰	omissão	21,84‰
vocabulary	7,99‰	inserção	16,91‰
questions, negatives, auxiliaries	5,78‰	inversão	2,72‰

Quadro 3.2. 4: erros mais comuns no nível de curso intermediário segundo o sistema de classificação baseado em Shepherd (2001) e o SO2I, proposto nesta pesquisa.

O quadro indica que os aprendizes do nível intermediário cometem erros com o uso de preposições de partículas (conjunções); tempo e aspecto verbal; escolha de palavras; e formação de questões e uso de negação e auxiliares de uma das maneiras a seguir:

a) substituindo uma palavra por outra que pertença ou não à mesma classe gramatical;

b) omitindo palavras, prefixos ou sufixos;

c) inserindo palavras, prefixos ou sufixos;

d) invertendo palavras que compõem a colocação pretendida.

3.2.5 Nível de curso intermediário superior

3.2.5.1 Segundo o sistema de classificação baseado em Shepherd (2001)

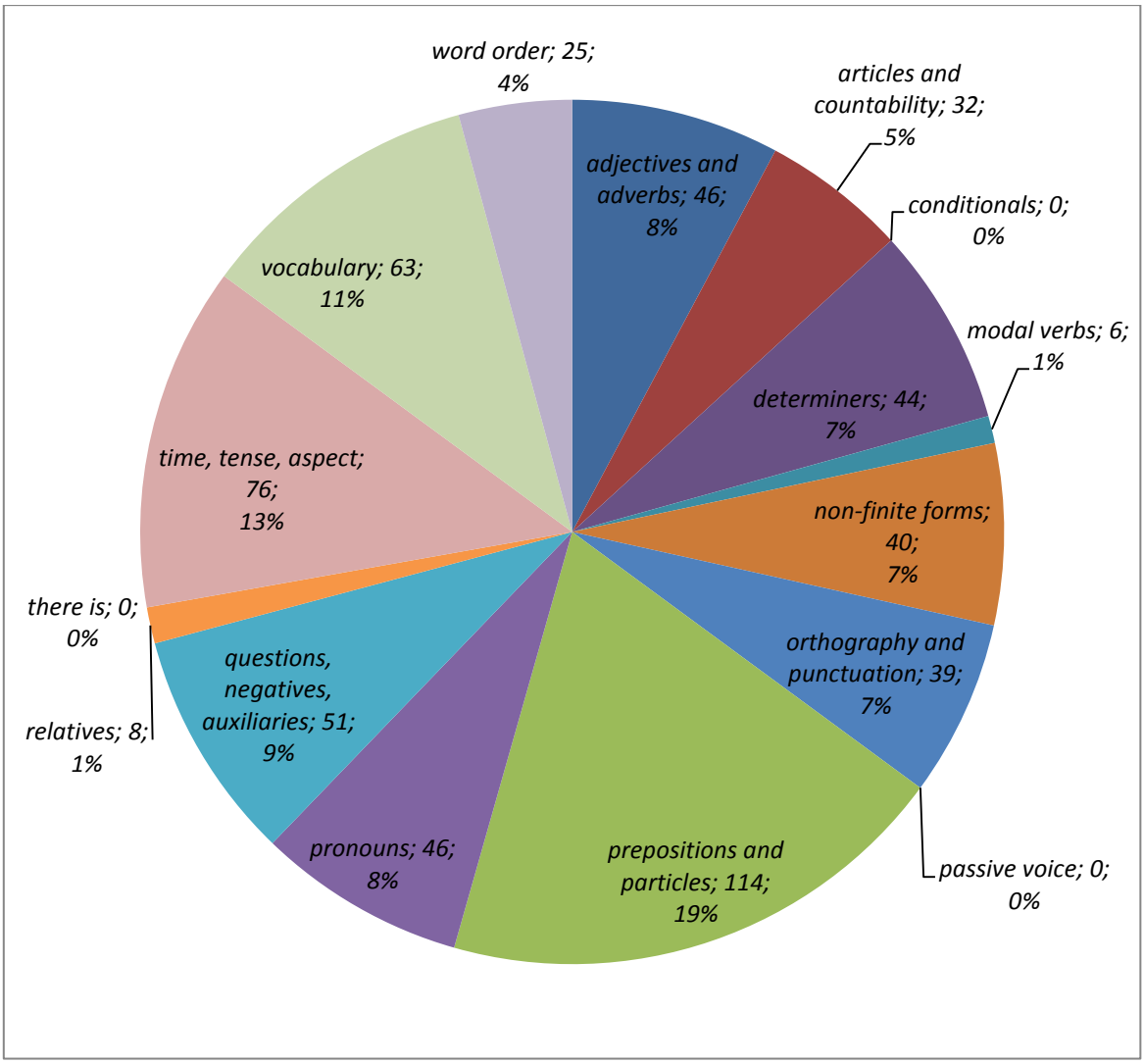


Figura 3.2.17: gráfico com os os erros mais cometidos pelos aprendizes no nível intermediário superior.

Os valores numéricos no gráfico acima indicam a somatória bruta das ocorrências para cada categoria.

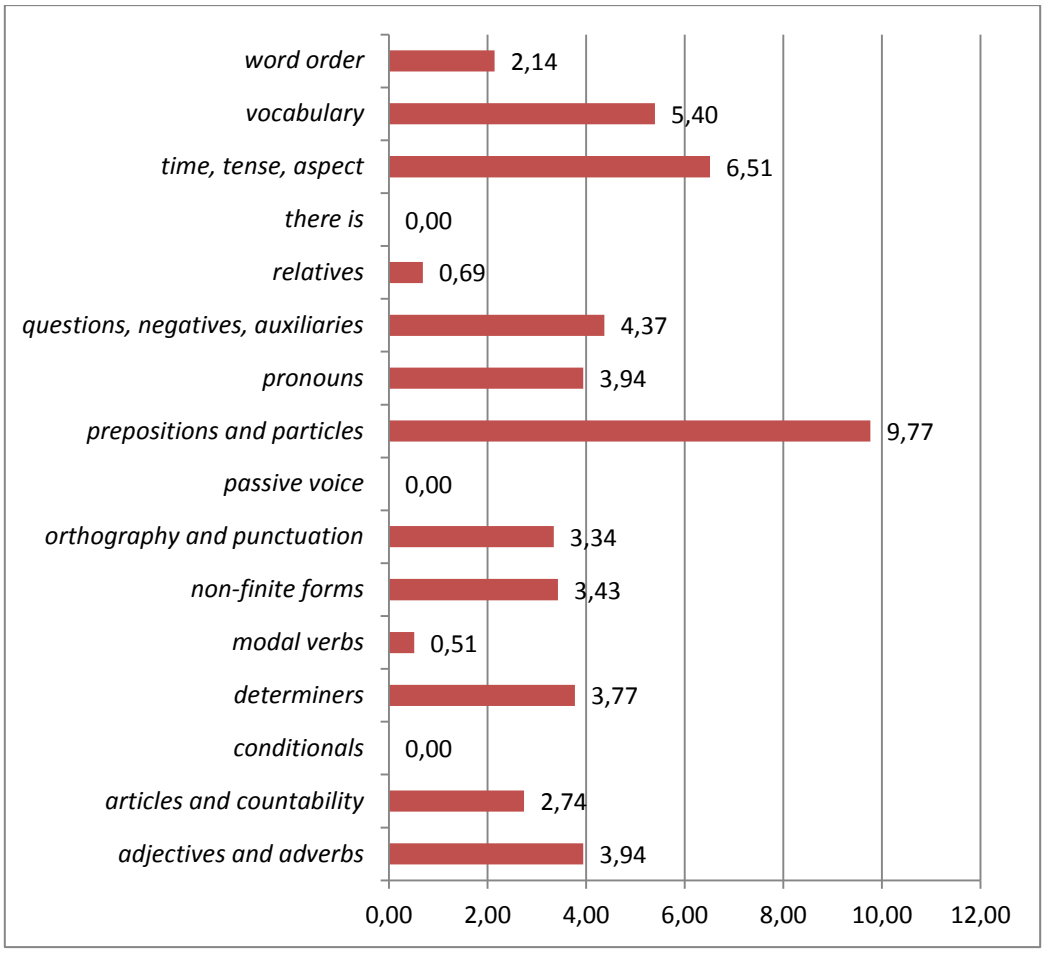


Figura 3.2.18: gráfico com os os erros mais cometidos pelos aprendizes no nível intermediário superior (valores normalizados).

Os valores no gráfico acima ao lado das barras indicam as ocorrências normalizadas por mil. O gráfico mostra que os erros mais comuns no nível intermediário superior segundo o sistema de classificação baseado em Shepherd (2001) são:

a) uso inadequado de preposições ou partículas (conjunções): 9,77‰ (114 ocorrências);

b) tempo e aspecto verbal: 6,51‰ (76 ocorrências);

c) escolha de palavras (*vocabulary*): 5,40‰ (63 ocorrências no córpus COBRA7_recorte);

d) formação de questões e o uso de negações e auxiliares: 4,37‰ (51 ocorrências).

3.2.5.2 Segundo o sistema de classificação SO2I, desenvolvido neste trabalho

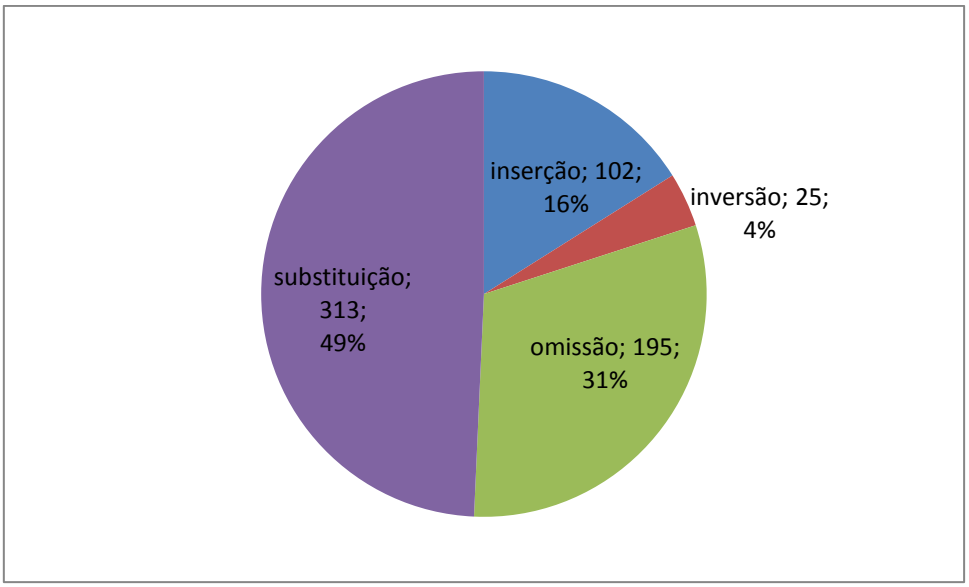


Figura 3.2.19: os erros mais comuns no nível intermediário superior segundo o sistema de classificação desenvolvido nesta pesquisa.

Os números indicam os valores absolutos das ocorrências, isto é, o número total de erros em cada categoria.

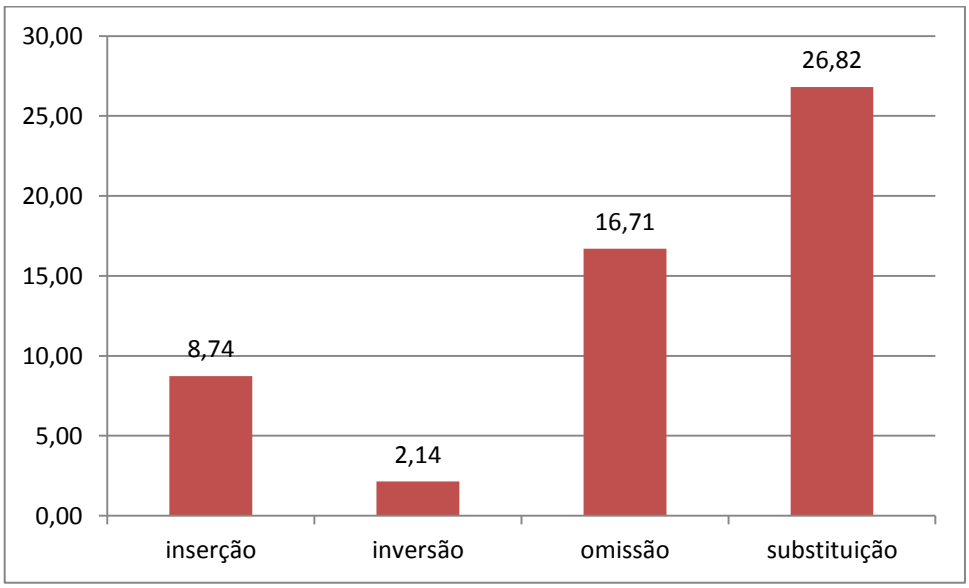


Figura 3.2.20: os erros mais comuns no nível intermediário superior segundo o sistema de classificação desenvolvido nesta pesquisa.

Os números no gráfico acima sobre as barras indicam a frequência normalizada por mil. O gráfico indica que os erros mais cometidos pelos aprendizes no nível de curso intermediário superior segundo o sistema de classificação desenvolvido nesta pesquisa são:

a) substituição: 26,82‰ (313 ocorrências);

b) omissão: 16,71‰ (195 ocorrências);

c) inserção: 8,74‰ (102 ocorrências);

d) inversão: 2,14‰ (25 ocorrências).

3.2.5.3 Comparação entre os sistemas de classificação usados nesta pesquisa

Baseado em Shepherd (2001):		**Sistema SO2I:**	
prepositions and particles	9,77‰	substituição	26,82‰
time, tense, aspect	6,51‰	omissão	16,71‰
vocabulary	5,40‰	inserção	8,74‰
questions, negatives, auxiliaries	4,37‰	inversão	2,14‰

Quadro 3.2.5: erros mais comuns no nível de curso intermediário superior segundo o sistema de classificação baseado em Shepherd (2001) e o SO2I, proposto nesta pesquisa.

O quadro indica que os aprendizes do nível intermediário superior cometem erros com o uso de preposições de partículas (conjunções); tempo e aspecto verbal; escolha de palavras; e formação de questões e uso de negação e auxiliares de uma das maneiras a seguir:

a) substituindo uma palavra por outra que pertença ou não à mesma classe gramatical;

b) omitindo palavras, prefixos ou sufixos;

c) inserindo palavras, prefixos ou sufixos;

d) invertendo palavras que compõem a colocação pretendida.

3.2.6 Nível de curso avançado

3.2.6.1 Segundo o sistema de classificação baseado em Shepherd (2001)

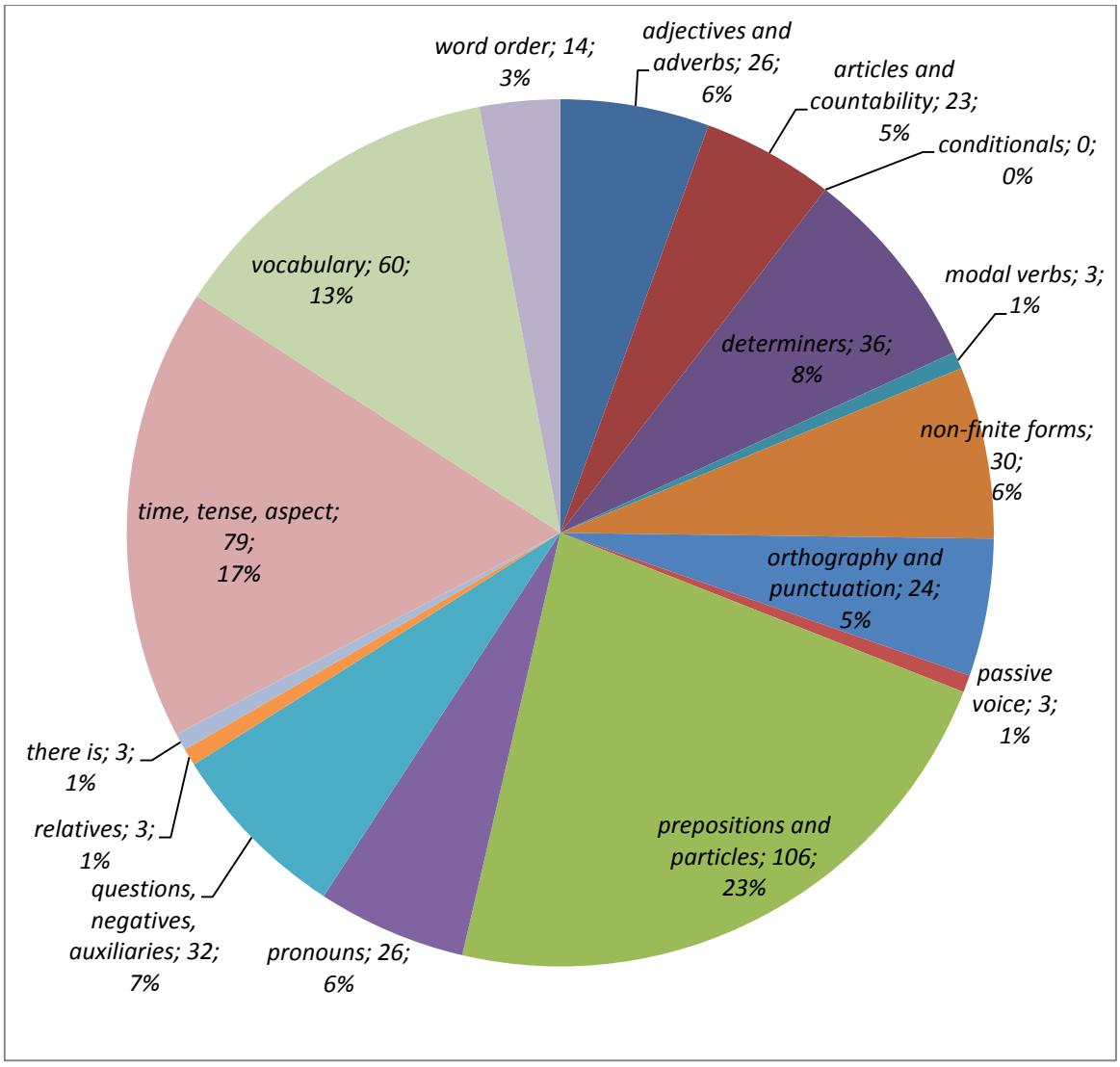


Figura 3.2.21: gráfico com os os erros mais cometidos pelos aprendizes no nível avançado.

Os valores numéricos no gráfico indicam a somatória bruta das ocorrências.

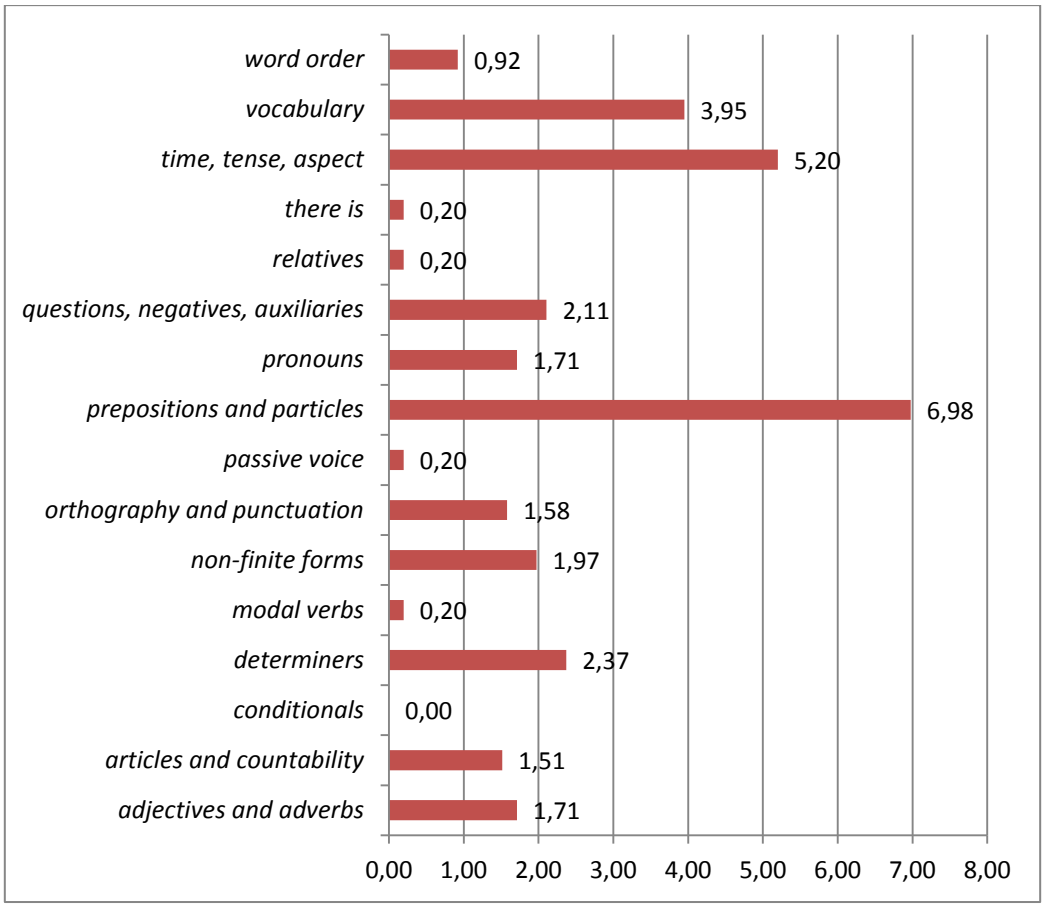


Figura 3.2.22: gráfico com os os erros mais cometidos pelos aprendizes no nível avançado (valores normalizados).

Os valores no gráfico acima ao lado das barras indicam as ocorrências normalizadas por mil. O gráfico mostra que os erros mais comuns no nível avançado segundo o sistema de classificação baseado em Shepherd (2001) são:

a) uso inadequado de preposições ou partículas (conjunções): 6,98‰ (106 ocorrências);

b) tempo e aspecto verbal: 5,20‰ (79 ocorrências);

c) escolha de palavras (*vocabulary*): 3,95‰ (60 ocorrências no córpus COBRA7_recorte);

d) uso inadequado de determinantes: 2,37‰ (36 ocorrências).

3.2.5.2 Segundo o sistema de classificação SO2I, desenvolvido neste trabalho

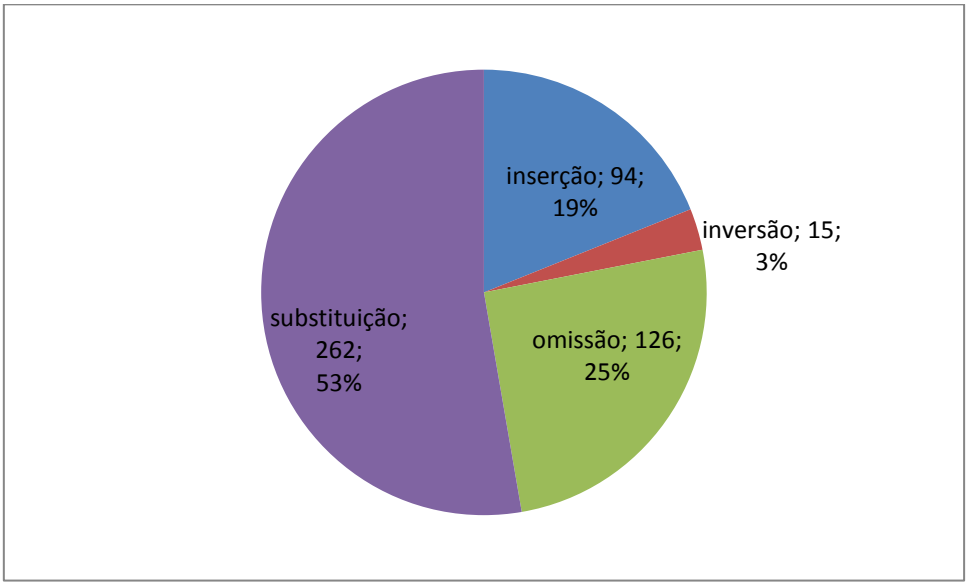


Figura 3.2.23: os erros mais comuns no nível avançado segundo o sistema de classificação desenvolvido nesta pesquisa.

Os números no gráfico acima indicam os valores absolutos das ocorrências, isto é, o número total de erros em cada categoria.

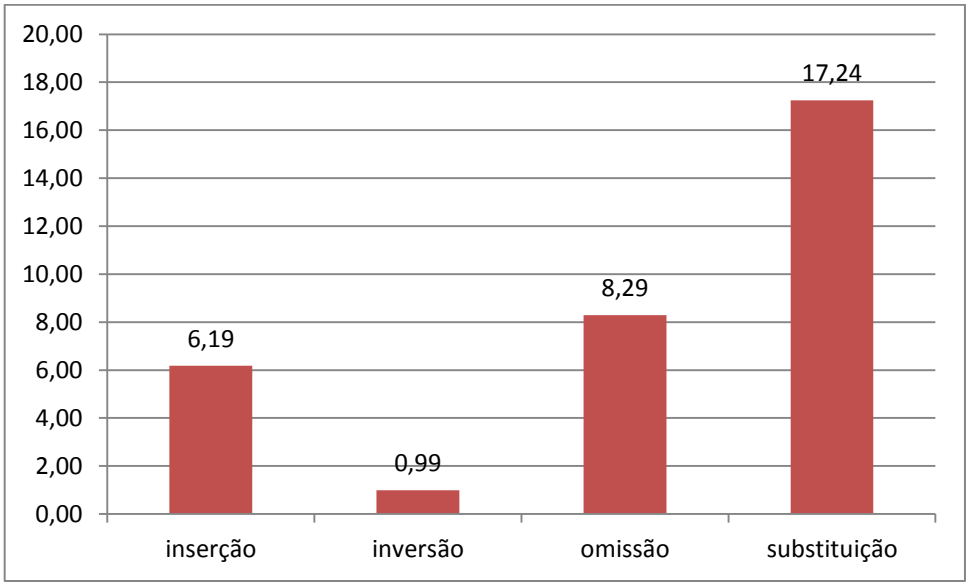


Figura 3.2.24: os erros mais comuns no nível avançado segundo o sistema de classificação desenvolvido nesta pesquisa (valores normalizados).

Os números no gráfico acima sobre as barras na figura acima indicam a frequência normalizada por mil. O gráfico mostra que os erros mais cometidos pelos aprendizes no nível de curso avançado segundo o sistema de classificação desenvolvido nesta pesquisa são:

a) substituição: 17,24‰ (262 ocorrências);

b) omissão: 8,29‰ (126 ocorrências);

c) inserção: 6,19‰ (94 ocorrências);

d) inversão: 0,99‰ (15 ocorrências).

3.2.5.3 Comparação entre os sistemas de classificação usados nesta pesquisa

Baseado em Shepherd (2001):		**Sistema SO2I:**	
prepositions and particles	6,98‰	substituição	17,24‰
time, tense, aspect	5,20‰	omissão	8,29‰
vocabulary	3,95‰	inserção	6,19‰
determiners	2,37‰	inversão	0,99‰

Quadro 3.2.6: quadro-resumo das ocorrências de erros mais comuns no nível de curso avançado segundo o sistema de classificação baseado em Shepherd (2001) e o SO2I.

O quadro indica que os aprendizes do nível avançado cometem erros com o uso de preposições de partículas (conjunções); tempo e aspecto verbal; escolha de palavras; e uso de determinantes de uma das maneiras a seguir:

a) substituindo uma palavra por outra que pertença ou não à mesma classe gramatical;

b) omitindo palavras, prefixos ou sufixos;

c) inserindo palavras, prefixos ou sufixos;

d) invertendo palavras que compõem a colocação pretendida.

3.2.7 Variação dos cinco erros mais comuns ao longo dos níveis de curso

3.2.7.1 Sistema de classificação baseado em Shepherd (2001)

A) Preposições e partículas

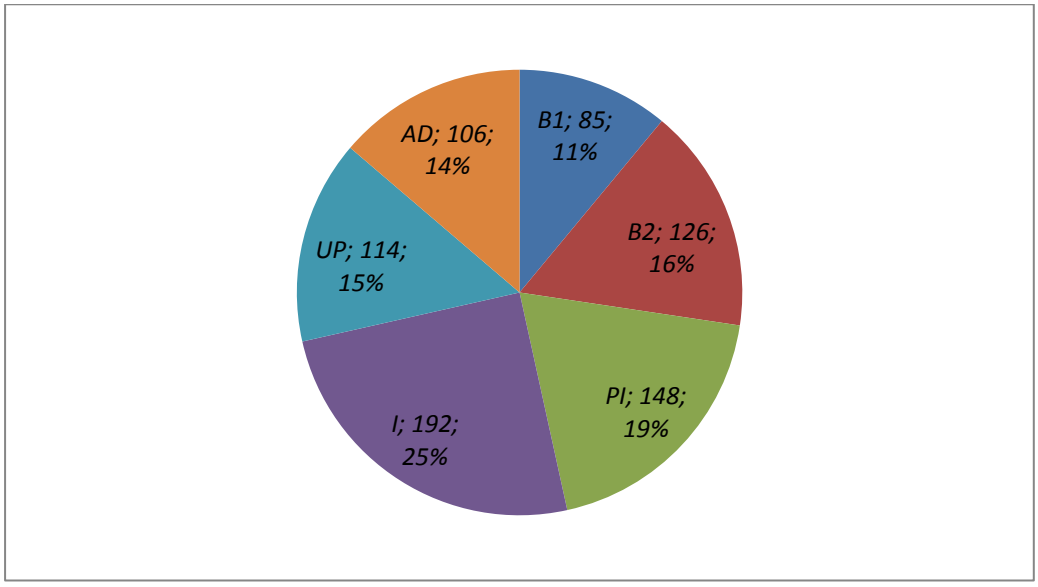


Figura 3.2.25: gráfico comparativo entre-níveis para o uso de preposições ou partículas[112].

Os valores na figura acima indicam a somatória bruta das ocorrências.

[112] As siglas na figura representam os níveis de curso básico 1, básico 2, pré-intermediário, intermediário, intermediários superior e avançado.

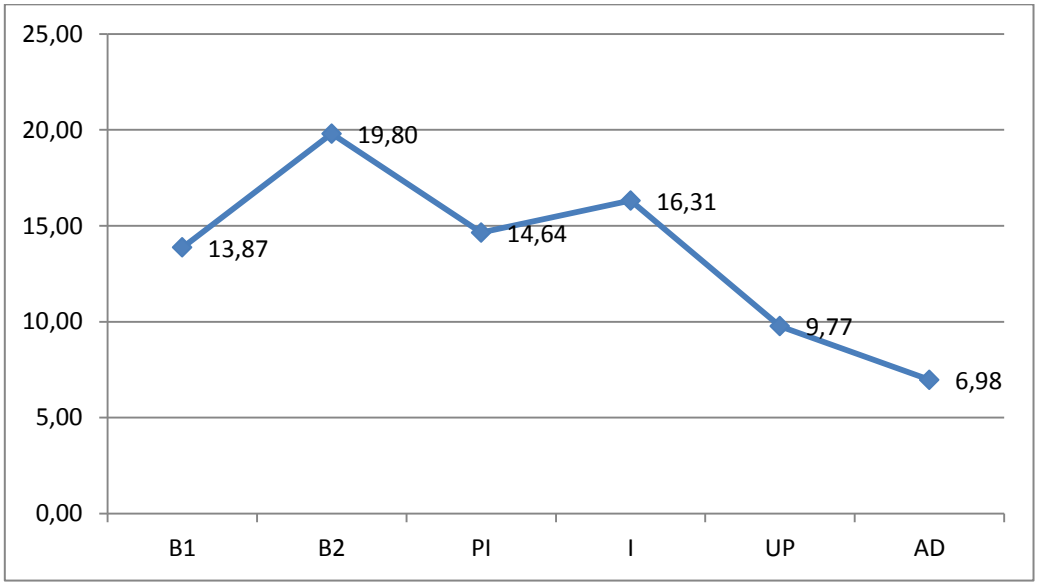


Figura 3.2.26: gráfico comparativo entre-níveis para o uso de preposições ou partículas[113] (valores normalizados).

Os valores numéricos indicam as ocorrências normalizadas por mil. O gráfico mostra que no nível básico 1 há 13,87 erros a cada mil palavras com relação ao uso de preposições. Esse valor cresce para 19,80‰ no nível seguinte e a partir do pré-intermediário começa a cair progressivamente (14,64‰, 16,31‰, 9,77‰, e 6,98‰). Isso mostra que no segundo nível, provavelmente pela necessidade do uso de colocações específicas como as usadas para se fornecer direções (*next to*, *far from* e outras), o aprendiz brasileiro comete mais erros no uso de preposições. Todavia, esse tipo de erro tende a diminuir ao longo dos estágios. Contudo, não está ainda totalmente solucionado no nível avançado.

A ocorrência de erros varia com relação ao gráfico anterior porque no primeiro os números representam as ocorrências absolutas. No gráfico acima, por sua vez, os valores foram normalizados por mil. Tal procedimento torna os valores individuais estatisticamente comparáveis, de modo que oferecem um resultado mais confiável cientificamente. Tal diferença poderá também ser observada nos próximos gráficos.

B) Vocabulário

[113] As siglas na parte inferior representam, respectivamente, os níveis de curso básico 1, básico 2, pré-intermediário, intermediário, intermediários superior e avançado.

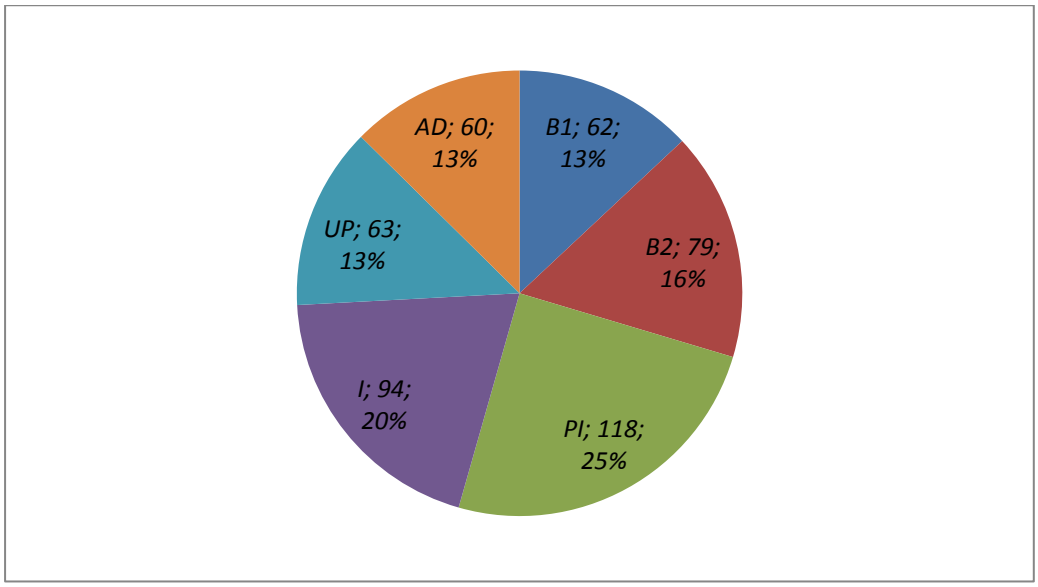


Figura 3.2.27: gráfico comparativo entre-níveis para o uso de escolha lexical[114].

Os valores acima indicam a somatória bruta das ocorrências por nível de curso.

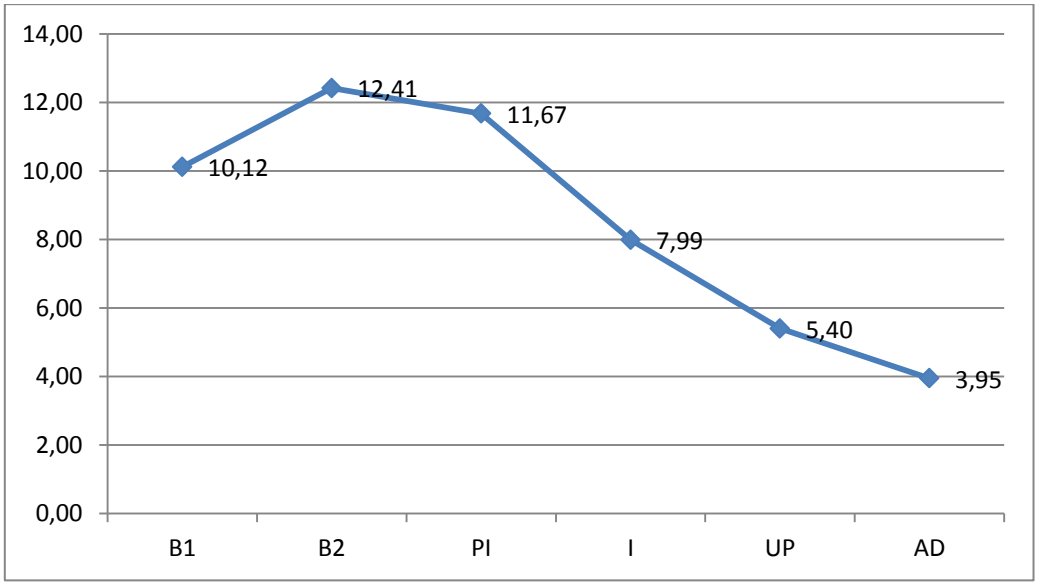


Figura 3.2.28: gráfico comparativo entre-níveis para o uso de escolha lexical (valores normalizados).

[114] As siglas na figura representam os níveis de curso básico 1, básico 2, pré-intermediário, intermediário, intermediários superior e avançado.

Os valores numéricos indicam as ocorrências normalizadas por mil. O gráfico mostra que no nível básico 1 há 10,12 erros a cada mil palavras com relação ao uso de preposições. Esse valor cresce para 12,41‰ no nível seguinte e a partir do pré-intermediário começa a cair progressivamente (11,67‰, 7,99‰, 5,40‰, e 3,95‰). Isso mostra que no segundo nível, provavelmente por ser uma transição entre o básico e o pré-intermediário, espera-se que o aprendiz brasileiro se comunique mais e, consequentemente, comete mais erros de escolha lexical. Todavia, embora haja uma redução de esse tipo de erro até o nível avançado, o problema não é totalmente sanado nesse nível.

C) Tempo e aspecto verbal

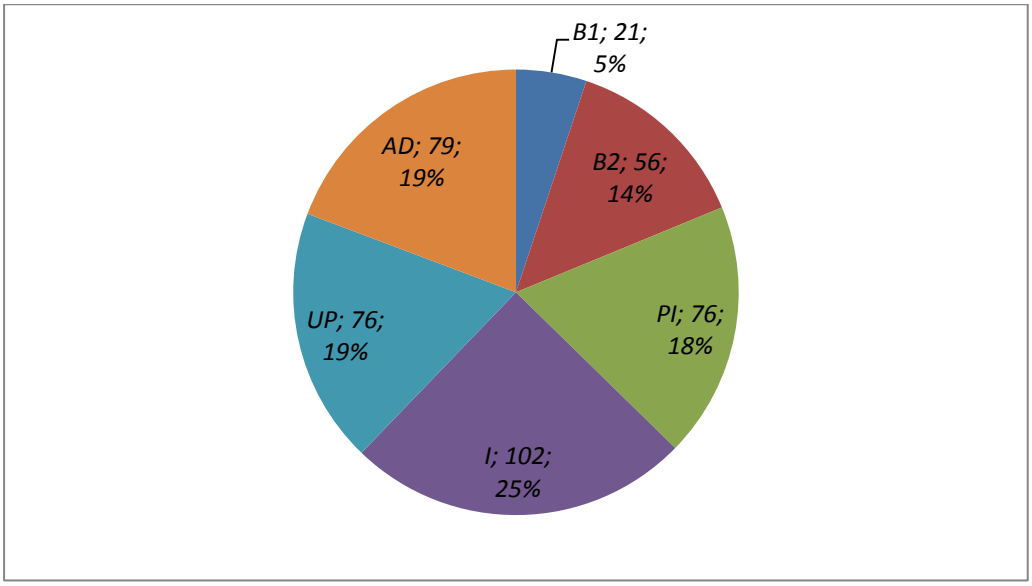


Figura 3.2.29: gráfico comparativo entre-níveis para o emprego de tempo e aspecto verbal[115].

Os valores numéricos indicam a somatória bruta das ocorrências por nível de curso.

[115] As siglas na figura representam os níveis de curso básico 1, básico 2, pré-intermediário, intermediário, intermediários superior e avançado.

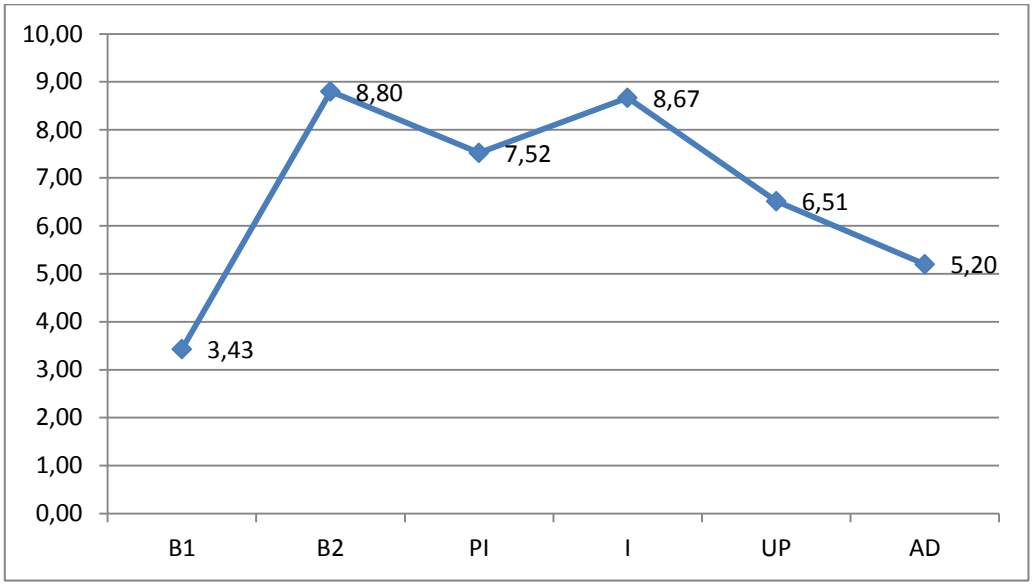


Figura 3.2.30: gráfico comparativo entre-níveis para o emprego de tempo e aspecto verbal (valores normalizados).

Os valores na figura indicam as ocorrências normalizadas por mil. Ao contrário da queda expressa ao longo dos níveis com relação aos erros citados anteriormente, com relação ao uso de tempos e aspectos verbais há um grande aumento no número de erros do básico 1 para o básico 2 (3,43‰ para 8,80‰). No nível pré-intermediário há uma pequena queda no número de ocorrências para esse tipo de erro (7,52‰), mas há um novo aumento no nível intermediário (8,67‰), seguido, novamente, de quedas progressivas nos níveis seguintes (6,51‰ e 5,20‰). Talvez o aumento desse tipo de ocorrência no nível intermediário esteja, novamente, atrelado à exposição do aprendiz a novos tempos verbais como é o caso do *present perfect*, do *used to* e de outros.

É importante notar, porém, que mesmo no nível avançado há mais erros relativos ao uso de tempo e aspecto verbal do que no nível básico 1. Digo isso porque é esperado que no nível avançado os estudantes tenham domínio da língua estrangeira aprendida.

D) Determinantes

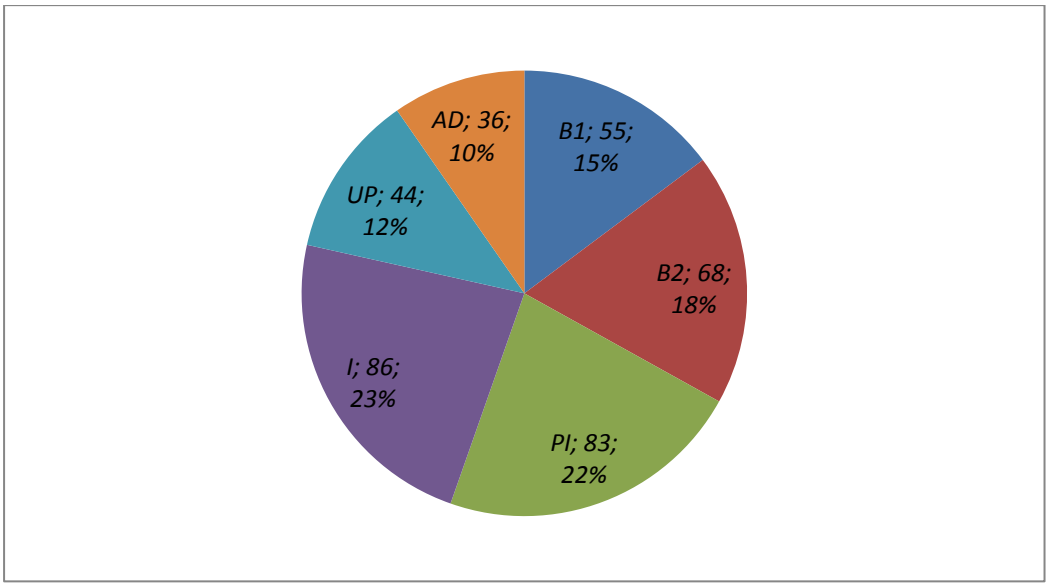


Figura 3.2.31: gráfico comparativo entre-níveis para o uso de determinantes[116].

Os valores na figura acima indicam a somatória bruta das ocorrências por nível de curso.

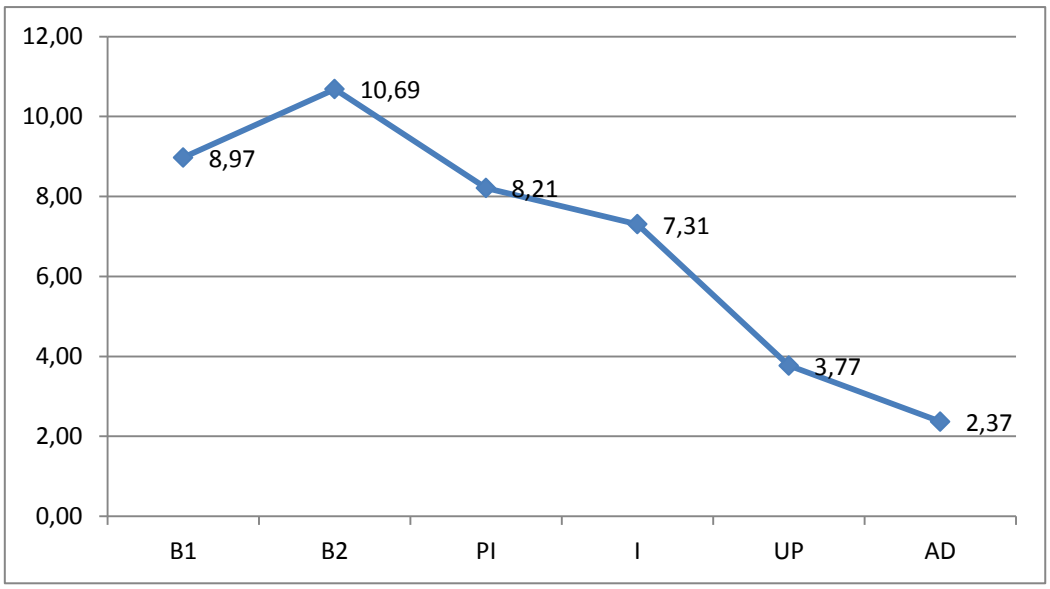


Figura 3.2. 32: gráfico comparativo entre-níveis para o uso de determinantes (valores normalizados).

[116] As siglas na figura representam os níveis de curso básico 1, básico 2, pré-intermediário, intermediário, intermediários superior e avançado.

Os valores numéricos no gráfico acima indicam as ocorrências normalizadas por mil. Com relação ao uso de determinantes, há, no básico 1, um número razoável de erros (8,97‰). No nível seguinte há um pequeno aumento (10,69‰). Nos níveis pré-intermediário e intermediário os números continuam semelhantes (8,21‰ e 7,31‰). Nos níveis seguintes o número de erros dessa natureza cai significativamente (3,77‰ e 2,37‰).

Os dados mostram que, nos primeiros níveis, os determinantes causam confusão para o aprendiz brasileiro e, por isso, talvez mereçam maior atenção por parte do professor de inglês.

E) Questões, negações e auxiliares

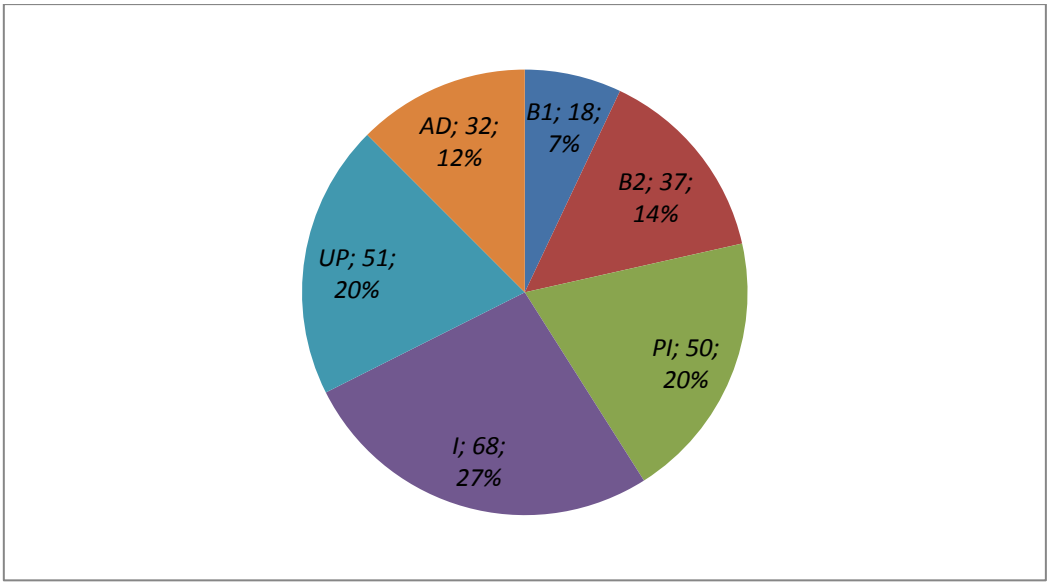


Figura 3.2.33: gráfico comparativo entre-níveis para o uso de questões, negações ou auxiliares[117].

Os números no gráfico indicam a somatória bruta das ocorrências por nível de curso.

[117] As siglas na figura representam os níveis de curso básico 1, básico 2, pré-intermediário, intermediário, intermediários superior e avançado.

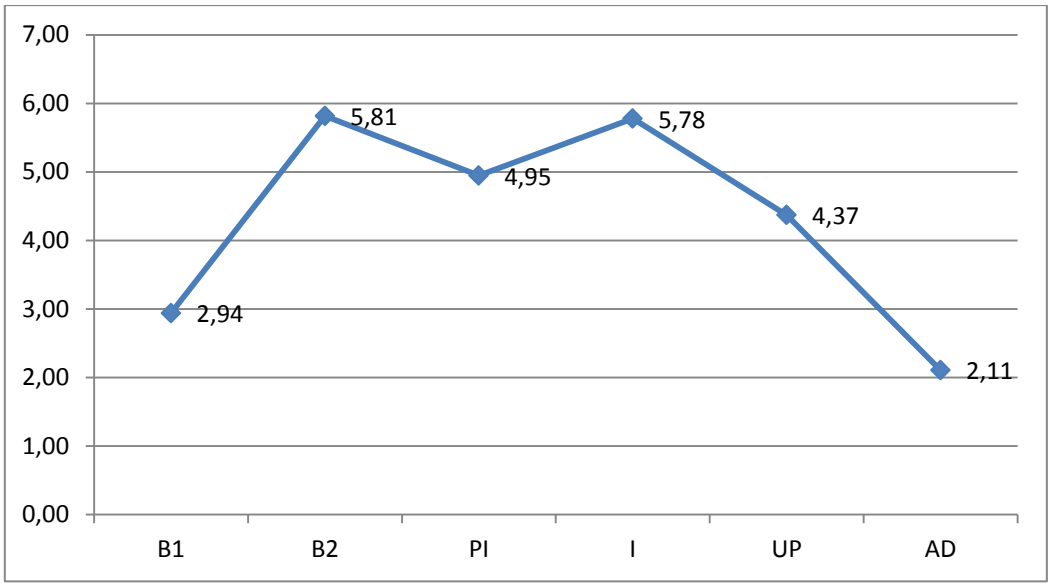


Figura 3.2.34: gráfico comparativo entre-níveis para o uso de questões, negações ou auxiliares (valores normalizados).

Os números na figura acima indicam as ocorrências normalizadas por mil. Os dados mostram que há certa diferença entre o número de erros no nível de curso básico 1 (2,94‰) e avançado (2,11‰). Entre esses dois níveis, todavia, há um sensível aumento no número desse tipo de erro, que praticamente se estabiliza.

Isso pode indicar que, até o nível avançado, o aprendiz brasileiro tem problemas em se familiarizar com o uso de verbos auxiliares diversos, bem como o uso de diversas palavras de negação como *no* e *not*. Essa informação sugere que podem-se criar alternativas para o ensino e a prática desses elementos visando a redução desse tipo de erro.

3.2.7.2 Sistema de classificação SO2I

Os cinco erros supracitados (uso de preposições e partículas, vocabulário, tempo e aspecto verbal, determinantes e questões, negações e auxiliares) apareceram nas redações de uma das seguintes maneiras (da mais para a menos frequente):

a) substituindo-se a palavra por uma que pertencia ou não à mesma classe gramatical;

b) omitindo-se a palavra quando era necessária;

c) inserindo-se a palavra quando não era necessária;

d) invertendo-se a ordem das palavras que compunham colocações.

Essa informação é relevante porque enquanto que o sistema de classificação baseado em Shepherd (2001) indica quais erros gramaticais aparecem nas redações de aprendizes que compõem o córpus COBRA-7_recorte, o sistema de classificação desenvolvido nesta pesquisa permite um foco mais colocacional, pois foca na substituição de palavras que compõem uma colocação, suas omissões, inserções desnecessárias e inversões, usos que ocasionam erros. Por isso, ao mesmo tempo em que a metodologia de identificação de erros proposta nesta pesquisa permite ao professor-pesquisador identificar os erros cometidos pelos seus aprendizes ao confrontá-los com um córpus de comparação, permite a ele ter uma ideia mais clara sobre os "passos" dados pelos aprendizes para gerar esses erros e, consequentemente, pode trazer uma luz para o desenvolvimento de técnicas e atividades que visem reduzi-los ou mesmo otimizar o aprendizado de idiomas como um todo.

A seguir mostrarei como os erros se apresentaram entre os níveis de curso segundo este sistema de classificação:

A) Substituição

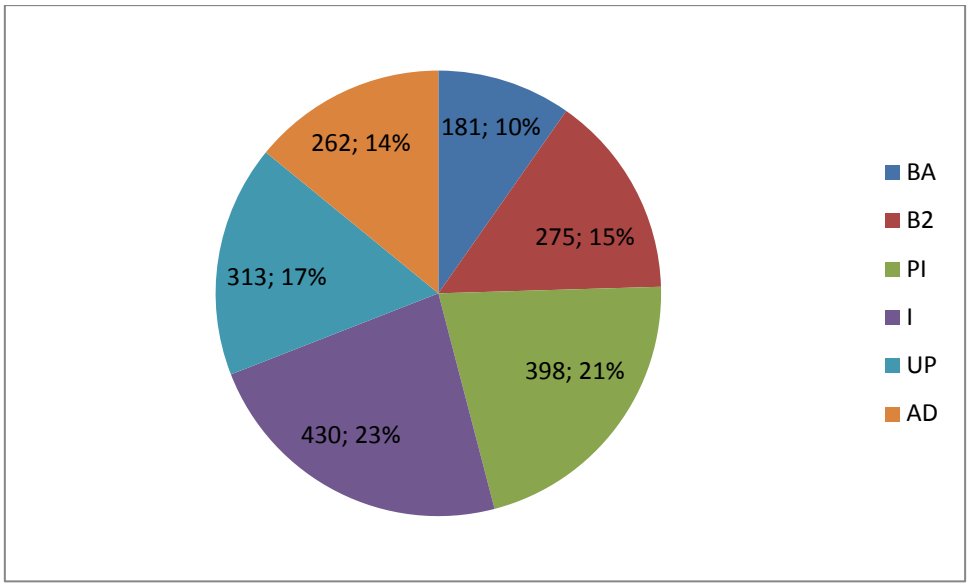


Figura 3.2.35: gráfico comparativo entre-níveis para substituição.

As siglas na figura representam os níveis de curso básico 1, básico 2, pré-intermediário, intermediário, intermediários superior e avançado.Os números no gráfico indicam a somatória bruta das ocorrências por nível de curso.

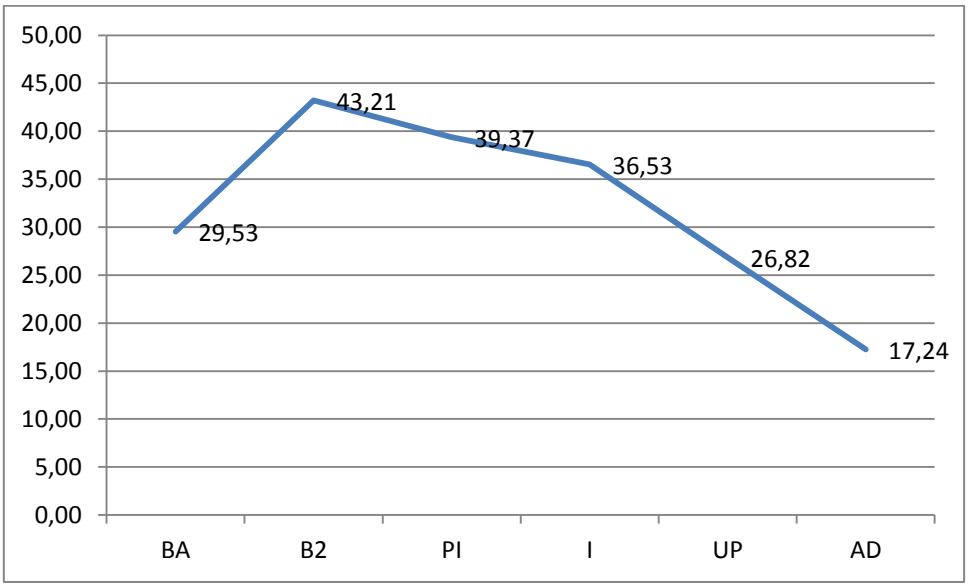


Figura 3.2.36: gráfico comparativo entre-níveis para substituição (valores normalizados).

Os valores numéricos no gráfico acima indicam as ocorrências normalizadas por mil.

De acordo com o gráfico acima, há um aumento sensível de erros de substituição do nível de curso básico 1 (29,53‰) para o básico 2 (43,21‰), a partir do qual há um progressivo declínio de ocorrências desse tipo (39,37‰, 36,53‰, 26,82‰, e 17,24‰). Todavia, tal ocorrência se faz ainda presente no nível de curso avançado.

B) Inserção

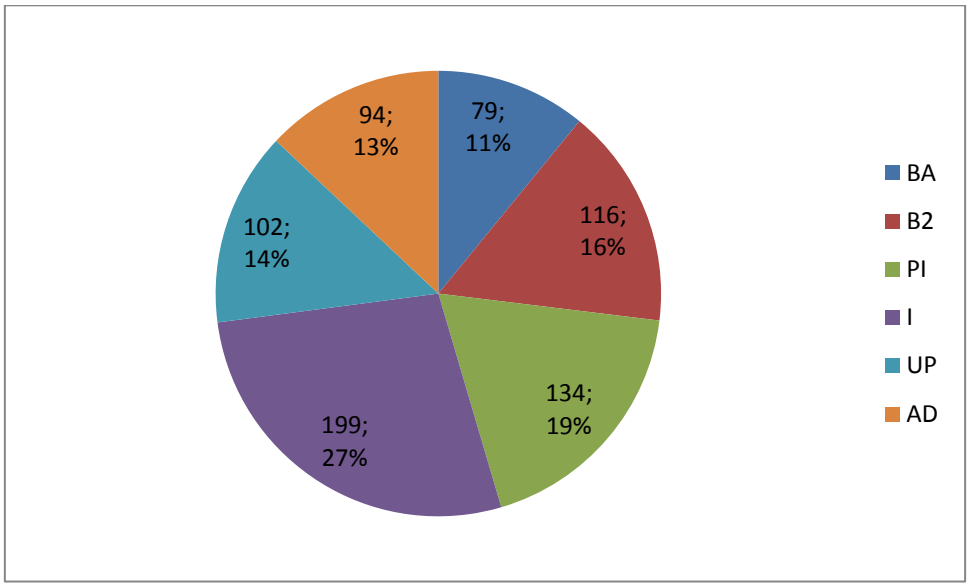


Figura 3.2.37: gráfico comparativo entre-níveis para inserção.

As siglas na figura representam os níveis de curso básico 1, básico 2, pré-intermediário, intermediário, intermediários superior e avançado.Os números no gráfico indicam a somatória bruta das ocorrências por nível de curso.

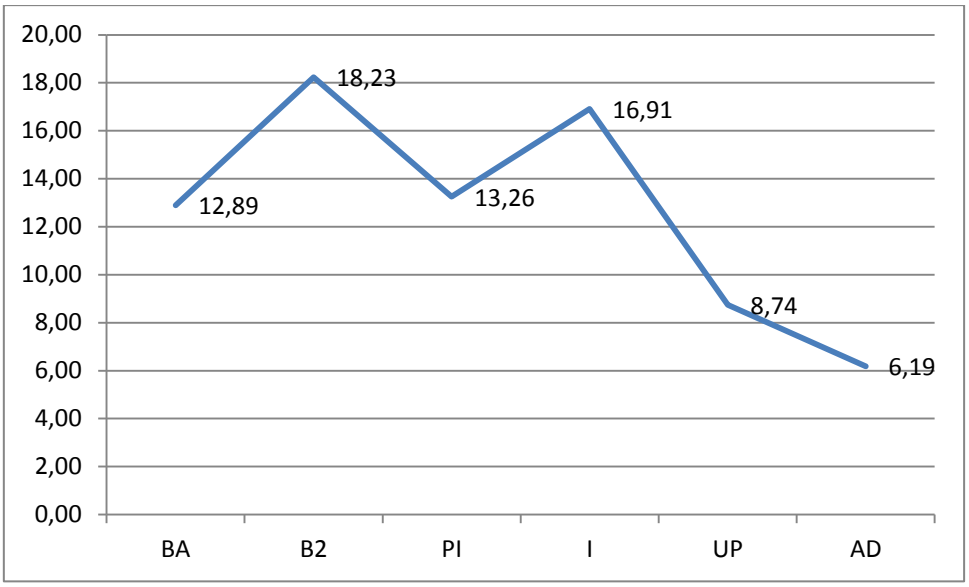


Figura 3.2.38: gráfico comparativo entre-níveis para inserção (valores normalizados).

Os valores numéricos no gráfico acima indicam as ocorrências normalizadas por mil.

Os resultados mostram um aumento sensível de erros de inserção do nível de curso básico 1 (12,89‰) para o básico 2 (18,23‰), uma queda no nível seguinte (pré-intermediário) (13,26‰), seguido de um novo aumento próximo ao do nível anterior (16,91‰) e, a partir de então, um progressivo declínio de ocorrências desse tipo. Todavia, tal ocorrência se faz ainda presente no nível de curso avançado (6,19‰).

C) Omissão

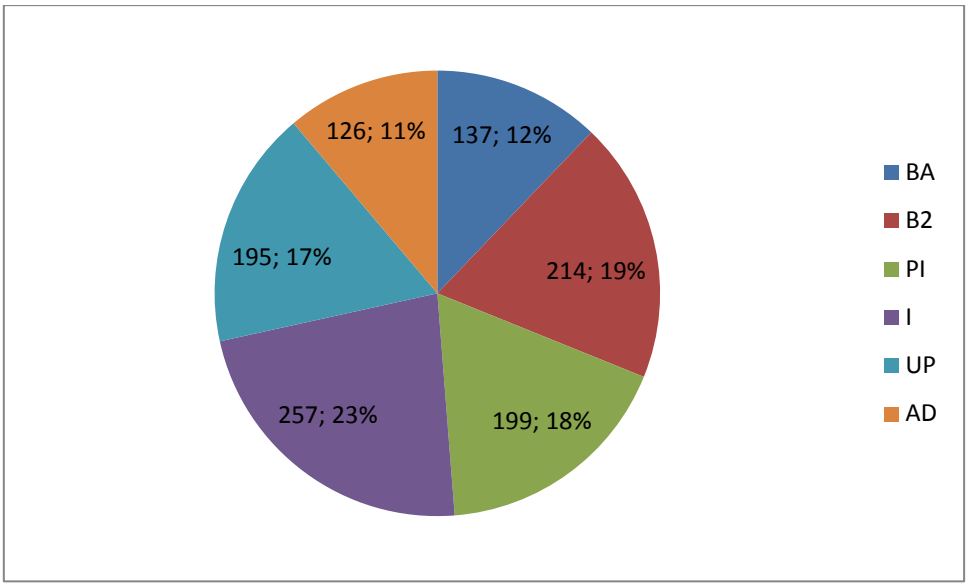


Figura 3.2.39: gráfico comparativo entre-níveis para omissão.

As siglas na figura representam os níveis de curso básico 1, básico 2, pré-intermediário, intermediário, intermediários superior e avançado.Os números no gráfico indicam a somatória bruta das ocorrências por nível de curso.

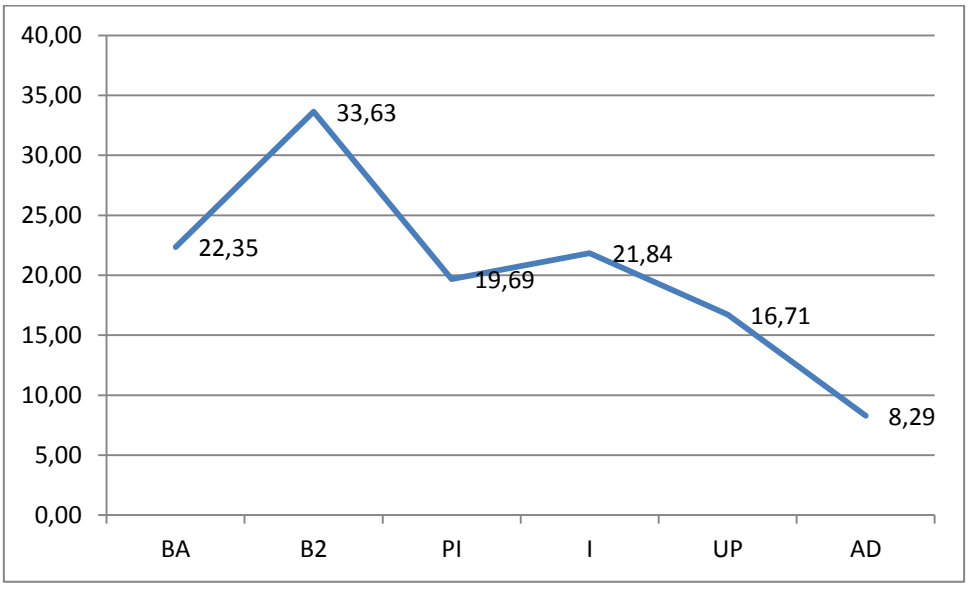


Figura 3.2.40: gráfico comparativo entre-níveis para omissão (valores normalizados).

Os valores numéricos no gráfico acima indicam as ocorrências normalizadas por mil.

Os resultados mostram um aumento sensível de erros de omissão do nível de curso básico 1 (22,35‰) para o básico 2 (33,63‰), uma queda no nível seguinte (pré-intermediário) (19,69‰), seguido de um pequeno aumento próximo ao do nível anterior (21,84‰) e, a partir de então, um progressivo declínio de ocorrências desse tipo. Todavia, tal ocorrência se faz ainda presente no nível de curso avançado (8,29‰).

D) Inversão

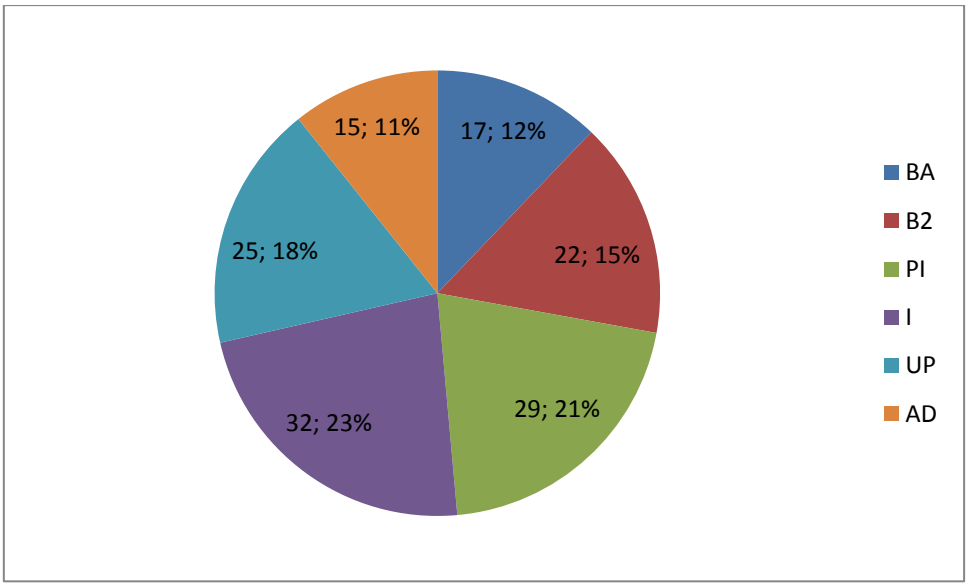


Figura 3.2.41: gráfico comparativo entre-níveis para inversão.

As siglas na figura representam os níveis de curso básico 1, básico 2, pré-intermediário, intermediário, intermediários superior e avançado.Os números no gráfico indicam a somatória bruta das ocorrências por nível de curso.

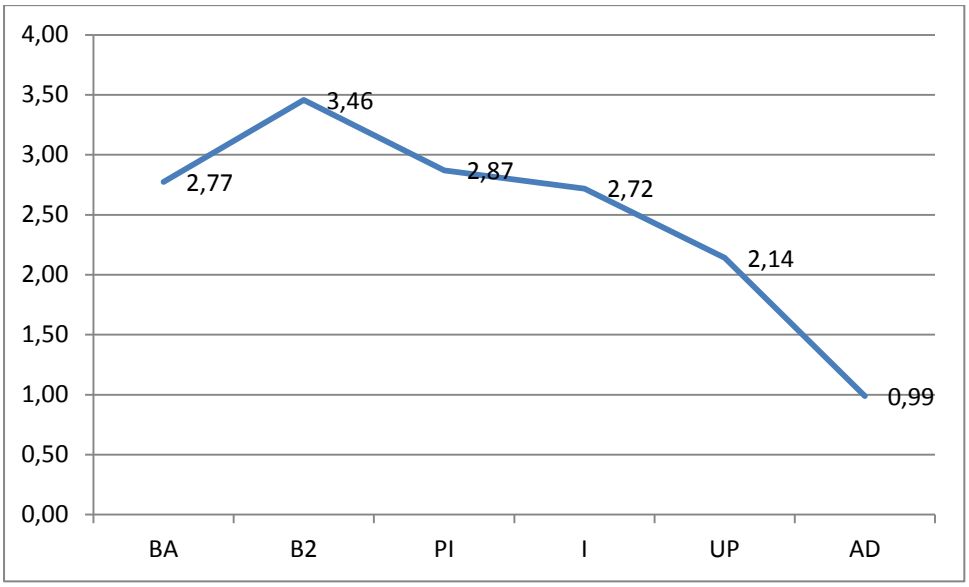


Figura 3.2.42: gráfico comparativo entre-níveis para inversão (valores normalizados).

Os valores numéricos no gráfico acima indicam as ocorrências normalizadas por mil.

Os resultados mostram um aumento sensível de erros de inversão do nível de curso básico 1 (2,77‰) para o básico 2 (3,46‰), uma queda progressiva nos níveis seguinte (2,87‰, 2,72‰, 2,14‰, 099‰). Todavia, tal ocorrência se faz ainda presente no nível de curso avançado.

3.3 Questão 3: Qual nível de curso apresenta maior diversidade de erros no córpus COBRA-7_recorte?

Nesta pesquisa trabalhei com 17 tipos de erro, somando-se os 13 baseados no modelo de Shepherd (2001) e os 4 desenvolvidos nesta pesquisa (sistema SO2I).

A seguir faço uma verificação de quantos desses tipos de erro aparecem em cada um dos níveis de curso estudados nesta pesquisa.

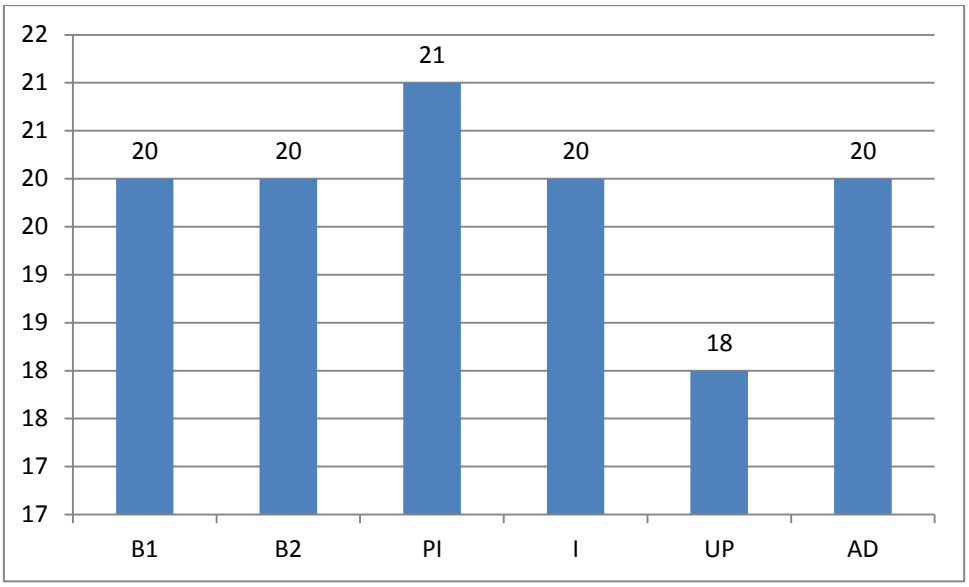


Figura 3.3.1: qual nível de curso apresenta a maior diversidade de erros segundo os sistemas de classificação baseado em Shepherd (2001) e SO2I, desenvolvido nesta pesquisa[118].

O gráfico mostra que o nível de curso com a maior diversidade de erros foi o pré-intermediário. Esse nível foi o único que apresentou erros de estruturação lógica de condicionais (*if clauses*). Ao mesmo tempo que isso já era esperado, pois trata-se exatamente do nível de curso no qual as *if clauses* são estudadas, indaguei-me sobre haver, nos níveis superiores, relutância dos aprendizes em usar esse tipo de formação. Uma busca nas linhas de concordância do córpus COBRA-7_recorte (arquivo "Controle_corpus_v3.xls", disponível no CD-rom anexo) com o termo de busca "*if*" revelou que os aprendizes utilizaram os condicionais, porém de forma correta.

Mesmo assim, vale dizer que, no nível pré-intermediário foi encontrado um único erro de uso de condicional, o que pode sugerir insegurança do aprendiz brasileiro para usar esse tipo de construção. Essa suposição poderá ser melhor investigada em estudos posteriores desenvolvidos pelos próprios professores ou linguistas de córpus.

3.4 Síntese dos achados

Nesta seção apresento uma síntese dos achados. Para isso usarei uma tabela contendo os sinais matemáticos de soma, subtração e igualdade. Tais sinais indicam se uma determinada ocorrência de erro, respectivamente, aumentou, se manteve, ou foi reduzida. O cálculo foi feito

[118] Os números referem-se à somatória bruta categorias nas quais encontraram-se erros em cada um dos níveis de curso. O número total de categorias é 17.

dividindo-se a ocorrência de uma categoria de erro em um determinado nível pela do nível imediatamente anterior. No caso de erros com o uso de preposições ou partículas, por exemplo (ver tabela a seguir), dividi a ocorrência normalizada dos erros no nível de curso básico 2 (19,80) pela do básico 1 (13,87). O resultado (1,42) indica que houve aumento nesse tipo de erro do primeiro nível supracitado para o segundo. A pontuação para o uso desses sinais matemáticos seguiu a seguinte lógica:

A) se o resultado da divisão do nível y pelo nível x com relação à ocorrência do erro 1 for menor ou igual a 0,5, então haverá dois sinais de subtração (“- -”), que indicam que a ocorrência caiu pela metade;

B) se o resultado da divisão do nível y pelo nível x com relação à ocorrência do erro 1 for menor ou igual a 0,75, então o sinal será de subtração (“-”), indicando redução;

C) se o resultado da divisão do nível y pelo nível x com relação à ocorrência do erro 1 for menor ou igual a 1,25, então o sinal será de igualdade (“=”), indicando que a ocorrência se manteve considerando-se o nível anterior;

D) se o resultado da divisão do nível y pelo nível x com relação à ocorrência do erro 1 for menor ou igual a 1,75, então o sinal será de soma (“+”), indicando que houve aumento na ocorrência de erros do nível imediatamente anterior para o posterior;

E) se o resultado da divisão do nível y pelo nível x com relação à ocorrência do erro 1 for maior que 1,75, então usarei um sinal duplo de soma (“++”), indicando que o número de ocorrências de erros do nível imediatamente anterior para o posterior dobrou;

Erro	B1	B2	PI	I	UP	AD
Preposições ou partículas	13.87	+	-	=	-	-
		19.80	14.64	16.31	9,77	6,98
Vocabulário	10.12	=	=	-	-	-
		12.41	11.67	7.99	5.40	3.95
Tempo e aspecto verbal	3.43	++	=	=	=	=
		8.80	7.52	8.67	6.51	5.20
Determinantes	8.97	=	=	=	-	-
		10.69	8.21	7.31	3.77	2.37
Questões, negacões ou auxiliares	2.94	++	=	=	=	- -
		5.81	4.95	5.78	4.37	2.11
Substituição	29.53	+	=	=	-	-
		43.21	39.37	36.53	26.82	17.24
Inserção	12.89	+	-	+	-	-
		18.23	13.26	16.91	8.74	6.19
Omissão	22.35	+	-	=	=	- -
		33.63	19.69	21.84	16.71	8.29
Inversão	2.77	=	=	=	=	- -
		3.46	2.87	2.72	2.14	0.99

Quadro 3.4.1: frequência normalizada de erros em cada nível e diferenca entre níveis adjacentes.

Como se percebe, houve uma diminuição constante (ou seja, três ou mais colunas de um mesmo tipo de erro com sinal de -) em relação aos erros com o uso de preposições ou partículas, escolha lexical (vocabulário), e inserção. Houve manutenção (ou seja, três ou mais colunas de um mesmo tipo de erro com sinal de =) em relacão aos com o uso de tempo ou aspecto verbal, determinantes, questões, negacões ou auxiliares, e inversão. Por outro lado, houve oscilação com

relação aos erros de substituição e omissão. Isso significa que não houve um padrão como nos outros casos mas, de modo geral, um aumento seguido de manutenção e uma redução.

Neste capítulo foram apresentados e interpretados os resultados obtidos para cada uma das questões de pesquisa que nortearam o trabalho. O capítulo a seguir apresentará uma discussão dos resultados e o fechamento do trabalho.

Considerações Finais

O presente capítulo faz um fechamento do trabalho, retomando seus pontos principais, apontando limitações, e fazendo sugestões de pesquisa futura e possíveis aplicações pedagógicas dos resultados.

Conforme dito na Introdução, na área de ensino de idiomas os professores se deparam, diariamente, com erros nas produções escritas dos seus aprendizes brasileiros. Todavia, os professores em geral tendem a encarar esses negativamente, sem considerar que eles podem não somente se repetir com outros aprendizes mas também ser a expressão de uma característica linguística comum aos aprendizes brasileiros.

A pesquisa descrita aqui teve como objetivo justamente enfocar essa problemática ao identificar e classificar os erros na escrita de aprendizes brasileiros de inglês e, como desdobramentos possíveis, prover aos professores e pesquisadores um sistema de identificação e classificação de erros, com vistas a auxiliá-los em seu trabalho e informar a produção de materiais didáticos locais, isto é, voltados a aprendizes brasileiros.

As questões de pesquisa investigadas foram:

1- Quais os erros mais comuns no córpus COBRA-7_recorte?

2- Qual a variação de erro entre os níveis de curso dos aprendizes no córpus COBRA-7_recorte?

3- Qual nível de curso apresenta maior diversidade de erros no córpus COBRA-7_recorte?

Para responder a essas questões a pesquisa encontrou suporte teórico na Linguística de Córpus, área que "se coupa da coleta e exploração de corpora, ou conjuntos de dados linguísticos textuais, em formato legível por computador, que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística" (Berber Sardinha, 2000b, p. 3).

A metodologia consistiu na criação do córpus COBRA-7 a partir de redações de aprendizes brasileiros de inglês como língua estrangeira matriculados em uma rede de escolas de idiomas e armazenadas em um servidor *online*. Como a anotação dos erros encontrados nesse córpus seria manual e devido ao tempo restrito do mestrado, optei por um recorte de 50 redações por cada nível

de curso, totalizando 300 redações. Essa amostra foi chamada de COBRA7_recorte e constituiu o córpus de análise deste trabalho.

Das observações das 300 redações foram extraídos 3.854 erros, identificados por meio do método desenvolvido nesta pesquisa, que consistiu nos seguintes passos: 1. leitura da redação; 2. identificação da suspeita de erro; 3. depreensão da coligação na qual o erro está inserido. Neste passo, manter a própria palavra central do erro (nódulo) e a(s) preposições que o acompanha(m) ou antecede(m); 4. busca da coligação no sítio do COCA utilizando o nódulo e duas palavras à esquerda e à direita dele com o intuito de verificar se a colocação é usada pelos falantes nativos; 5. confirmação ou não de que se trata de um erro.

Quando um erro foi identificado por meio do método acima, procedi à sua classificação. Para isso utilizei um sistema baseado na proposta de Shepherd (2001) (sistema 1), que os agrupa em categorias gramaticais. Como tal sistema possuía muitas categorias e, consequentemente, dificuldade de distinção entre elas, optei pela criação de um sistema mais enxuto e eficaz (sistema de classificação SO2I, ou sistema 2, composto por quatro categorias: substituição, omissão, inserção e inversão), conforme comprovado pelo teste estatístico Kappa de concordância entre avaliadores.

Os resultados mostraram que os erros mais comuns no córpus COBRA-7_recorte foram (do maior para o menor): uso de preposições ou conjunções, escolha lexical, uso de tempo e aspecto verbal, e uso de determinantes (sistema 1). Tais erros se apresentaram de uma das quatro formas a seguir: substituição, omissão, inserção, ou inversão (sistema 2).

Mostraram também que de modo geral os erros apresentam um crescimento em número e diversidade no nível pré-intermediário e tendem a cair ao longo dos níveis de curso. Tal característica pode sugerir aos professores e pesquisadores que o pré-intermediário é o nível que requer maior atenção ao desempenho dos aprendizes e maior sensibilidade à resposta que apresentam aos temas e aspectos léxico-gramaticais estudados. Tal dificuldade pode ser justificada considerando que se trata de um nível de curso tranasitório ente o básico (que apresenta estrutura e léxico elementares) e o intermediário (que apresenta tempos e aspects verbais e vocabulário diversos e exige maior capacidade argumentativa). Com relação aos erros no uso de tempo e aspecto verbal, este estudo revelou que tendem a oscilar entre 3,43‰ e 8,80‰ (números normalizados) e também não se solucionam ao término do nível de curso avançado (5,20‰), o que pode indicar que o enfoque lexical da escola de idiomas de onde procederam as redações que originaram o córpus COBRA-7_recorte pode ter negligenciado a gramática. Portanto, este estudo reafirma a importância de léxico e gramática no ensino de idiomas.

Com relação ao desenvolvimento do método de identificação e classificação de erros, a pesquisa mostrou que o sistema de classificação SO2I pode ser usado de modo independente, não necessitando, assim, de um sistema paralelo como o baseado em Shepherd (2001), isso porque se trata de um sistema baseado na mecânica textual que independe de nomenclaturas e classes gramaticais que podem ser de difícil compreensão para os aprendizes, como a experiência docente tem mostrado. Por isso o SO2I parece ser um sistema eficiente na correção de redações, uma vez que o professor informa ao aprendiz em que parte da sua redação há uma inversão de termos, uma substituição, uma omissão, ou uma inserção desnecessária, e o próprio aprendiz corrige seu erro por meio de processo reverso.

A relevância destes achados refere-se ao fato de que o foco lexical de muitas escolas de inglês hoje parece ter negligenciado a gramática. Além disso, o método de identificação e classificação de erros aqui desenvolvido pode facilitar o trabalho do professor na correção das redações dos seus aprendizes e desenvolver neles, devido à repetição e exposição, a percepção de reconhecimento dos próprios erros.

O presente trabalho possui limitações. A primeira delas é o fato de o córpus de estudo COBRA-7 não pôde ser analisado globalmente por ser inviável em termos de tempo, como justificado anteriormente. Se o córpus tivese sido analisado inteiramente, talvez os resultados apresentassem alguma alteração, na medida em que analisei, no córpus COBRA-7_recorte, aproximadamente 10% de todo o córpus compilado. A segunda limitação diz respeito ao próprio recorte utilizado. Talvez outras escolhas e outras metodologias poderiam ter revelado outras necessidades para os aprendizes, mas pesquisas futuras poderão preencher esta lacuna.

Há vários aspectos que podem ser explorados em pesquisa futura. O primeiro deles diz respeito à reprodução do método de identificação e classificação de erros aqui sugerida para que a comunidade científica possa ter uma ideia mais clara de quanto os resultados aqui revelados refletem o perfil do aprendiz brasileiro de inglês como língua estrangeira. O segundo aspecto diz respeito a um possível levantamento lexical por nível de curso e/ou por gênero. Por fim, os erros apontados nesta pesquisa podem ser usados em programas computacionais que tenham por objetivo a correção automática de erros de uso de língua.

Pode-se pensar em algumas possíveis aplicações pedagógicas para os resultados desta pesquisa. A primeira delas pode ser o uso do sistema SO2I nas salas de aula de inglês como língua estrangeira. A segunda possível aplicação pode ser o desenvolvimento de materiais didáticos que procurem reduzir os erros aqui apresentados em cada nível de curso. Por fim, uma terceira aplicação possível poderia ser uma conscientização constante do professor de que léxico e gramática são inseparáveis, mas o foco em um pode causar deficiência no outro.

O trabalho aqui apresentado pretende ter apresentado uma contribuição para a Linguística de Córpus na medida em que utilizou seus pressupostos teóricos tanto para a compilação do córpus quanto para a análise. Além disso, ao ter desenvolvido um córpus médio formado por composições feitas por aprendizes brasileiros de inglês como língua estrangeira, o trabalho pretende ter contribuído especificamente para as pesquisas com córpora de aprendizes. Por fim, esta pesquisa, ao enfocar as características específicas das composições dos aprendizes brasileiros cujas redações compuseram o córpus COBRA-7_recorte, espera ter preenchido uma lacuna importante na literatura da análise de erros de aprendizes de inglês como língua estrangeira.

É importante que os professores de inglês se tornem cientes da missão que possuem enquanto pesquisadores e, portanto, investigadores das produções de seus próprios aprendizes e, a partir disso, desenvolvam técnicas que gerem maior autonomia nos aprendizes e levem-nos a identificar e corrigir os seus próprios erros.

Referências bibliográficas

BAKER, P., HARDIE, A., MCENERY, T. (2006). *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

BALBÁS, M. S. (2003). Análise de erros, baseada na Linguística de Corpus, da escrita de aprendizes brasileiros universitários de espanhol como língua estrangeira. Dissertação de mestrado. LAEL. Pontifícia Universidade Católica, São Paulo, Brasil.

BENNET, G. R. (2010). Principles of Corpus Linguistics. In: *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. Michigan ELT. Disponível na Internet no endereço <http://www.press.umich.edu/titleDetailDesc.do?id=371534>. Acesso em 15 fev. 2012.

BERBER SARDINHA, T. (1998). Size of a representative corpus. (Resumo de discussão sobre córpora, lista de discussão). 26/08. Disponível na Internet no site <http://www.hd.uib.no/corpora/1998-3/0107.html>. Acesso em 18 ago. 2011.

_____. (2000) Computador, corpus e concordância no ensino de léxico-gramática de língua estrangeira. In: LEFFA, V. (Ed.) *As palavras e sua companhia: o léxico na aprendizagem*. pp 45-72. Pelotas: EDUCAT, Universidade Católica de Pelotas. Disponível na Internet no endereço <http://www2.lael.pucsp.br/~tony/temp/publications/2000vilson.pdf>. Acesso em 18 ago. 2011.

_____. (2000b). Linguística de Corpus: Histórico e Problemática. *D.E.L.T.A.* São Paulo, Vol. 16, nº 2, pp. 323-367.

_____. (2004). *Linguística de Corpus*. Barueri: Manole.

_____. (2011). *Linguística de Corpus*. Barueri: Manole.

BERBER SARDINHA, T., SHEPHERD, T. (2011). *Corpus Linguistics and writing assessment*. Apresentação convidada na *Roundtable on "Corpus linguistics for 21st century language learning"*. *AESLA Conference 2011*, Salamanca, Espanha.

BERGH, G. & ZANCHETTA, E. (2008). Web linguistics. In: Lüdeling, A. & Kytö, M. (Ed.). *Corpus Linguistics: an International Handbook*. (pp. 309-327). Walter de Gruyter: Berlin/New York.

BIBER, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*. Oxford: Orford University Press, v.8, n. 4, outubro, pp. 243-57.

_____. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

BIBER, D., CONRAD, D. & REPPEN, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

BIBER, D. *et al*. (1999). *Longman grammar of spoken and written English*. London: Longman.

BOLINGER, D. (1976). Meaning and Memory. *Forum Linguisticum 1.1*. (pp. 1-14). Harvard: Harvard University.

BROWN, H. D. (2007). *Teaching by Principles – An Interactive Approach to Language Pedagogy*. New York: Pearson/Longman.

CASSEMIRO, E. (2009). Marcas da língua materna na produção escrita em língua inglesa. Dissertação de mestrado. LAEL. Pontifícia Universidade Católica, São Paulo, Brasil.

CICCHETTI, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, Vol. 6, No. 4, 284-290.

COHEN, J. (1960). *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, 20, 37-46. Disponível na Internet no endereço <http://www.4shared.com/document/NslFTMNJ/Jacob_Cohen_-_A_coefficient_of.html>. Acesso em 18 ago. 2011.

CONDI, R. (2005). Dois corpora, uma tarefa. O percurso de coleta, análise e utilização de corpora eletrônicos na elaboração de uma tarefa para ensino de inglês como língua estrangeira. Dissertação de mestrado. LAEL. Pontifícia Universidade Católica, São Paulo, Brasil.

CORDER, S. P. (1981) *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

DAVIES, M. (2008-). *The Corpus of Contemporary American English (COCA): 425 million words, 1990-present*. Disponível na Internet no endereço <http://www.americancorpus.org>. Acesso em 19 ago. 2012.

DELEGÁ-LUCIO, D. (2006). A relexicalização de adjetivos nas redações de alunos de inglês: um estudo baseado em Corpus de aprendiz. Dissertação de mestrado. LAEL. Pontifícia Universidade Católica, São Paulo, Brasil.

DUTRA, D. P., SILERO, R. P. (2010). Descobertas linguísticas para pesquisadores e aprendizes: a linguística de Corpus e o ensino de gramática. Revista Brasileira de Linguística aplicada. Vol. 10 n 4. Disponível na internet no endereço <http://www.scielo.br/scielo.php?pid=S1984-63982010000400005&script=sci_arttext#nt01>. Acesso em 09 mar. 2012.

EASTWOOD, J. (2002). *Oxford Guide to English Grammar*. Oxford: Oxford University Press.

FIRTH, J. R. (1957). *Papers in Linguistics - 1934-1951*. Oxford: Oxford University Press.

_____. (1968). A synopsis of linguistic theory, 1930-55. In. Palmer F.R. (Ed.) *Selected Papers of J. R. Firth 1952-59*. (pp. 168-205). London/Harlow: Longman.

GARDNER, H. (1985). *Frames of mind*. New York, Basic Books Inc.

GIULLI, A., SIGNORINI, A. (2005). The indexable web is more than 11.5 billion pages. Disponível na Internet no endereço: <http://www.cs.uiowa.edu/~asignori/pubs/web-size/>. Acesso em 24 mar. 2011.

GRANGER, S. (1998). The computer learner corpus : a versatile new source of data for SLA research. In: *Learner English on Computer*. (pp. 3-18). London: Longman.

_____. (2002). A bird's eye view of learner corpus research. In: HUNG, J., PETCH-TYSON, S. (Eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. (pp. 3-33). Amsterdam: Benjamins.

_____. (2008). Learner corpora. In: LÜDELING, A. & KYTÖ, M. (Ed.). *Corpus Linguistics: an International Handbook*. (pp. 259-275). Walter de Gruyter: Berlin/New York.

GRAYSON, K. e RUST, R. (2001). Interrater reliability. Journal of Consumer Psychology, 10(1&2), 71-73.

GREENBAUM, S. (1996). *The Oxford English Grammar*. New York: Oxford University Press.

HALLIDAY, M. A. K. (1991). Corpus studies and probabilistic grammar. In: AIJMER, K., ALTENBERG, B. (Orgs.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. (pp. 30-43). Londres: Longman.

_____. (1992). Language as system ans language as instance: the corpus as a theoretical construct. In: SVARTVIK, J. (Org.). *Directions in corpus linguistics. Proceedings of Nobel Symposium*. 82, Stockholm, 4-8.

HOEY, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

_____. (1997). From concordance to text structure: new used for computer corpora. In: LEWANDOWSKA-TOMASZCZYK, B., MELIA, P. J. (Orgs.). *PALC'97 – Practical Applications in Language Corpora*. pp. 2-22. Lodz, Lodz University Press.

_____. (2000). A world beyond collocation: new perspectives on vocabulary teaching. In: LEWIS, M. (Org.). *Teaching collocation: further developments in the lexical approach*. (pp. 224-43). Hove, LTP.

LADO, R. (1957). *Linguistics Across Cultures*. Michigan: Ann Arbor: University of Michigan Press.

LEECH, G. N. (1974). *Semantics: The Study of Meaning*. Harmondsworth: Penguin.

MICHAELIS, L. (2006). Time and Tense. In: AARTS, B. e MCMAHON, A. (eds.). *The Handbook of English Linguistics*. Oxford: Blackwell. Disponível na internet no endereço <http://spot.colorado.edu/~michaeli/MichaelistenseHEL.pdf>. Acesso em 20 jul. 2011.

NESSELHAUF, N. (2004). Learner corpora and their potencial for language teaching. In: SINCLAIR, J. (Ed.). *How to use Corpora in Language Teaching*. Amsterdam: John Benjamins Publishing Company.

_____. (2005). *Collocations in a Learner Corpus*. Amsterdam/Philadelphia: John Benjamins.

O'KEEFFE, A., MCCARTHY, M., CARTER, T. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

PARTINGTON, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdã/Filadélfia: John Benjamins.

SACCONI, L. A. (1996). *Minidicionário Sacconi da língua portuguesa*. São Paulo: Atual.

SCOTT, M., 2008, *WordSmith Tools version 5*, Liverpool: Lexical Analysis Software.

SELINKER, L. (1972). Interlanguage. In: Richards, Jack C. (Org.). (1974). *Error Analysis: Perspectives on Second Language Acquisition*. (pp. 31-54). London: Longman Group Limited.

SHEPHERD, D. (2001). Portuguese speakers. In: SWAN, M., SMITH, B. *A Teacher's Guide to Interference and Other Problems*. (pp. 113-128). Cambridge: Cambridge University Press.

SIEMENS, G. (2005). Connectivism: A learning theory for a digital age. Disponível na Internet no endereço: <http://www.itdl.org/Journal/Jan_05/article01.htm>. Acesso em 15 fev. 2012.

SIMON, S. (2005/2008). What is a Kappa coefficient? (Cohen's Kappa). Disponível na Internet no site < http://www.childrensmercy.org/stats/definitions/kappa.htm>. Acesso em 15 fev. 2012..

SINCLAIR, J. (1966). Beginning the study of lexis. In: BAZELL, C. E. *In Memory of J. R. Firth*. (pp. 410-30). Londres: Longman.

_____. (1987). Collocation : a progress report. In: STEELE, R., THREADGOLD, T. *Language Topics: Essays in Honour of Michael Halliday*. (pp. 319-32. V.2). Amsterdã/Filadélfia: John Benjamins.

_____. (1987b). *Looking up*. ?: Collins Cobuild.

_____. (1988). Naturalness in language. *ELR Journal*. v.2. pp. 11-20. Birmingham: Birmingham University Press..

_____. (1991). *Corpus, Concordance, Collocation*. Oxford: Orford University Press.

_____. (1996). *EAGLES. Preliminary Recommendations on Corpus Typology*. Disponível na Internet no endereço <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.5014&rep=rep1&type=pdf>. Acesso em 10 jan. 2012..

_____. (2004). *Trust the Text*. New York: Routledge.

_____. (2004b). (Ed.). *How to use Corpora in Language Teaching*. Amsterdam: John Benjamins Publishing Company.

SINCLAIR, J., TEUBERT, W. (2007). Interview with John Sinclair, conducted by Wolfgang Teubert. In: TEUBERT, W., RAMESH, K. (Eds.). *Corpus Linguistics. Critical Concepts in Liguistics*. (Vol. 1, cap. 12). Abingdon: Routledge.

SWAN, M., SMITH, B. (2001). *Learner English – A Teacher's Guide to Interference and Other Problems*. Cambridge: Cambridge University Press.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

UR, P. (2010). English as a lingua franca: a teacher's perspective. *Cadernos de Letras (UFRJ)* n. 27 - dez. Disponível na Internet no endereço: <http://www.letras.ufrj.br/anglo_germanicas/cadernos/numeros/122010/textos/cl301220100penny.pdf>. Acesso em 05 jan. 2012.

Anexos

Anexo 1:

Aceite do Comitê de Ética em Pesquisa – PUC-SP.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO
COMITÊ DE ÉTICA EM PESQUISA DA PUC-SP
SEDE CAMPUS MONTE ALEGRE

Protocolo de Pesquisa nº 230/2011

Faculdade de Filosofia, Comunicação, Letras e Artes
Programa de Estudos Pós-Graduados em Linguística Aplicada: Estudo da Linguagem
Orientador(a): Prof.(a). Dr.(a). Antonio Paulo Berber Sardinha
Autor(a): Wendel Mendes Dantas

PARECER sobre o Protocolo de Pesquisa, em nível de Dissertação de Mestrado, intitulado ***Classificando mau-usos nas produções escritas de aprendizes de inglês como língua estrangeira do córpus COBRA-SEVEN***

CONSIDERAÇÕES APROVADAS EM COLEGIADO

Em conformidade com os dispositivos da Resolução nº 196 de 10 de outubro de 1996 e demais resoluções do Conselho Nacional de Saúde (CNS) do Ministério da Saúde (MS), em que os critérios da relevância social, da relação custo/benefício e da autonomia dos sujeitos da pesquisa pesquisados foram preenchidos.

O Termo de Consentimento Livre e Esclarecido permite ao sujeito compreender o significado, o alcance e os limites de sua participação nesta pesquisa.

A exposição do Projeto é clara e objetiva, feita de maneira concisa e fundamentada, permitindo concluir que o trabalho tem uma linha metodológica bem definida, na base do qual será possível retirar conclusões consistentes e, portanto, válidas.

No entendimento do CEP da PUC-SP, o Projeto em questão não apresenta qualquer risco ou dano ao ser humano do ponto de vista ético.

CONCLUSÃO

Face ao parecer consubstanciado apensado ao Protocolo de Pesquisa, o Comitê de Ética em Pesquisa da Pontifícia Universidade Católica de São Paulo – PUC/SP – Sede Campus Monte Alegre, em Reunião Ordinária de **12/09/2011**, **APROVOU** o Protocolo de Pesquisa nº **230/2011**.

Cabe ao(s) pesquisador(es) elaborar e apresentar ao CEP da PUC-SP – Sede Campus Monte Alegre, os relatórios parcial e final sobre a pesquisa, conforme disposto na Resolução nº 196 de 10 de outubro de 1996, inciso IX.2, alínea "c", do Conselho Nacional de Saúde (CNS) do Ministério da Saúde (MS), bem como cumprir integralmente os comandos do referido texto legal e demais resoluções do Conselho Nacional de Saúde (CNS) do Ministério da Saúde (MS).

São Paulo, 12 de setembro de 2011.

Prof. Dr. Edgard de Assis Carvalho
Coordenador do Comitê de Ética em Pesquisa da PUC-SP

Rua Ministro Godói, 969 – Sala 63-C (Andar Térreo do E.R.B.M.) – Perdizes – São Paulo – SP – CEP: 05015-001
Tel/Fax: (11) 3670-8466 – e-mail: cometica@pucsp.br – site: http://www.pucsp.br/cometica

Anexo 2:

Planilhas do programa computacional Microsoft Excel 2010 usados nesta pesquisa.

Devido ao grande número de páginas das planilhas contendo os erros encontrados nas produções escritas estudadas nesta pesquisa esses anexos estão disponibilizados no CD-rom anexo.