PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO

PUC-SP

TELMA DE LURDES SÃO BENTO FERREIRA

LINGUÍSTICA DE CORPUS E AUTENTICIDADE DE LIVROS DIDÁTICOS:

O CASO DO PORTUGUÊS COMO LÍNGUA ESTRANGEIRA (PLE)

MESTRADO EM LINGUÍSTICA APLICADA E ESTUDOS DA LINGUAGEM

SÃO PAULO

2010

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO

PUC-SP

TELMA DE LURDES SÃO BENTO FERREIRA

LINGUÍSTICA DE CORPUS E AUTENTICIDADE DE LIVROS DIDÁTICOS:

O CASO DO PORTUGUÊS COMO LÍNGUA ESTRANGEIRA (PLE)

MESTRADO EM LINGUÍSTICA APLICADA E ESTUDOS DA LINGUAGEM

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 a orientação do Prof. Dr. Antônio Paulo Berber Sardinha.

SÃO PAULO

2010

Dissertação defendida e aprovada em: __ / __ / __

Banca Examinadora

_____

_____

_____

Ao meu marido, Ricardo, e à nossa filha, Maria Sofia.

Ao meu pai, Manuel, e à minha sogra, Maria Joaquina.

À minha mãe, Maria de Lurdes (in memoriam)*.*

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus pela luz, proteção e força concedidas.

Ao Prof. Dr. Antônio Paulo Berber Sardinha pela acolhida, dedicação, profissionalismo e sabedoria.

Aos professores do LAEL, em especial à Profª Drª Maria Antonieta Alba Celani.

Às professoras doutoras Tânia Maria Granja Shepherd e Solange Maria Sanches Gervai pelas valiosas contribuições.

Ao colega José Lopes Moreira Filho pelos *scripts* elaborados especialmente para esta dissertação.

Aos colegas orientandos, em especial às queridas Márcia Veirano, Rosana Teixeira, Denise Delegá Lúcio e Solange Contrera pelas sugestões.

Às queridas amigas e coautoras do *Muito Prazer*, Vera Lúcia Ramos e Glaucia Roberta Rocha Fernandes, pelos caminhos percorridos e por aqueles que ainda iremos percorrer.

À minha família, em especial ao meu querido marido, pela paciência, amor e apoio infinitos, e ao meu pai e à minha sogra pelo carinho e pelos serviços de "babás" da Maria Sofia, enquanto eu escrevia esta dissertação.

À minha filha, Maria Sofia, por ser minha "companheirinha" (apesar de não entender o que a mamãe fazia tanto tempo em frente ao computador).

Ao Evandro Lisboa Freire pela revisão minuciosa e ao Rafael Martins pela diagramação.

À Maria Lúcia dos Reis e à Márcia Martins pela dedicação e atenção.

À CAPES pelo auxílio financeiro.

A todos que, direta ou indiretamente, contribuíram para a concretização deste trabalho.

Sobre a tarefa de quem ensina

É uma tarefa que requer de quem com ela se compromete um gosto especial de querer bem não só aos outros, mas ao próprio processo que ela implica. É impossível ensinar sem essa coragem de querer bem, sem a valentia dos que insistem mil vezes antes de uma desistência. É impossível ensinar sem a capacidade forjada, inventada, bem cuidada de amar (...). É preciso ousar, no sentido pleno desta palavra, para falar em amor sem temer ser chamado de piegas, de meloso, de acientífico, senão de anticientífico. É preciso ousar para dizer cientificamente e não blá-blá-blantemente, que estudamos, aprendemos, ensinamos, conhecemos com o nosso corpo inteiro. Com os sentimentos, com as emoções, com os desejos, com os medos, com as dúvidas, com a paixão e também com a razão crítica. Jamais com esta apenas. É preciso ousar para jamais dicotomizar o cognitivo do emocional. É preciso ousar para ficar ou permanecer ensinando por longo tempo nas condições que conhecemos, mal pagos, desrespeitados e resistindo ao risco de cair vencidos pelo cinismo. É preciso ousar, aprender a ousar, para dizer não à burocratização da mente a que nos expomos diariamente.

(Paulo Freire)

SUMÁRIO

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

RESUMO

Esta pesquisa pretende mostrar os resultados da análise de um material didático, de cuja autoria participo, para o ensino de Português como Língua Estrangeira (PLE) (Fernandes et al., 2008). A pesquisa teve como objetivo principal a investigação dos aspectos indicativos de autenticidade do material didático analisado, partindo do pressuposto de que mesmo textos não autênticos podem apresentar elementos característicos de autenticidade e que tais elementos podem ser detectados por meio de metodologia de Linguística de Corpus. Para tanto, a pesquisa está embasada na Linguística de Corpus e nos conceitos de autenticidade (Berber Sardinha, 2007; Nunan, 1989), idiomaticidade (Sinclair, 1991) e de pacotes lexicais (Biber et al., 1999).

Desenvolvemos e aplicamos uma metodologia de identificação de autenticidade em *corpora*, que, em síntese, baseia-se na análise da lexicogramática dos textos envolvidos em busca dos padrões que possam fornecer evidências de autenticidade (ou não) do material didático, visto que se espera que a frequência e a quantidade dos padrões encontrados reflita o uso real da linguagem.

Os resultados da análise indicaram que nem todo texto não autêntico é um mau exemplo de lexicogramática, visto que os textos não autênticos do material didático apresentaram muitas ocorrências de pacotes lexicais recorrentes na linguagem autêntica. Ao mesmo tempo, o simples fato de ser autêntico não garante que os pacotes lexicais que o texto contenha sejam típicos da linguagem falada ou escrita.

A pesquisa pretende contribuir para a área visto que não há, até o momento, nenhuma pesquisa que enfoque a análise de autenticidade em materiais didáticos para o ensino de PLE.

Palavras-chave: Linguística de Corpus, ensino de Português como Língua Estrangeira, autenticidade, pacotes lexicais.

ABSTRACT

This study aims to show the results of an analysis of a textbook, of which I am co-author, for the teaching of Portuguese as a Foreign Language (PFL) (Fernandes et al., 2008). The main aim of the research was the investigation of indicative aspects of authenticity in the teaching material analyzed, starting from the premise that even non-authentic texts may show characteristic elements of authenticity, and that these elements can be detected using the methods from Corpus Linguistics. As such, this research is based on Corpus Linguistics and the concepts of authenticity (Berber Sardinha, 2007; Nunan, 1989), idiomaticity (Sinclair, 1991), and lexical bundles (Biber et al., 1999).

We developed and applied a methodology for identification of authenticity in corpora that, in summary, is based on the lexico-grammatical analysis of the texts involved in a search for patterns that might provide evidence of authenticity (or otherwise) of teaching material, given that the frequency and quantity of the patterns found are expected to reflect the actual usage of language.

Results indicated that not every non-authentic text is a bad example of lexico-grammar, since this kind of text included in the teaching material showed many instances of lexical bundles recurrent in authentic language. At the same time, the mere fact of being authentic does not guarantee that the lexical bundles contained in a text are typical of spoken or written language.

The research aims to contribute to the area since to date there has been no research focusing the analysis of authenticity in teaching materials for PFL.

Keywords: Corpus Linguistics, teaching of Portuguese as a foreign language, authenticity, lexical bundles.

INTRODUÇÃO

É possível perceber a utilização cada vez maior de *corpora* na elaboração e análise de materiais didáticos. A pesquisa com Linguística de Corpus tem informado, influenciado e revolucionado, de acordo com alguns pesquisadores, materiais de ensino e trabalhos de referência, particularmente no contexto de ILE (Inglês como Língua Estrangeira) (Braun et al., 2006). Sobre os estudos baseados em *corpus*, Braun et al. afirmam que "as observações baseadas em *corpus* têm ajudado a descobrir e remover discrepâncias entre o que é ensinado nos materiais didáticos e o que é, de fato, usado" (Braun et al., 2006:1)[1 e 2].

No entanto, pesquisas mostram que o uso de *corpora* no ensino de PLE[3] ainda não é efetivamente explorado. Embora encontremos um número razoável de pesquisas (cf. Paes Almeida, 2007), o uso de *corpora* eletrônicos ainda é pouco explorado. Das pesquisas encontradas, destacamos a de Berber Sardinha (1999) que é, até onde sabemos, o primeiro artigo que trata de *corpora* e ensino de línguas e o *corpus* foi o primeiro usado para lecionar PLE com a metodologia da Linguística de Corpus.

Com relação à análise de materiais didáticos para o ensino de PLE, há várias pesquisas que analisam os materiais disponíveis no mercado. No entanto, a maior parte delas é manual (*page-by-page approach*), ou seja, não utiliza métodos automáticos nem se beneficia do uso de *corpora* da língua. Encontramos somente duas pesquisas que utilizam-se de *corpus* para análise: Cavalcante (2006), que analisa as formas verbais em um livro didático, e Carvalho (2007), que se utilizou de um *corpus* de materiais didáticos com o intuito de responder qual é a imagem do Brasil construída neles.

---

1. "Corpus-based observations have helped to uncover and remove discrepancies between what is taught in schoolbooks and what is actually used."
2. Todas as traduções de trechos em inglês citados são de nossa autoria.
3. Nesta pesquisa, adotamos a abreviação PLE (Português como Língua Estrangeira) seguindo a indicação de Paes Almeida (2007), que afirma que além da abreviação normalmente ser utilizada no ensino de português fora do Brasil, onde o português é uma língua estrangeira, ela pode ser utilizada como genérica.

Esta pesquisa empreendeu a análise do material didático *Muito Prazer – Fale o Português do Brasil* (doravante MP), de cuja autoria participo (Fernandes et al., 2008). A pesquisa tem por pressupostos teóricos principais a Linguística de Corpus e os conceitos de autenticidade (Berber Sardinha, 2007; Nunan, 1989), idiomaticidade (Sinclair, 1991) e de pacotes lexicais (Biber et al., 1999) com o objetivo específico de investigar os aspectos indicativos de autenticidade do material didático analisado, partindo do pressuposto de que mesmo textos originalmente elaborados para fins didáticos podem apresentar elementos característicos de autenticidade, ou seja, a autenticidade é uma questão de grau, sendo que os textos podem ser mais ou menos autênticos, e não simplesmente autênticos ou não autênticos. A ideia de grau implica que há probabilidade de textos não autênticos terem maior ou menor aproximação com a linguagem atestada em *corpora* eletrônicos. Nossa proposta é justamente verificar o grau de aproximação observado nos textos do material didático, que contém tanto textos autênticos quanto não autênticos.

A linguagem autêntica possui uma característica que Sinclair (1991) chama de idiomaticidade (*idiom principle*), isto é, um conjunto de fatores que a tornam natural, normalmente associados à escolha das combinações lexicogramaticais que são empregadas. Sendo assim, uma das maneiras para inferir a idiomaticidade das escolhas lexicais, de acordo com Berber Sardinha (2007: 277-278), é por meio da quantidade de ‘pacotes lexicais’ presentes no texto. Um pacote lexical, em linhas gerais, é uma sequência de palavras fixas, de extensão variável, muitas vezes chamada de ‘chunk’ (Lewis, 2000) ou ‘cluster’ (Scott & Tribble, 2006). Esses padrões podem ser bem formados ou não, como “bom-dia” ou “que possui um”, e devem ocorrer com certa frequência para serem considerados pacotes lexicais.

Dessa maneira, recorremos a *corpora* eletrônicos em busca dos padrões que pudessem fornecer evidências de autenticidade (ou não) do material didático, visto que se espera que a frequência e a quantidade dos padrões encontrados reflita o uso real da linguagem. Para isso, esta pesquisa desenvolveu e aplicou uma metodologia de identificação de autenticidade em *corpora*, que, em síntese, baseia-se na análise da lexicogramática dos textos envolvidos, promovendo a comparação

dos trigramas e pacotes lexicais[4] presentes no material didático com os dos *corpora* de referência.

Assim, as questões de pesquisa que nortearam o trabalho são as seguintes:

1. Quantos trigramas e pacotes lexicais existem nos textos (falados e escritos) do material didático?
2. Quais desses são convergentes (i.e., existem no *corpus* de referência) e divergentes (i.e., não existem no *corpus* de referência)?
3. A proporção de uso dos convergentes é equivalente nos *corpora*?
4. Com base nas respostas às perguntas acima, qual é o grau de autenticidade dos textos do material didático?

A fim de responder a essas perguntas, utilizamos as ferramentas computacionais do pacote WordSmith Tools versão 3.0 (Scott, 1997) e *scripts*[5] em Shell e Python especialmente desenvolvidos para esta pesquisa.

Isso posto, segue-se uma breve explanação acerca da organização desta dissertação.

O primeiro capítulo foi dedicado à fundamentação teórica da pesquisa e está dividido em três seções principais: a primeira apresenta princípios teóricos da Linguística de Corpus, bem como o uso de *corpora* no ensino de línguas estrangeiras e na análise de materiais didáticos; a segunda apresenta os conceitos de autenticidade, de textos autênticos e não autênticos, e de idiomaticidade utilizados na pesquisa; e a terceira trata da definição de 'pacotes lexicais' (*lexical bundles*).

O segundo capítulo apresenta a metodologia empregada na pesquisa, incluindo a descrição dos *corpora*, bem como a especificação dos procedimentos de análise dos dados. Primeiramente, são detalhados os *corpora* que compuseram o estudo e, em seguida, é especificado o processo de análise e seleção dos dados e as ferramentas utilizadas.

---

4. Os termos 'trigramas' e 'pacotes lexicais' aqui utilizados designam, respectivamente, sequências de três palavras fixas e sequências de palavras fixas de alta frequência.
5. *Scripts* criados pelo professor orientador e por um colega do grupo de pesquisa GELC (Grupo de Pesquisa em Linguística de Corpus), do qual a autora faz parte.

O terceiro capítulo apresenta as estatísticas gerais dos *corpora*, os resultados das análises quantitativa e qualitativa, bem como as descobertas feitas com relação aos trigramas e pacotes lexicais convergentes e divergentes encontrados no material didático comparados ao *corpus* de referência BP.

As considerações finais retomam os pontos relevantes da pesquisa e trazem, ainda, a discussão dos resultados. Incluímos apêndices e anexos para proporcionar ao leitor a oportunidade de ter acesso a dados complementares do material didático pesquisado no âmbito deste estudo.

CAPÍTULO 1

FUNDAMENTAÇÃO TEÓRICA

Este capítulo constitui o arcabouço teórico da pesquisa e está dividido em três seções principais: a primeira apresenta princípios teóricos da Linguística de Corpus, bem como o uso de *corpora* no ensino de línguas estrangeiras e na análise de materiais didáticos; a segunda apresenta os conceitos de autenticidade, de textos autênticos e não autênticos, e de idiomaticidade utilizados na pesquisa; e a terceira trata da definição de 'pacotes lexicais' (*lexical bundles*).

A pesquisa tem por pressupostos teóricos principais a Linguística de Corpus e os conceitos de autenticidade (Berber Sardinha, 2007; Nunan, 1989), idiomaticidade (Sinclair, 1991) e de pacotes lexicais (Biber et al., 1999) com o objetivo específico de investigar os aspectos indicativos de autenticidade do material didático analisado, partindo do pressuposto de que mesmo textos originalmente elaborados para fins didáticos podem apresentar elementos de autenticidade.

1.1 Linguística de Corpus

O trabalho aqui proposto tem como fundamentação teórica principal a Linguística de Corpus (LC) que, de acordo com Berber Sardinha (2004: 3),

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

Segundo Fox (1998), nos últimos anos houve uma revolução na maneira como a língua pode ser estudada. É possível compilar grandes *corpora* que possibilitem aos pesquisadores a análise da linguagem como está sendo usada hoje e como foi usada em períodos específicos do passado, sendo possível deixar de lado suas intuições e verificar o que os dados lhes dizem. Como Leech (1992: 107 apud

Kennedy 1998: 7) observou, “o foco do estudo está (...) na observação da linguagem em uso que leva à teoria, e não o contrário” [1 e 2].

De acordo com Berber Sardinha, a LC trabalha dentro de um quadro conceitual formado por uma abordagem empirista, ou seja, que dá primazia aos dados provenientes da observação da linguagem, e uma visão de linguagem como sistema probabilístico. Segundo o autor, essa visão pressupõe que embora muitos traços linguísticos sejam possíveis teoricamente, não ocorrem com frequência relevante (Berber Sardinha, 2004).

Kennedy (1998) indica que, muitas vezes, as evidências para as teorias da linguagem vêm da intuição e introspecção. No caso de uma pesquisa baseada em *corpora*, a evidência vem diretamente dos textos, sendo que a LC se preocupa não somente com palavras, estruturas e usos possíveis, mas com o que é mais provável ocorrer na linguagem em uso. Assim, o que importa à LC não é a possibilidade de algo ocorrer, mas sim a *probabilidade* de ela ocorrer. O foco principal da LC é empírico e ela se preocupa com os padrões da língua conforme esta é usada, determinando o que é típico e o que é incomum em determinadas circunstâncias.

1.1.1 Definição de *corpus*

A definição mais completa de *corpus*, segundo Berber Sardinha (2004: 18-9), é a seguinte:

> Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise (Sanchez, 1995: 8-9).

De acordo com o autor, essa definição de *corpus* pode ser apontada como a mais completa porque menciona a origem (dados autênticos), o propósito (objeto de estudo linguístico), a composição (conteúdo criteriosamente escolhido), a

1. “The focus of study is (…) on observation of language in use leading to theory rather than vice-versa.”
2. Todas as traduções de trechos em inglês citados são de nossa autoria.

formatação (dados legíveis por computador), a representatividade (de uma língua ou variedade linguística) e a extensão (vasto o suficiente para ser representativo).

1.1.2 Tipos de *corpora*

Os *corpora* podem ter tamanhos e formatos variados, visto que são compilados para pesquisas e necessidades diferentes. Berber Sardinha (2004: 20-21) agrupa os principais tipos de *corpus* segundo os seguintes critérios:

- Modo: Falado ou Escrito.
- Tempo: Sincrônico ou Diacrônico, Contemporâneo ou Histórico.
- Seleção: De amostragem (estático, amostra finita da linguagem como um todo), Monitor (dinâmico ou orgânico), Equilibrado (textos distribuídos em quantidades semelhantes).
- Conteúdo: Especializado, Regional ou Dialetal, Multilíngue.
- Autoria: De aprendiz (falantes não nativos), De língua nativa.
- Disposição interna: Paralelo (textos comparáveis, por exemplo, original e tradução), Alinhado (tradução abaixo de cada linha do original).
- Finalidade: De estudo (*corpus* a ser descrito), De referência (para contrastar com o *corpus* de estudo), De treinamento ou teste (para desenvolvimento de aplicações e ferramentas de análise).

Nesta pesquisa, nosso *corpus* de estudo (material didático analisado) contém as seguintes características:

- Falado[3] e Escrito.
- Contemporâneo.
- De amostragem.
- Estático, i. e., não há crescimento ou diminuição, visto que nosso *corpus* trata da amostra do livro didático.
- De língua nativa.

3. Em sentido estrito, trata-se de uma simulação de transcrição da linguagem oral.

1.1.2.1 Corpora *disponíveis on-line*

Para as pesquisas em LC, há a necessidade da disponibilidade de *corpora* eletrônicos (ou, se for o caso, a compilação de um *corpus* para estudo). Alguns dos principais *corpora* eletrônicos da língua portuguesa disponíveis on-line são elencados na Tabela 1.1.

Tabela 1.1 – *Corpora* disponíveis on-line

Corpus	**Palavras**	**Composição**	**Localização**
Corpus Brasileiro	1 bilhão	Português brasileiro, escrito e falado	PUC-SP
Banco do Português (v. 2.0)	660 milhões	Português brasileiro, escrito e falado	PUC-SP
Projecto AC/DC (Acesso a *Corpora* / Disponibilização de *Corpora*)	360 milhões	Português escrito e falado, com predominância da variedade europeia	Projeto Linguateca
Modern Portuguese	315 mil	Português literário (romances)	Brigham Young University
CetemPublico (*Corpus* de Extractos de Textos Electrónicos MCT/Público)	229 milhões	Português europeu	Projeto Linguateca
Corpus Unesp / Araraquara / Usos do Português	200 milhões	Português brasileiro, escrito	Unesp (Araraquara)
CRPC (*Corpus* de	152,6 milhões	Português de	CLUL – Centro de

Referência do Português Contemporâneo)		vários países lusófonos, com predominância da variedade europeia	Linguística da Universidade de Lisboa
Corpus do Português Brasileiro Contemporâneo	100 milhões	Português brasileiro, escrito e falado	Unesp (Araraquara)
Corpus do Português	45 milhões	Português do século XIV ao século XX	Brigham Young University
Portext	30 milhões	Português escrito de vários países	Universidade de Nice
Modern Newspapers	28 milhões	Português escrito, jornalístico e entrevistas publicadas em jornais	Brigham Young University
CetenFolha (*Corpus* de Extractos de Textos Electrónicos Nilc/Folha de S. Paulo)	24 milhões	Português brasileiro retirado do jornal *Folha de S. Paulo*	Projeto Linguateca
Comet (*Corpus* Multilíngue para Ensino e Tradução)	5 milhões	Português escrito comparável com inglês	USP
CR-LW (*Corpus* de Referência Lácio-Web)	5 milhões	Português brasileiro, escrito	Nilc (USP, UFSCAR, Unesp (Araraquara))
Historical	2,8 milhões	Português escrito	Brigham Young

Portuguese Prose		(1300 a 1900)	University
TychoBrahe Parsed Corpus of Historical Portuguese	1,9 milhão	Português antigo (1550 a 1850)	Unicamp
Borba-Ramsey Corpus of Brazilian Portuguese	1,67 milhão	Português brasileiro, escrito	Brigham Young University
Corpus Internacional do Português	1 milhão	Português europeu	Universidade de Lisboa
Compara	Informação não disponível	*Corpus* paralelo – de originais e traduções – português e inglês	Projeto Linguateca
Cordial (*Corpus* de Discurso para a Análise de Língua e Literatura)	Informação não disponível	Português escrito	UFMG
Nupill (Núcleo de Pesquisas em Informática, Linguística e Letras)	Informação não disponível	Português escrito	UFSC
Nurc (Projeto de Estudo da Norma Linguística e Letras)	Informação não disponível	Português brasileiro, falado	USP, UFRJ, UFBA, UFPE, UFRGS
PHPB (Projeto para a História do Português Brasileiro)	Informação não disponível	Português escrito	UFPE, UFPBA, UFMG, UFRJ, EFSC, UFPB, USP
Português Falado	Informação não	Português	UFC, URCA

do Ceará	disponível	brasileiro, falado	
Varport (Análise Contrastiva de Variantes do Português)	Informação não disponível	Português escrito e falado, brasileiro e europeu	UFRJ, CLUL
Varsul (Variação Linguística Urbana da Região Sul)	Informação não disponível	Português falado	UFSC, UFRGS, UFPR

Fontes: Berber Sardinha (2004: 9-10) e COMET (2009).

Nesta pesquisa, trabalharemos com o *Corpus* do Banco do Português (LAEL/PUC-SP) como *corpus* de referência principal. Essa escolha se deve ao fato de ser um *corpus* contemporâneo de língua geral, hoje com cerca de 660 milhões de palavras do português do Brasil (versão 2.0), o segundo maior *corpus* de português do Brasil no momento, mas o primeiro em relação à análise de dados realizada. Composto de gêneros variados é, segundo Berber Sardinha (2004: 164), "um *corpus* orgânico, pois é aberto e seu conteúdo está em constante expansão e renovação".

1.1.3 Tipos de pesquisa com *corpus*

Segundo Tognini-Bonelli (2001 apud Shepherd, 2009) as abordagens de pesquisas em LC podem ser baseadas (*corpus-based*) ou dirigidas (*corpus-driven*) por *corpus*. As pesquisas baseadas em *corpus* se aproveitam do *corpus* para expor ou testar hipóteses e exemplificar teorias e descrições linguísticas pré-existentes. As abordagens dirigidas por *corpus*, em contrapartida, têm como ponto de partida o *corpus* e visam à observação dos dados que levam à hipótese e à generalização.

Com relação às pesquisas de desenvolvimento de materiais didáticos, Biber et al. (2002 apud Shortall, 2007) sugerem que as investigações baseadas em *corpus* podem informar os materiais didáticos, em particular quanto às construções gramaticais. Ainda de acordo com o autor, a principal vantagem de uma abordagem baseada em *corpus* é que garante que os alunos estão sendo expostos à linguagem

que realmente ocorre em interações no mundo real. Isso também significa que quaisquer regras gramaticais apresentadas no material representam o uso real. Sendo assim, os autores de materiais didáticos podem, por meio de pesquisas com *corpora*, confirmar suas intuições e incluir o que é importante e usado na língua. Para Hunston e Francis (1998: 45 apud Shortall, 2007) a abordagem dirigida por *corpus* na elaboração de materiais leva a descrições da linguagem baseadas em dados autênticos em vez de intuições dos autores e/ou comprometimento da língua.

Esta pesquisa foi dirigida por *corpus*, ou seja, partimos do *corpus* de estudo para verificar o grau de autenticidade dos textos, em especial os elaborados para fins didáticos.

1.1.4 Linguística de Corpus e ensino

Os *corpora* eletrônicos e seus programas estão provando ter cada vez mais influência no ensino de línguas como fontes de descrição de linguagem e materiais pedagógicos (Gabrielatos, 2005). A disponibilidade cada vez maior desses *corpora* e o emprego maior do computador no ensino e pesquisa motivaram uma mudança do nosso entendimento de questões-chave acerca do funcionamento, comportamento, descrição e ensino do léxico. Segundo Berber Sardinha (2004), desde os anos 1970 a descrição da linguagem baseada em *corpus* tem apresentado um crescimento contínuo na área do ensino e aprendizagem de línguas, na qual já há várias aplicações derivadas da LC destinadas especificamente ao ensino. Exemplos cada vez mais comuns são a utilização de bancos de dados de milhões de palavras na confecção e atualização de dicionários e gramáticas, em especial da língua inglesa (O'Keeffe et al., 2007).

Com relação à contribuição da LC para o ensino de uma segunda língua (L2), Conrad (2005: 395) afirma que ela está relacionada à importância que a LC coloca nos estudos empíricos de grandes bancos de dados da língua. Assim, a partir das observações do comportamento da linguagem em uso podemos desenvolver teorias e descrições da língua em questão. Especificamente em pesquisas de análise e desenvolvimento de materiais didáticos, como o caso da presente pesquisa, o uso de *corpora* de L1 (língua nativa ou primeira língua) em estudos linguísticos fornece

evidências convincentes de discrepâncias entre o uso real e as visões de linguagem baseadas na introspecção (Sinclair, 1997 apud Gabrielatos, 2005) e nos revela padrões que não haviam sido detectados por ela. Essa afirmação corrobora a de Hunston (2002: 13 apud Shortall, 2007), que indica que os *corpora* informam como a língua funciona de uma forma que não é acessível à intuição de um falante nativo e cita, como exemplo, a fraseologia. De acordo com O'Keeffe et al. (2007: 60), em linhas gerais, os *corpora* podem "revelar as preferências dos usuários da língua com relação aos padrões, ao escrever e falar, nos contextos representados nos *corpora* coletados"[4].

Para Gabrielatos (2003: 2), "a intuição do falante nativo nem sempre é confiável e a condição de falante nativo não nos garante, automaticamente, uma visão consciente, clara e abrangente da língua em todos seus contextos de uso"[5]. Além disso, ainda de acordo com o autor, é a pesquisa com *corpus* que fornece as evidências mais convincentes de discrepâncias entre as intuições e o uso real da língua. No caso do desenvolvimento de materiais didáticos, Mindt (1996 apud Shortall, 2007) afirma que os estudos baseados em *corpus* proporcionam a oportunidade de tornar esses materiais mais próximos da realidade. Além disso, é possível utilizar *corpora* de materiais didáticos para análise da linguagem à qual os alunos estão sendo expostos. Quando comparamos esses *corpora* a um *corpus* de L1, é possível contrastar o que está sendo ensinado com a linguagem em uso, o que facilita o desenvolvimento de materiais mais eficientes (Gabrielatos, 2005).

Há vários estudos que contrastam o conteúdo encontrado em materiais didáticos (doravante MDs) de diferentes línguas com *corpora* de falantes nativos, mas, até onde sabemos, há somente um estudo sobre o ensino de português para estrangeiros. Com relação às pesquisas que verificam a autenticidade dos textos de MDs por meio da análise dos padrões observados no material, há somente dois estudos em Inglês como Língua Estrangeira (ILE) (Allan, 2009 e Contrera, 2010). No entanto, não é do nosso conhecimento a existência de algum estudo para o ensino de Português como Língua Estrangeira (PLE).

---

4. "(…) Reveal the regular, patterned preferences of the language users represented in it, speaking and writing in the contexts in which the corpus was gathered."
5. "Native-speaker intuitions are not always dependable. Being a native speaker does not automatically give us a conscious, clear and comprehensive picture of our language in all its contexts of use."

1.1.4.1 Linguística de Corpus e ensino de Português como Língua Estrangeira

A LC, de acordo com Berber Sardinha (2000: 4-5), expõe alguns "mitos" acerca da descrição da linguagem que eram aceitos e difundidos nos livros didáticos e de referência como "verdades". Tal mitologia incluiria a crença de que:

(1) há dois níveis independentes de organização da linguagem, a sintaxe e o léxico;

(2) a sintaxe tem precedência sobre o léxico, servindo como base para o 'preenchimento' de 'lacunas' sintáticas;

(3) a fluência nativa é algo subjetivo que reside na mente dos falantes nativos e que não pode ser observada e descrita objetivamente;

(4) a frequência dos traços linguísticos enquanto reveladora de padronização e convencionalidade do uso da língua é irrelevante e, portanto, os alunos não precisam aprender sobre modos típicos de expressão em contextos específicos. Ainda de acordo com o autor, a posição que emerge da descrição da linguagem baseada em *corpus* diante dessa mitologia seria a seguinte:

(a) a linguagem não é estruturada pelo princípio de 'lacuna e preenchimento'[6] (Lewis, 2000; Sinclair, 1991 apud Berber Sardinha, 2000: 4);

(b) a linguagem é padronizada (Berber Sardinha, 2000: 4);

(c) a sensação de naturalidade e fluência nativa não são aspectos abstratos, mas possuem traços linguísticos demonstráveis por meio de padrões (Cowie, 1998 apud Berber Sardinha, 2000: 5);

6. Ou 'slot and filler' em inglês. De acordo com esse esquema, as lacunas sintáticas podem ser preenchidas lexicalmente de qualquer modo, desde que o conjunto de lacunas seja estruturalmente plausível (Berber Sardinha, 2004).

(d) a diferença entre sintaxe e léxico é mais uma conveniência metodológica do que uma realidade observável (Sinclair, 1991 apud Berber Sardinha, 2000: 5);

(e) a frequência dos traços linguísticos é pertinente para uma teoria da linguagem já que nem todas as possibilidades estruturais se realizam e as frequências dos traços ocorrentes variam sistematicamente (de Beaugrande, 1999; Halliday, 1991 e 1992 apud Berber Sardinha, 2000: 5).

No entanto, ao analisarmos alguns dos materiais didáticos de PLE disponíveis no mercado (ver Anexo 3), não identificamos, até o momento, nenhuma iniciativa de utilização dos preceitos da LC em sua confecção e, com relação às pesquisas, apesar de haver um número razoável sobre o ensino de PLE (cf. Paes Almeida, 2007), poucas delas utilizam *corpora* (ver subseção 1.1.5, Estado da Arte).

1.1.4.2 Linguística de Corpus e análise de materiais didáticos

Há várias pesquisas sobre análise de materiais didáticos, não só para o ensino de língua inglesa, mas, também, para o ensino de português para estrangeiros. No entanto, a maior parte dessas análises são manuais, ou seja, não se beneficiaram do uso de *corpora* da língua. Aijimer (2009) pesquisou os estudos de análise de materiais didáticos da língua inglesa e constatou que, mesmo com o aumento do interesse em pesquisas na área, a partir dos anos 1980, e tendo em vista que diferentes linhas de pesquisa podem ser realizadas, a maior parte das abordagens ainda é feita utilizando a metodologia manual, 'página por página' (*page-by-page approach*). A autora afirma que somente seis estudos recentes utilizaram métodos automáticos, ou seja, abordagem com *corpus* (*corpus approach*), entre eles Biber et al. (2004), que investigaram os 'pacotes lexicais' (*lexical bundles*). O estudo revelou que o discurso de sala de aula e o livro didático de 'Inglês para Fins Acadêmicos' (*English for Academic Purposes*) mostram características de linguagem específicas e resultados diferentes dos esperados pelos autores. Quando se utiliza *corpora* de materiais didáticos, temos uma vertente que se convenciona chamar de '*textbook corpora*', ou *corpora* de materiais didáticos, em português.

Ainda no ensino da língua inglesa, podemos citar a pesquisa de Shortall (2007), que fez uma tentativa de determinar se as evidências encontradas no livro didático analisado e no *corpus* estavam em conflito e até que ponto a gramática do material subrepresenta o uso da língua na comunicação do mundo real. Além disso, o autor discute até que ponto é justificável ignorar as evidências do *corpus* em favor de propósitos pedagógicos.

Encontramos também a pesquisa de Koprowski (2005); apesar de hoje em dia ser comum encontrarmos colocações, *phrasal verbs*, expressões idiomáticas e fixas e outras nos textos de ILE, a análise do autor sobre a utilidade de *chunks* observados em três livros didáticos contemporâneos concluiu que os autores podem ter feito um trabalho não satisfatório em sua seleção, visto que o processo de seleção foi altamente subjetivo e conduzido sem dados provenientes de *corpus*, sendo que eles se valeram da intuição, experiência e senso comum. O autor ainda chama a atenção para o fato de que enquanto aprender *chunks* pode ser algo desejável, é concebível que os alunos não estejam sendo expostos aos itens mais úteis.

Allan (2009) compilou um *corpus* de leituras simplificadas com o intuito de verificar se a autenticidade da linguagem à qual os alunos estão sendo expostos foi comprometida. A autora compara as 'porções lexicais' (*lexical chunks*) encontradas no British National Corpus (BNC) às do *corpus* de leituras simplificadas e conclui que, apesar de algumas diferenças, a frequência e o tipo de porções lexicais são suficientes para fornecer insumos que refletem a linguagem autêntica, sugerindo que as leituras simplificadas podem oferecer um equilibrio aceitável de acessibilidade e autenticidade.

Por fim, mencionamos a dissertação de mestrado de Contrera (2010), que pesquisou o emprego de lexicogramática autêntica em cinco livros didáticos para o ensino de ILE atuais e de décadas passadas, sob a perspectiva da LC. Para tanto, a autora analisou os pacotes lexicais no *corpus* de estudo (MDs), contrastando-os com os *corpora* de referência BNC e Google Corpus com o intuito de verificar quais são os livros compostos por um grau de autenticidade linguística superior em relação aos demais investigados. Por fim, a autora conclui que a LC pode mostrar ao pesquisador resultados que vão de encontro à sua intuição, visto que a autora acreditava que os livros mais atuais teriam lexicogramática mais autêntica. No

entanto, a análise mostrou que mesmo os livros de abordagem audiolingual, com textos visivelmente não autênticos, contêm lexicogramática que pode ser considerada autêntica.

Como a de Contrera, nossa pesquisa analisou os trigramas convergentes e divergentes do MD em relação aos *corpora* de referência. No entanto, nossa pesquisa difere daquela em relação à inclusão da análise dos pacotes lexicais altamente frequentes no *corpus* de referência, bem como a classificação dos achados pela frequência (em subusados, de uso equivalente e sobreusados no MD) e criação de um *corpus Baseline* para verificação da faixa de representatividade. Além disso, esta pesquisa ainda analisou os *subcorpora* falado e escrito do *corpus* de referência, separadamente.

Já no ensino de PLE, encontramos várias pesquisas que analisam os materiais didáticos disponíveis no mercado. Para citar algumas, temos a pesquisa de Júdice (2008), que analisou as representações do Brasil nos anos 1940 e 1990 e Furlan (2008), que analisou quem são os povos do Brasil nos livros didáticos para o ensino de PLE. No entanto, encontramos apenas duas pesquisas que utilizam-se de *corpus* para análise: Cavalcante (2006) e Carvalho (2007). (ver subseção 1.1.5.2 – Uso de *corpus* em pesquisas no ensino de Português como Língua Estrangeira).

Esta pesquisa desenvolveu e aplicou uma metodologia automatizada de análise (‘*corpus approach*’) que, em síntese, baseia-se na análise da lexicogramática dos textos envolvidos, promovendo a comparação dos trigramas e pacotes lexicais presentes no MD com os do *corpus* de referência (Banco de Português – BP) para a identificação do grau de autenticidade.

Sendo assim, seguimos a recomendação de Sinclair (1991: 39 apud Koprowski, 2005: 331), que indica que “os autores de materiais precisam ter dados comprovados de linguagem como seu ponto inicial. Se não for possivel, eles devem pelo menos confirmar seus dados baseados em intuição por meio de um *corpus*”[7].

7. “Materials writers need to begin with attested language data as their starting point. If this is too much to ask, then course designers might at least confirm their intuitively-based data with a corpus.”

1.1.5 Estado da arte

1.1.5.1 Ensino de Português como Língua Estrangeira no Brasil – brevíssimo histórico

De acordo com Mateus (2008), a língua portuguesa é a quinta do mundo em número de falantes (e a terceira entre as europeias), e é a língua nacional ou oficial em sete países espalhados por quatro continentes: Brasil, Portugal, Angola, Moçambique, São Tomé e Príncipe, Guiné-Bissau, Cabo Verde e Timor-Leste (e Macau até 2049). Filho (2006, apud Souza et al., 2008) afirma que o português é a quarta língua mais usada na internet, superando, por exemplo, os números referentes ao alemão, francês e italiano. Além disso, o Brasil constitui a economia maior e mais dinâmica, bem como é responsável pela grande maioria dos falantes de língua portuguesa.

No entanto, a história do ensino de português como língua estrangeira, conjugada a de seu material didático, é relativamente recente. De acordo com Amado (2008), o ensino de português para estrangeiros teve início em Portugal, no ano de 1934, com a primeira turma matriculada na Universidade de Lisboa. No Brasil, entretanto, o ensino somente teve início na década de 1950, sendo que, de acordo com Gomes de Mattos (1997), a quase totalidade dos pouquíssimos cursos de Português do Brasil oferecidos nessa época dependiam de textos escritos no exterior. O primeiro livro conhecido para ensino de português do Brasil para estrangeiros foi o *Spoken Portuguese*, produzido em 1946 nos Estados Unidos por um ítalo-americano, Vincenzo Cioffari. Na mesma época (1954), foi elaborado, aqui no Brasil, o material *Português para estrangeiros*, de Mercedes Marchant, da PUC-RS. Os materiais didáticos seguintes foram publicados somente nas décadas de 1960 e 1970: *Modern Portuguese* (1966, edição experimental), de uma equipe binacional na Universidade do Texas, em Austin, cuja edição comercial saiu em 1971; *Português contemporâneo 1*, de Abreu e Rameh; *Português: conversação e gramática*, de Magro e De Paula; e *Português 1*, da editora Berlitz (Morita, 1998). Tais livros apresentavam, em termos teóricos, o estruturalismo, em vigor naquela época, e, em termos práticos, exercícios com ‘drills’ (atividades com estratégia de repetição), textos não autênticos, assim como instruções e explicações gramaticais, geralmente em inglês.

O maior número de livros surgiu na década de 1980, com o aumento do número de estrangeiros no país. De lá para cá, não houve um acréscimo significativo em termos numéricos, sendo que hoje há cerca de 20 livros disponíveis no mercado (ver Anexo 3). Além disso, na maioria deles ainda há ênfase na gramática, fundamentação em textos não autênticos e conteúdos descontextualizados.

Com relação aos cursos de português para estrangeiros e ao seu público, grande parte do ensino no Brasil se dá em escolas privadas de línguas e em universidades. Nas escolas de línguas, o público é composto, em sua maioria, por executivos de diversas nacionalidades, a serviço de multinacionais, e suas esposas. Já nas universidades, o público é composto por estudantes em intercâmbio, vindos principalmente de países da América Latina e da África, mas, também, da Europa, Estados Unidos, Canadá e Ásia (Coreia e Japão).

1.1.5.2 Uso de corpus *em pesquisas no ensino de Português como Língua Estrangeira*

Quanto às pesquisas sobre o ensino de PLE no Brasil que utilizam *corpora*, temos um artigo de Berber Sardinha (1999) que é, até onde sabemos, o primeiro que trata de *corpora* e ensino de línguas. Esse artigo apresenta os resultados da exploração de um *corpus*, coletado a partir de notícias distribuídas pela internet, para o ensino de português do Brasil na Grã-Bretanha. O *corpus* foi o primeiro usado para lecionar PLE com a metodologia da LC, sendo que as informações retiradas da análise do *corpus* foram utilizadas para ilustrar, expandir e questionar as informações dadas nos materiais de referência, tais como gramáticas, livros-texto e dicionários. O autor argumenta que a principal motivação para usar um *corpus* em vez dos materiais existentes para ensino de PLE é que estes geralmente são baseados em exemplos inventados. Além disso, relatos anteriores quanto ao uso de *corpus* no ensino demonstraram que expor os alunos ao material de *corpus* trouxe benefícios importantes, visto que adotar concordâncias como uma técnica para exploração do *corpus* com alunos oferece a eles a oportunidade de fazer parte de atividades de descoberta que os torna pesquisadores ativos criando suas próprias explicações, que são mais bem aprendidas do que as regras prontas do livro-texto.

O autor conclui que o tipo de suporte disponível em materiais de referência existentes como livros didáticos, gramáticas e dicionários tende a ser inadequado para o aluno de português, já que ele não se baseia em amostras autênticas de linguagem como aquelas proporcionadas por um *corpus* eletrônico. Além disso, o autor acrescenta que, apesar de seu tamanho relativamente pequeno, o *corpus* forneceu evidências detalhadas para vários padrões, e essas evidências não estão disponíveis à intuição dos professores nativos.

Berber Sardinha (1997, comunicação pessoal)[8] demonstra como utilizar *corpora* para o ensino de línguas estrangeiras, em especial no ensino de PLE. Ele menciona os *corpora* de português existentes até aquele momento, bem como o *corpus* coletado e utilizado por ele e as ferramentas da LC a fim de demonstrar como é possível utilizar *corpora* para ensinar e explorar outras línguas além do inglês. Além disso, o autor menciona as vantagens (Johns, 1994) e limitações (Widdowson, 1991) de ensinar com *corpus* e concordâncias.

Além de Berber Sardinha, encontramos a dissertação de mestrado de Cavalcante (2006), que analisou a linguagem usada no material didático para ensino de PLE *Bem-vindo! A língua portuguesa no mundo da comunicação* (Ponce et al., 2003) e se ela corresponde, em termos de frequência de tempos e modos verbais, à linguagem falada e escrita no Brasil. Para isso, a autora contrastou os tempos e modos verbais presentes nas dez primeiras unidades do livro didático com aqueles usados no *corpus* Banco de Português (BP), do projeto Direct do LAEL/PUC-SP (a versão 1, menor, do mesmo *corpus* usado em nossa pesquisa) para verificar até que ponto a linguagem no livro apresentava-se em sintonia com o uso que os falantes nativos faziam dela. Para isso, Cavalcante fez uso de três *corpora*: um *corpus* de estudo (10 primeiras unidades do livro didático), um *corpus* de referência (BP) e um terceiro (BP etiquetado). A análise dos *corpora* indicou diferenças importantes entre a maneira como os verbos são apresentados no livro didático e como são usados pelos brasileiros. Os resultados indicaram que o livro didático apresentou tempos e modos verbais que não condizem com o português do Brasil, com tempos e modos verbais pouco comuns ganhando muito destaque. Assim, a autora conclui que aquilo que o

8. Concordancing Portuguese (1997) – apresentação em PowerPoint.

livro mostra como sendo a língua portuguesa não corresponde necessariamente à realidade do uso.

Carvalho (2007) analisou livros didáticos para o ensino de português para estrangeiros com relação à imagem da identidade brasileira construída, ou seja, a intenção da autora era responder quem são e o que fazem os brasileiros que os estrangeiros vão conhecer por meio do livro didático. Para isso, a autora analisou quantitativa e qualitativamente o vocabulário de oito livros didáticos com relação à identidade social e grupos sociais (etnias, raças, identidades regionais e atividades profissionais). Para isso utilizou-se de um *corpus* dos materiais didáticos selecionados e, como conclusão, acredita que se faz necessário uma maior preocupação por parte dos autores quanto à imagem do Brasil construída no livro didático, bem como uma seleção mais apurada dos textos a serem incluídos.

Temos, também, a pesquisa de Dell'sola (2002), que discute como os recursos disponíveis na internet podem ser utilizados como fontes de informação úteis no aprendizado da língua portuguesa. Além disso, a autora menciona a criação de um CD-ROM pela Universidade do Texas desenvolvido para o ensino de vocabulário comercial para aprendizes de PLE em nível intermediário ou avançado. Esse material foi lançado em 2000 e contém vídeo e transcrição de entrevistas com 27 falantes nativos de diferentes regiões brasileiras. Nessas entrevistas, as seguintes áreas são tratadas: Contabilidade, Propaganda, Banco, Organização e Estruturas de Empresas, Economia, Finanças, Recursos Humanos, Seguro, Investimento, Vendas, Bolsa de Valores, Comércio Internacional e Sindicatos. O CD-ROM também contém a transcrição e tradução para o inglês das entrevistas, seguidas de uma lista contendo os termos usados nas entrevistas. A autora afirma que além de oferecer ao aprendiz de PLE informação sobre negócios e vocabulário técnico, o material coloca esse aprendiz em contato com a fala autêntica de brasileiros que dominam o assunto e apresentam suas opiniões reais sobre os temas em sua língua materna.

A pesquisa realizada por Alencar (2004) teve como objeto de estudo o uso das expressões formulaicas e sua importância na descrição do PLE. Em seu trabalho, o autor percebe que essas rotinas conversacionais são utilizadas com frequência, principalmente na linguagem oral. No entanto, vê a necessidade de se definir critérios ou procedimentos para identificá-las e para verificar que espaço tais

expressões ocupam na descrição do PLE. O autor analisa alguns materiais didáticos com relação às expressões formulaicas e percebe que os materiais disponíveis no mercado, quando apresentam tais expressões, fazem menção incipiente, sendo que o máximo que o autor encontrou foram listas de estruturas e expressões que não apresentam uma organização clara para o aprendiz nem uma proposta de trabalho.

Como *corpus*, o autor utilizou-se dos diálogos transcritos da série *Os Normais*[9]. A pesquisa mostrou que há no português do Brasil uma grande quantidade de expressões que possuem uma função específica dentro da comunicação cotidiana. Dessa constatação, originou-se a identificação e sistematização das expressões formulaicas contidas no *corpus* proposto para que elas possam ser compreendidas e utilizadas com tranquilidade por professores e alunos.

Até onde sabemos, não há nenhuma pesquisa que enfoca a análise de autenticidade em MDs para o ensino de PLE (com ou sem utilização de *corpus*).

1.2 Autenticidade e Ensino de Línguas

De acordo com Breen (1995), há quatro tipos de autenticidade:

- Autenticidade dos textos;
- Autenticidade da interpretação de tais textos pelos aprendizes (ou seja, autenticação/validação dos textos pelos alunos);
- Autenticidade das tarefas; e
- Autenticidade da situação social da sala de aula (ou seja, exploração da sala de aula como um local no qual os participantes possam, juntos, dividir seus problemas, conquistas e processo de aprendizagem).

Nesta pesquisa, trabalhamos com as definições de textos autênticos e não autênticos, que serão discutidas na próxima subseção.

9. Exibida pela Rede Globo de Televisão de 2001 a 2003.

1.2.1 Materiais autênticos e não autênticos

Cada vez mais se fala no uso de materiais autênticos para o ensino de idiomas e, hoje, todos concordam que seu uso em sala de aula é benéfico para o processo de aprendizagem (Guariento e Morley, 2001; Berber Sardinha, 2007; Shortall, 2007), embora essa prática nem sempre tenha sido unânime. No entanto, por ser um conceito abstrato, há muita divergência no que pode ser considerado autêntico e o assunto é bastante discutido entre os pesquisadores da área. Fizemos uma breve pesquisa com alguns professores de PLE sobre o que eles consideram textos autênticos. Em linhas gerais, esses professores acreditam que autenticidade diz respeito a algo real, natural e que não sofreu alterações. O *Dicionário Aurélio*, por sua vez, além de fornecer como possível acepção para 'autêntico' algo que é verdadeiro e real, traz também a ideia de algo que é legalizado e autenticado, ou seja, no caso do ensino de línguas estrangeiras, para ser autêntico é necessário que algo seja validado por alunos e professores.

Para Berber Sardinha (2007), assim como para muitos linguistas, um texto autêntico é aquele que não foi criado com a finalidade de ensinar língua, sendo que possui todos os defeitos e virtudes da vida real. Essa definição vai ao encontro da de Morrow (1977: 13 apud Taylor, 1994: 4) que afirma que "um texto autêntico é um prolongamento da linguagem real, produzido por falantes nativos, para um público real e elaborado para transmitir uma mensagem real"[10]. Definição semelhante é dada por Nunan (1989: 54 apud Taylor, 1994: 4) que afirma que podemos considerar autêntico "qualquer material que não foi elaborado para o propósito de ensinar a língua em questão"[11].

Berber Sardinha (2007) ainda acrescenta que muitos livros didáticos geralmente não se utilizam de textos autênticos principalmente por sentirem necessidade de controlar o vocabulário e a gramática do conteúdo do curso, com base no conceito de que um texto torna-se mais adequado na medida em que incorpora apenas certa quantidade ou tipo de vocabulário e/ou de estruturas gramaticais.

---

10. "An authentic text is a stretch of real language, produced by a real speaker or writer for a real audience and designed to convey a real message of some sort."
11. "(…) Any material which has not been specifically produced for the purposes of language teaching."

Brown e Menasche (2006) propõem graus de autenticidade em vez de posicionar os textos como autênticos ou não autênticos. Eles sugerem cinco níveis de autenticidade, que vão desde 'autenticidade genuína', 'autenticidade alterada', 'autenticidade adaptada', 'autenticidade simulada' até 'inautenticidade'[12]. Os autores defendem esses vários níveis de autenticidade porque acreditam que é difícil caracterizar os textos simplesmente como autênticos ou inautênticos e, na prática, em sala de aula, a autenticidade completa é impossível de ser atingida.

Mishan (2004), por outro lado, faz uma distinção entre textos autênticos e autenticidade do uso da língua, ou seja, como o aluno se relaciona com o texto e com a atividade realizada. Esse conceito vai ao encontro do de Breen (1985), que afirma que a autenticidade deve ser considerada resultado da interdependência entre textos, aprendizes, tarefas de aprendizagem e situação social da sala de aula. De acordo com Breen (1985), há um conjunto de fatores que precisam ser levados em conta, inclusive a validação/autenticação do aluno. Ele afirma, ainda, que o que é autêntico é relativo aos nossos propósitos e aos pontos de vista dos diferentes participantes na sala de aula e que a questão da autenticidade de um texto é quase inseparável do questionamento de para quem esse texto é autêntico.

Quanto aos benefícios da utilização de textos autênticos, Mishan (2004) afirma que os textos autênticos fornecem a melhor fonte de insumos ricos e variados para aprendizes de idiomas, têm impacto nos fatores afetivos essenciais para o aprendizado, como a motivação, a empatia e o envolvimento emocional e resultam em um aprendizado mais duradouro. Para Wilkins (1976: 79), "o uso de textos autênticos, tanto escritos quanto falados, ajuda a fazer uma ponte entre o conhecimento em sala de aula e 'a capacidade do aluno em participar de eventos da vida real'" (apud Guariento e Morley, 2001: 347)[13]. Eles dão aos alunos o sentimento de que estão aprendendo a língua "real", que estão em contato com uma entidade viva, a língua-alvo como ela é usada pela comunidade que a fala. No entanto, os autores acreditam que a simplificação bem feita dos textos pode ser usada,

12. 'Genuine input authenticity'; 'altered input authenticity'; 'adapted input authenticity'; 'simulated input authenticity'; e 'inauthenticity'.
13. "The use of authentic texts, embracing both the written and spoken word, is helping to bridge the gap between classroom knowledge and 'a student's capacity to participate in real world events'."

especialmente em níveis mais inferiores, se quisermos obter respostas autênticas nos alunos.

Com relação aos textos não autênticos, Berber Sardinha (2007) afirma que seriam aqueles que em geral possuem exemplos “fictícios” e frases vazias de sentido e descontextualizadas, mas bem construídas e corretas gramaticalmente, existentes somente em escolas de idiomas e utilizadas para manipulação gramatical – úteis na escola, mas que não preparam os alunos para a língua efetivamente usada fora da sala de aula. De acordo com Shortall (2007) os textos autênticos, diferentemente dos não autênticos encontrados em materiais didáticos, têm o entusiasmo da comunicação real e não a esterilidade dos diálogos elaborados para ilustrar padrões gramaticais. Por outro lado, o autor acredita que o uso de linguagem não autêntica nos materiais didáticos deve-se ao fato de que o material autêntico nem sempre corresponde a um material sistematicamente tratável para ensino como o não autêntico. Um exemplo disso seria utilizar textos transcritos de uma conversa em um material didático. Diferente da linguagem autêntica, o autor acredita que a linguagem não autêntica dos materiais didáticos parece ser mais acessível aos alunos, mais sistemática na sua apresentação gradual e mais fácil de ensinar. No entanto, pelo menos no ensino de ILE, aparentemente os materiais didáticos mais atuais estão começando a incorporar mais características do discurso natural em seus diálogos não autênticos (Gillmore, 2004).

Para esta pesquisa e coleta do *corpus* de estudo, embasamo-nos nas definições propostas por Berber Sardinha (2007) e Nunan (1989) de que podemos considerar como autêntico qualquer material que não foi elaborado com propósitos pedagógicos. Sendo assim, entendemos serem textos não autênticos aqueles exemplos de linguagem elaborados para utilização em sala de aula e nosso *corpus* de estudo (MD), compilado para esta pesquisa, foi dividido em dois *subcorpora*: textos autênticos e não autênticos, com base nessas definições.

Sendo assim, seguindo os achados de Allan (2009) e Contrera (2010), acreditamos que os textos não autênticos podem conter elementos característicos da autenticidade, ou seja, os textos podem ser mais ou menos autênticos, e não simplesmente autênticos ou não autênticos. Sendo assim, pretendemos verificar nesta pesquisa o grau de autenticidade dos textos do material didático estudado

com base nos padrões lexicogramaticais. O estudo do grau de autenticidade sustenta-se na visão de linguagem como sistema probabilístico, pois a ideia de grau implica que há probabilidade de textos não autênticos terem maior ou menor aproximação com a linguagem atestada em *corpora* eletrônicos. Nossa proposta é justamente verificar o grau de aproximação observado nos textos do material didático, que contém tanto textos (orais e escritos) autênticos quanto não autênticos.

1.2.2 Idiomaticidade

De acordo com Hunston (2002: 136) as técnicas de *corpus* são usadas para resolver problemas da vida real e os métodos podem ser resumidos em:

- observar as frequências da ocorrência;
- observar as regularidades das co-ocorrências;
- observar as regularidades do uso.

A partir dessas observações de frequência e regularidade no *corpus* podemos chegar à identificação de padrões. De acordo com Berber Sardinha (1999: 294), a frequência dos itens não está disponível aos falantes nativos por meio da introspecção, e precisa ser obtida por meio de um *corpus*. Como Sinclair e Renouf (1988: 151) comentaram, essa característica é comum a todos os usuários de qualquer língua:

> O ser humano, ao contrário da crença popular, não é bem organizado para isolar, de maneira consciente, o que é central e típico de uma língua; qualquer coisa fora do comum é claramente percebida, mas os eventos rotineiros são apreciados de maneira subliminar[14].

Assim, normalmente, é muito mais fácil notar quando algo nos soa estranho ou incomum, como quando um aluno diz algo não condizente com o padrão. No entanto, é difícil percebermos o que é mais comum e o que devemos ou não ensinar aos alunos, visto que os padrões que podemos achar relevantes podem se mostrar nada significativos quando confrontados no *corpus*. Desse modo, a observação dos padrões é tida como de suma importância no ensino de língua estrangeira, pois a

14. “The human being, contrary to popular belief, is not well organized for isolating consciously what is central and typical in the language; anything unusual is sharply perceived, but the humdrum everyday events are appreciated subliminally.”

sensação de ‘naturalidade’ na fala ou na escrita depende em grande parte do emprego de padrões (Fox 1998: 33 apud Berber Sardinha, 2000: 4).

Como Sinclair (1991: 108) observou:

> A maior parte do texto é composta de palavras comuns em padrões comuns ou em leves variações desses padrões comuns. A maior parte das palavras mais frequentes não tem sentido(s) independente(s), mas são componentes de um rico repertório de padrões de multipalavras que fazem um texto. Isso é totalmente desconhecido dos procedimentos da gramática convencional[15].

A idiomaticidade, ou ‘princípio idiomático’ (*idiom principle*), de acordo com Sinclair (1991: 110) está relacionada ao vasto número de combinações pré-existentes que constituem escolhas únicas e que estão disponíveis ao usuário de uma língua. Diferentemente do ‘princípio da livre escolha’ (*open-choice principle*) que vê os textos como uma série de lacunas que podem ser preenchidas virtualmente com qualquer item lexical, a idiomaticidade sugere que as palavras tendem a se combinar de acordo com um limitado número de escolhas. Assim, as escolhas mais originais ou idiossincráticas tendem a soar menos ‘naturais’ do que as combinações de alta frequência.

Sendo assim, a idiomaticidade refere-se a quão ‘natural’ soa um texto (Sinclair, 1991 apud Berber Sardinha, 2007), o que não tem a ver com a gramaticalidade (um texto oral, por exemplo, com vários problemas gramaticais pode soar bastante natural). Em outras palavras, a idiomaticidade é uma característica da linguagem autêntica e pode ser definida como um conjunto de fatores que a tornam natural, normalmente associados à escolha das combinações lexicogramaticais empregadas.

No entanto, a idiomaticidade não pode ser entendida adequadamente por meio de nossa experiência, intuição ou conhecimento de língua. Assim,

> quando *produzimos* nossa língua materna, em um grande número de gêneros com que estamos familiarizados, temos perfeito comando inconsciente da idiomaticidade; porém, quando *analisamos* conscientemente a idiomaticidade, nossa intuição é pouco confiável (Sinclair, 1991 apud Berber Sardinha, 2007: 4).

---

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

Sendo assim, uma das maneiras para inferir a idiomaticidade das escolhas lexicais, de acordo com Berber Sardinha (2007: 277-278), é por meio da quantidade de 'pacotes lexicais' presentes no texto. Um pacote lexical (ver seção 1.3 – Pacotes lexicais), em linhas gerais, é uma sequência de palavras fixas, de extensão variável, muitas vezes chamada de '*chunk*' (Lewis, 2000) ou '*cluster*' (Scott & Tribble, 2006). Esses padrões podem ser bem formados ou não, como "bom-dia" ou "que possui um", e devem ocorrer com certa frequência para ser considerados pacotes lexicais.

Dessa maneira, recorremos a *corpora* eletrônicos em busca dos padrões que pudessem fornecer evidências da autenticidade (ou não) do material didático, visto que se espera que a frequência e a quantidade dos padrões encontrados reflita o uso real da linguagem.

1.3 'Pacotes Lexicais' *(Lexical Bundles)*

De acordo com Biber et al. (1999), existem diferentes tipos de expressões multipalavras e estas se distinguem de acordo com sua idiomaticidade e invariabilidade. Em um extremo temos as expressões idiomáticas, as quais são expressões relativamente fixas com sentidos que não podem ser depreendidos de suas partes. O exemplo clássico da língua inglesa, de acordo com Tagnin (2005), é a expressão *kick the bucket*, que, em português, não corresponde a "chutar o balde", mas sim a "morrer". Em português existe uma expressão idiomática correspondente, i. e., "bater as botas".

Além das expressões idiomáticas, há os casos de combinações lexicais consagradas, de duas ou mais palavras de conteúdo, os quais o linguista J. R. Firth denominou '*collocations*', ou colocações em português. As colocações são palavras que geralmente "andam juntas", que parecem combinar-se naturalmente, sem ter uma explicação para tal fato, tais como "açúcar mascavo", "praça pública", "criar problemas" e "acreditar cegamente" (Tagnin, 2005). De acordo com Biber et al. (1999), diferentemente das expressões idiomáticas, as colocações são associações estatísticas que tendem a co-ocorrer em conjuntos específicos de colocados em vez de expressões relativamente fixas.

Já as coligações, ainda de acordo com Tagnin (2005), são combinações consagradas de elementos linguísticos em que o colocado, ou seja, a palavra que não conhecemos ou que não nos ocorre de imediato e que é determinada pela base, é gramatical. Exemplos de coligações são “obedecer a”, “cumpridor de” e “bom em”.

No entanto, a base da presente pesquisa são os padrões que co-ocorrem em sequências mais longas, chamados de ‘pacotes lexicais’ (*lexical bundles*). De acordo com Biber et al. (1999), os pacotes lexicais são sequências de três ou mais palavras que mostram uma tendência estatística de co-ocorrerem juntas em determinados tipos de textos e, na maior parte dos casos, não são unidades estruturais completas (por exemplo, “a ver com” e “acordo com a”) nem expressões que os falantes reconheceriam como idiomáticas ou fixas.

Os pacotes lexicais são definidos por sua frequência e, para ser considerado um pacote lexical recorrente, a combinação de palavras tem de ocorrer, pelo menos, dez vezes por milhão de palavras. Além disso, somente as combinações ininterruptas (não divididas por pontuação ou trocas de turno) podem ser tratadas como pacotes lexicais em potencial (Biber et al., 1999).

Quanto às diferenças em frequência entre as expressões idiomáticas e os pacotes lexicais, Biber et al. (1999) constataram, no *corpus* em inglês por ele utilizado, que todos os pacotes pesquisados são muito mais comuns do que as expressões idiomáticas (raras, por exemplo, em conversações). Sendo assim, ao transferirmos essa constatação para o ensino de línguas, percebemos que os alunos podem estar tentando dominar expressões idiomáticas raras como *in a nutshell* e *beat about the bush* em vez de padrões que eles realmente irão precisar no dia a dia.

Encontramos vários estudos sobre pacotes lexicais, no entanto, a maior parte é limitada a estudos na língua inglesa (Biber, Conrad e Cortes, 2004; Biber, 2006 e 2009; Cortes, 2007; Hyland, 2007 e 2008; Nekrasova, 2009; Shepherd e Viana, 2006; Shepherd, 2009; Berber Sardinha e Shepherd, 2008). Em língua portuguesa, há alguns trabalhos sobre pacotes lexicais em *corpora* de aprendiz de português como língua materna, como Shepherd et al. (2006, 2007 e s.d. [no prelo]) e um trabalho sobre pacotes lexicais usados em linguagem jornalística (Araujo, 2010). No

entanto, não é do nosso conhecimento a existência de estudos de pacotes lexicais em PLE.

Esta seção marca o final da apresentação do arcabouço teórico utilizado nesta pesquisa. O próximo capítulo apresenta a metodologia empregada na análise dos dados, bem como a descrição dos *corpora* de estudo e referência.

CAPÍTULO 2

METODOLOGIA

Neste capítulo apresentamos a metodologia empregada na pesquisa, incluindo a descrição dos *corpora*, bem como a especificação dos procedimentos de análise dos dados. Primeiramente, são detalhados os *corpora* que compuseram o estudo e, em seguida, é especificado o processo de análise e seleção dos dados e as ferramentas utilizadas.

A seguir, descrevemos o material de pesquisa, o livro didático *Muito prazer – fale o português do Brasil*[1] e, em seguida, detalhamos os procedimentos de coleta e organização do *corpus* de estudo (*corpus* MD).

2.1 Material de Pesquisa: Livro Didático

2.1.1 Apresentação do material

O objetivo do *Muito prazer – fale o português do Brasil* (MP)**,** segundo as autoras, é "capacitar o aluno, de qualquer nacionalidade, que deseja aprender o português do Brasil a comunicar-se com precisão e fluência" (Fernandes et al., 2008: 17). Para tanto, as autoras afirmam apresentar o léxico e a gramática essenciais para uma boa comunicação em português, por meio de atividades estimulantes e contextualizadas, que apresentam a linguagem em uso na comunicação dos brasileiros.

O material constitui um curso para alunos de nível iniciante e intermediário e também pode ser utilizado por autodidatas. Além disso, as autoras afirmam que os exemplos e atividades elaborados a partir da linguagem corrente do português do Brasil procuram mostrar como certas palavras e expressões se comportam em

1. As demais coautoras autorizaram o uso do livro e a citação nominal de seu titulo nesta pesquisa.

determinados contextos. O material conta com dois CDs de áudio e seu roteiro de gravação encontra-se no fim do livro.

Visto que o material não foi elaborado com base em *corpus* (apesar de haver textos autênticos em partes do material), uma das dúvidas que deu origem a esta pesquisa é relativa à aproximação do material não autêntico da linguagem atestada em *corpora* eletrônicos.

2.1.2 Orientações teóricas do MP

No MP foram consideradas as Abordagens Comunicativa e Lexical abaixo detalhadas.

2.1.2.1 Abordagem Comunicativa

O '*Communicative Language Teaching*' (CLT) ou Ensino Comunicativo, em português (também conhecido como 'Abordagem Comunicativa'), de acordo com Richards e Rodgers (2001), marca o início de uma grande mudança de paradigma no campo do ensino de idiomas, no século XX, e suas ramificações podem ser percebidas ainda hoje. Os princípios gerais do CLT hoje são amplamente aceitos no mundo todo.

A grande aceitação dessa abordagem e o modo relativamente variado com que é interpretada e aplicada podem ser atribuídos ao fato de que os praticantes de diferentes tradições educacionais conseguem identificar-se com ela e, consequentemente, interpretá-la de várias maneiras, partilhando, no entanto, da mesma teoria de ensino de língua estrangeira.

A Abordagem Comunicativa tem como ponto central a visão de linguagem como comunicação (Richards e Rodgers, 2001). O objetivo do ensino é desenvolver a competência comunicativa. Isso significa que o aluno adquirirá conhecimento e habilidade para usar a língua de acordo com o contexto, escolhendo o que é mais

adequado. Os proponentes dessa abordagem veem o aprendizado de idiomas como a aquisição de meios linguísticos para realizar variadas funções.

Como visão de linguagem, pode-se afirmar que o Ensino Comunicativo possui uma base teórica rica e, até certo ponto, eclética. Algumas características dessa visão são:

1. A língua é um sistema usado para expressar significados.
2. A principal função da língua é permitir interação e comunicação.
3. A estrutura da língua reflete seus usos funcionais e comunicativos.
4. As unidades principais da língua não são somente suas características gramaticais e estruturais, mas também categorias de significado funcional e comunicativo, como podem ser observadas no discurso.

Ainda de acordo com os autores, apesar de haver vasta bibliografia sobre a visão de linguagem no Ensino Comunicativo, pouco foi escrito sobre a teoria de aprendizado. Entretanto, podemos discernir alguns elementos teóricos em algumas práticas comunicativas. Por exemplo, acredita-se que atividades que envolvem comunicação real e aquelas em que a língua é usada na realização de tarefas significativas promovem o aprendizado. Outro elemento teórico que pode ser identificado é a crença de que a linguagem que é significativa para o aluno apoia o processo de aprendizado. Assim, as atividades são escolhidas de acordo com quão bem proporcionam ao aluno o uso autêntico e significativo da língua (em vez de prática meramente mecânica de padrões).

2.1.2.2 Abordagem Lexical

A Abordagem Lexical, desenvolvida por Lewis (1997), pode ser resumida deste modo: a linguagem não consiste em gramática tradicional e vocabulário, mas em porções (‘*chunks*’) pré-fabricadas de mais de uma palavra que, quando combinadas, produzem um texto coerente e contínuo. Essas porções, de acordo com o autor, são sequências de palavras que constituem maneiras naturais ou comuns de expressar

ideias ou propósitos específicos pelos falantes nativos. Há várias combinações de palavras diferentes que podem expressar uma mensagem, mas há somente uma ou duas dessas combinações que são normais e naturais e estas são as que devemos ensinar aos nossos alunos. Portanto, o foco principal desta abordagem é a crença de que os alunos necessitam aprender uma grande quantidade dessas combinações ou porções e o autor identifica três tipos básicos. São eles:

- colocações (p. ex., “cão e gato” – e não “gato e cachorro”, “pão-duro”, “vinho tinto”);
- expressões fixas (“Muito Prazer!”, “De nada!”); e
- expressões semifixas (“Como eu ia dizendo...”, “Uma salva de palmas para...”).

A abordagem promove a atenção dos alunos para essas sequências de blocos pré-fabricados e os encoraja a manter anotações dessas palavras e expressões em seus ‘cadernos lexicais’ (*lexical notebooks*). Além disso, mais atenção será dada:

- ao léxico – diferentes tipos de porções de mais de uma palavra;
- à compreensão auditiva (em níveis mais básicos) e à leitura (em níveis mais avançados);
- ao português provável e não ao português possível. Por exemplo, a combinação “cometer um crime” é possível e provável, mas “cometer uma boa ação” vai soar estranha, apesar de ser gramaticalmente possível;
- à organização de cadernos lexicais para revelar padrões e facilitar sua recuperação;
- à linguagem que os alunos podem encontrar fora da sala de aula; e
- ao preparo dos alunos para que eles consigam se beneficiar do texto tanto quanto possível.

2.1.3 Características das unidades

Há 20 unidades no MP e a cada 4 unidades uma unidade de revisão e outra de pronúncia são apresentadas, totalizando 10 unidades adicionais. Como dito anteriormente, foram utilizados alguns textos autênticos no material didático, em

especial os textos de leitura das últimas unidades. Na subseção 2.2.1.1 (Divisão do *corpus* MD em autêntico e não autêntico) descrevemos quais partes do MD são autênticas e quais são não autênticas.

As unidades são divididas em três lições (A, B e C) e uma parte final que as relaciona e as revisa, de acordo com o tópico principal da unidade. Cada lição (exceto da Unidade 1) é composta por:

- PANORAMA: seu objetivo é introduzir e contextualizar o assunto que será abordado, utilizando o conhecimento prévio do aluno, a fim de prepará-lo para o conteúdo que será apresentado (ver Figura 2.1).

Figura 2.1: Panorama da unidade 7, lição A.

- DIÁLOGO: os diálogos foram elaborados para tentar recriar situações da vida real no país, com uma linguagem apropriada para diferentes tipos de contextos (registros formal e informal). Por meio deles, o aluno entra em contato com as estruturas gramaticais e o vocabulário que serão praticados

nos exercícios seguintes. Além disso, o aluno terá oportunidade de praticar pronúncia e compreensão auditiva (ver Figura 2.2).

Figura 2.2: Diálogo da unidade 5, lição B.

- CONSTRUÇÃO DO CONTEÚDO: primeiramente por meio de exercícios escritos controlados e depois com exercício oral mais livre, o aluno poderá consolidar as estruturas estudadas e aplicá-las, a fim de aumentar sua competência comunicativa (ver Figura 2.3).

UNIDADE 16 IMAGINE FAZER UMA VIAGEM DE BICICLETA!

LIÇÃO C CONSTRUÇÃO DO CONTEÚDO

A. Use as palavras abaixo para substituir as palavras grifadas. Conjugue os verbos, caso seja necessário.

aproveitar conseguir reparar avisar
combinar nada disso fechado

1. A: Vamos passar uma semana no SPA?
   B: Combinado.
2. O Lúcio pediu para te informar que aquele albergue está lotado.
3. Não sei se será possível ficar em uma pousada com acesso à Internet.
4. A recepcionista notou que estamos carregando malas demais.
5. A: Quero fazer um tour de ônibus pela cidade.
   B: De jeito nenhum! Estou esgotada.
6. Por que não tiramos vantagem disso e fazemos um passeio de barco?
7. Boné não fica bonito com terno. Pelo menos não nesse país.

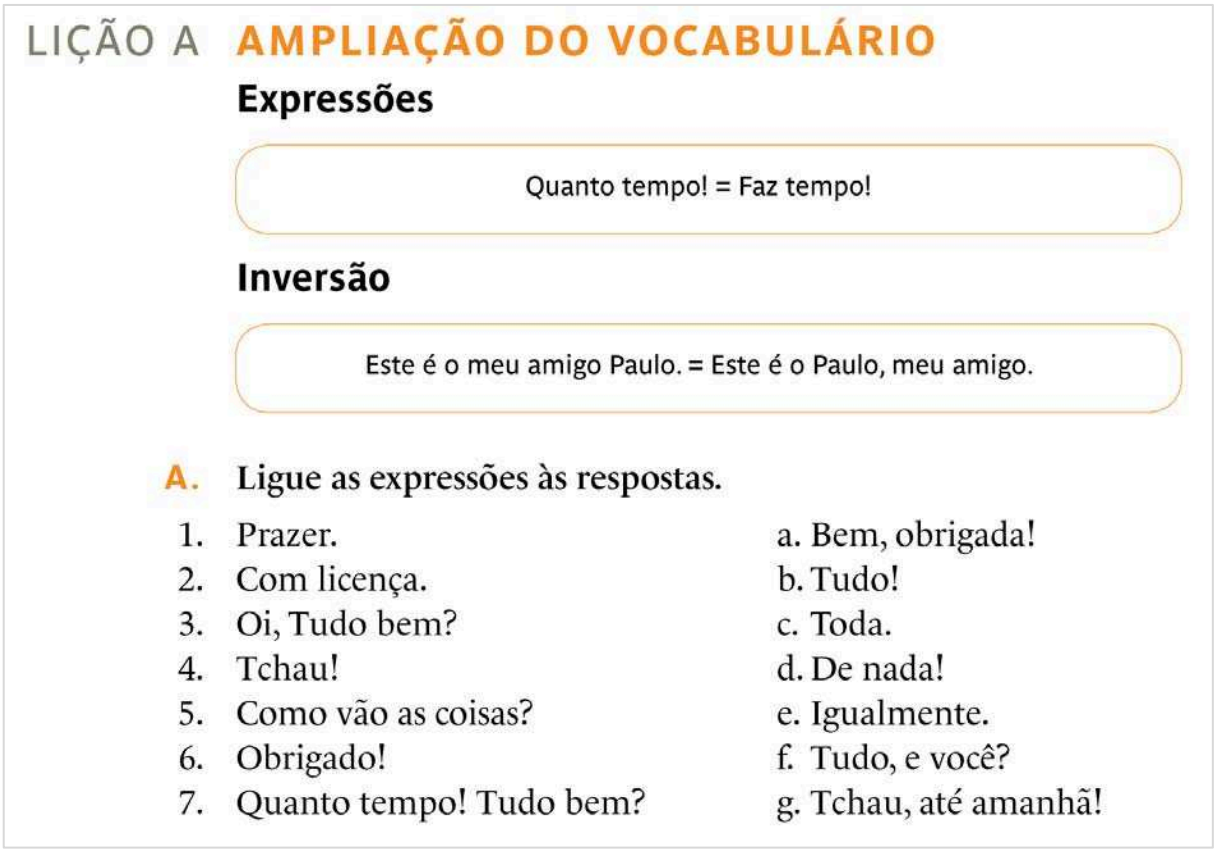
B. Complete as frases com as suas informações.

1. Em meu país, _____ (ROUPA) não combina com _____ (ROUPA).
2. Eu sempre reparo _____ quando vou às compras.
3. Às vezes quando viajo, aproveito para _____ _____.
4. Eu geralmente não consigo _____ quando estou fora de casa.
5. Fico furioso (a) quando não me avisam _____ _____.

C. Oral: Combine com um/a ou dois/duas colegas um pequeno *tour* pela sua cidade. Proponha lugares para visitar, coisas para fazer. Use as expressões "nada disso" e "combinado" para resolver os detalhes.

Figura 2.3: Construção do conteúdo da unidade 16, lição C.

- AMPLIAÇÃO DO VOCABULÁRIO: nessa seção, o aluno aprende palavras relacionadas ao assunto da lição de maneira ativa, ou seja, pode utilizá-las em exercícios orais ou reconhecê-las em exercícios de compreensão auditiva (ver Figura 2.4).



LIÇÃO A **AMPLIAÇÃO DO VOCABULÁRIO**

Expressões

Quanto tempo! = Faz tempo!

Inversão

Este é o meu amigo Paulo. = Este é o Paulo, meu amigo.

A. Ligue as expressões às respostas.

1. Prazer.	a. Bem, obrigada!
2. Com licença.	b. Tudo!
3. Oi, Tudo bem?	c. Toda.
4. Tchau!	d. De nada!
5. Como vão as coisas?	e. Igualmente.
6. Obrigado!	f. Tudo, e você?
7. Quanto tempo! Tudo bem?	g. Tchau, até amanhã!

Figura 2.4: Ampliação do vocabulário da unidade 2, lição A.

Na parte final da unidade (lições A, B e C), o aluno revê o conteúdo das três lições. Essa parte é dividida em COMPREENSÃO AUDITIVA, APLICAÇÃO ORAL DO CONTEÚDO, LEITURA, REDAÇÃO e CONSOLIDAÇÃO LEXICAL.

- COMPREENSÃO AUDITIVA: nessa seção, o aluno tem mais uma oportunidade de reconhecer e internalizar estruturas e vocabulário vistos anteriormente (ver Figura 2.5).

LIÇÕES A, B e C **COMPREENSÃO AUDITIVA**

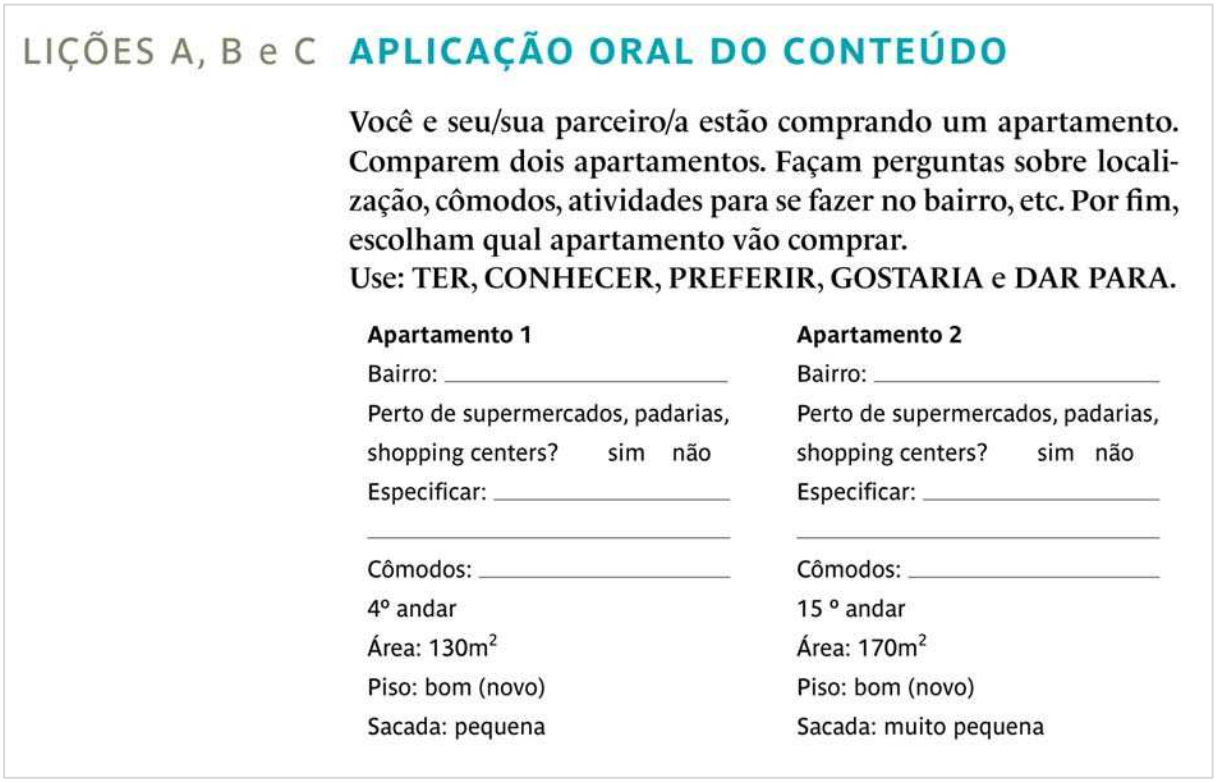
1/24 Ouça os três diálogos e responda as perguntas.

1. a. Qual a nacionalidade do marido da Geni? _____
   b. Qual a idade dele? _____
   c. Qual a língua materna dele? _____

2. a. Que horas são? _____
   b. Onde está a filha da Martinha? _____
   c. Qual a idade dela? _____

Figura 2.5: Parte da compreensão auditiva da unidade 3, lições A, B e C.

- APLICAÇÃO ORAL DO CONTEÚDO: nessa seção, o aluno, novamente, tem a oportunidade de aplicar comunicativamente o conteúdo da unidade e, dessa forma, consolida seu conhecimento e melhora sistematicamente sua fluência oral (ver Figura 2.6).



LIÇÕES A, B e C **APLICAÇÃO ORAL DO CONTEÚDO**

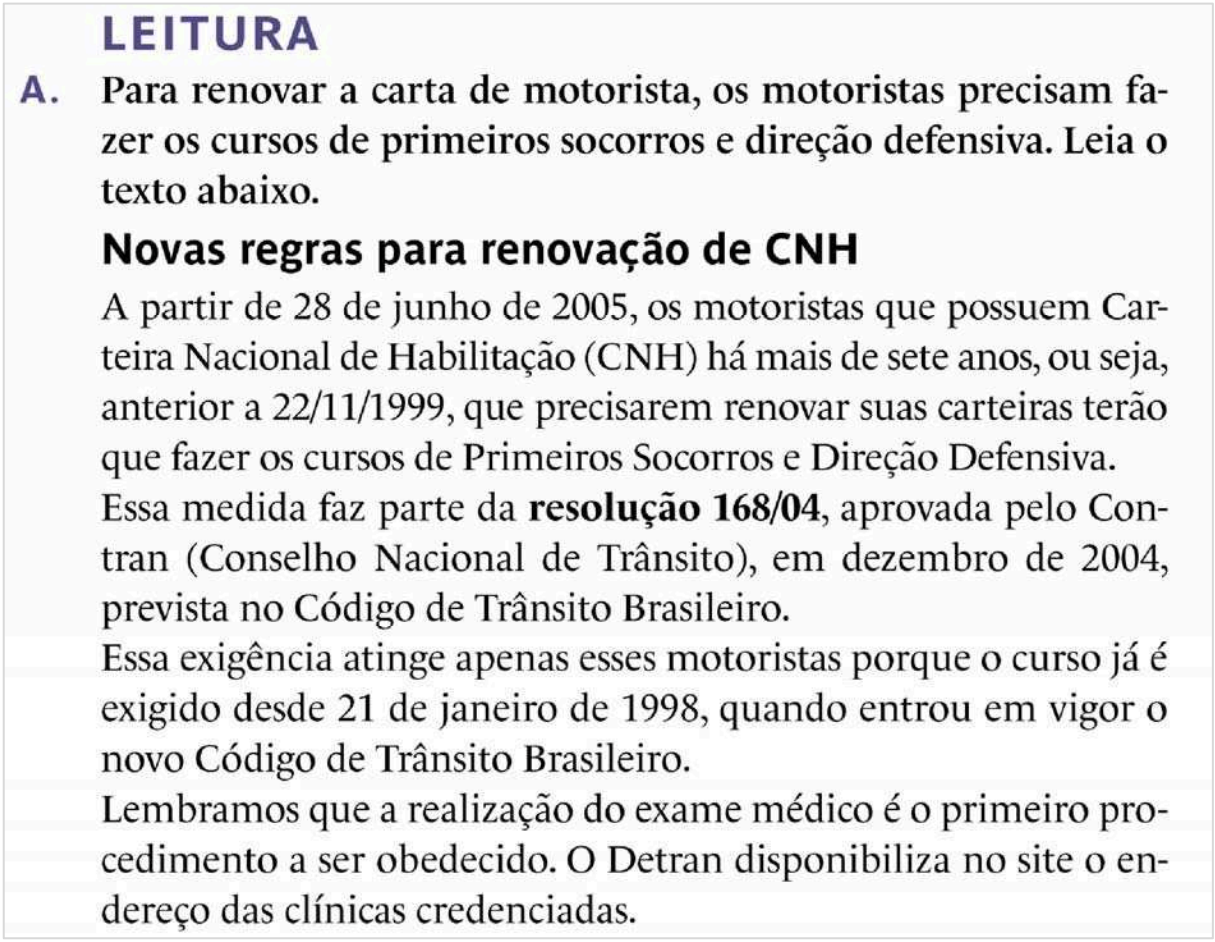
Você e seu/sua parceiro/a estão comprando um apartamento. Comparem dois apartamentos. Façam perguntas sobre localização, cômodos, atividades para se fazer no bairro, etc. Por fim, escolham qual apartamento vão comprar.
Use: TER, CONHECER, PREFERIR, GOSTARIA e DAR PARA.

Apartamento 1	**Apartamento 2**
Bairro: _____	Bairro: _____
Perto de supermercados, padarias, shopping centers? sim não	Perto de supermercados, padarias, shopping centers? sim não
Especificar: _____	Especificar: _____
_____	_____
Cômodos: _____	Cômodos: _____
4º andar	15 º andar
Área: 130m^2	Área: 170m^2
Piso: bom (novo)	Piso: bom (novo)
Sacada: pequena	Sacada: muito pequena

Figura 2.6: Aplicação oral do conteúdo da unidade 8, lições A, B e C.

- LEITURA: os textos da leitura, em sua grande maioria, foram obtidos de fontes autênticas (jornais, revistas, internet) e adaptados ao nível do conhecimento linguístico do aluno. Além dos exercícios de compreensão que os seguem, também há exercícios que fazem com que o aluno fale um pouco mais de si e de sua realidade (ver Figura 2.7).

Como dito anteriormente, as leituras das últimas unidades foram retiradas de fontes autênticas, sem adaptações ou simplificações, sendo que esses textos compõem a maior parte do *subcorpus* autêntico (MDA). Na subseção 2.2.1.1 será especificada a proporção de material não autêntico e autêntico. O não autêntico tem proporção muito maior que o autêntico e esse é mais um motivo para verificarmos a autenticidade do material.



LEITURA

A. Para renovar a carta de motorista, os motoristas precisam fazer os cursos de primeiros socorros e direção defensiva. Leia o texto abaixo.

Novas regras para renovação de CNH

A partir de 28 de junho de 2005, os motoristas que possuem Carteira Nacional de Habilitação (CNH) há mais de sete anos, ou seja, anterior a 22/11/1999, que precisarem renovar suas carteiras terão que fazer os cursos de Primeiros Socorros e Direção Defensiva.

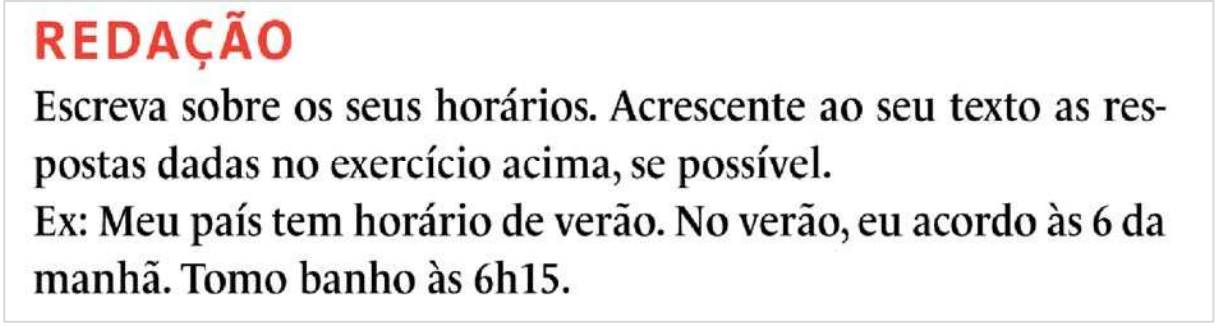
Essa medida faz parte da **resolução 168/04**, aprovada pelo Contran (Conselho Nacional de Trânsito), em dezembro de 2004, prevista no Código de Trânsito Brasileiro.

Essa exigência atinge apenas esses motoristas porque o curso já é exigido desde 21 de janeiro de 1998, quando entrou em vigor o novo Código de Trânsito Brasileiro.

Lembramos que a realização do exame médico é o primeiro procedimento a ser obedecido. O Detran disponibiliza no site o endereço das clínicas credenciadas.

Figura 2.7: Trecho da leitura da unidade 17, lições A, B e C.

- REDAÇÃO: a proposta de atividade escrita tem a finalidade de fazer com que o aluno utilize o vocabulário e a gramática aprendidos, até aquele momento, e escreva sobre um tópico visto na unidade (ver Figura 2. 8).



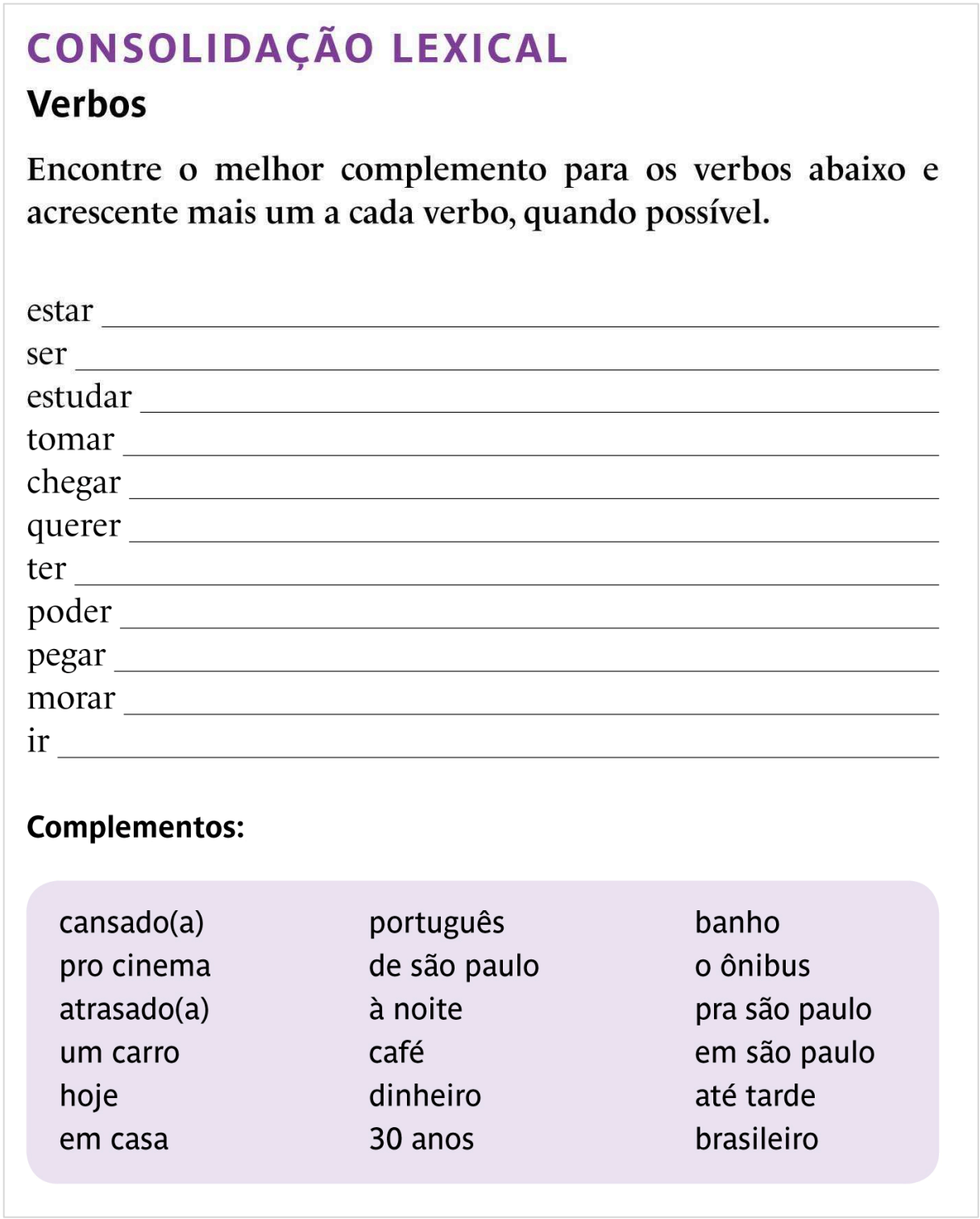
REDAÇÃO

Escreva sobre os seus horários. Acrescente ao seu texto as respostas dadas no exercício acima, se possível.

Ex: Meu país tem horário de verão. No verão, eu acordo às 6 da manhã. Tomo banho às 6h15.

Figura 2.8: Atividade de redação da unidade 3, lições A, B e C.

- CONSOLIDAÇÃO LEXICAL: inspirada na Abordagem Lexical (Lewis, 1997), essa seção tem a finalidade de organizar o vocabulário aprendido na unidade, de modo que o aluno fixe melhor as combinações mais frequentes de palavras e as estruturas estudadas (ver Figura 2.9).



CONSOLIDAÇÃO LEXICAL

Verbos

Encontre o melhor complemento para os verbos abaixo e acrescente mais um a cada verbo, quando possível.

estar _____
ser _____
estudar _____
tomar _____
chegar _____
querer _____
ter _____
poder _____
pegar _____
morar _____
ir _____

Complementos:

cansado(a)	português	banho
pro cinema	de são paulo	o ônibus
atrasado(a)	à noite	pra são paulo
um carro	café	em são paulo
hoje	dinheiro	até tarde
em casa	30 anos	brasileiro

Figura 2.9: Consolidação lexical da unidade 5, lições A, B e C.

Além disso, em todas as unidades, há os quadros “Note que...” e “Na conversação...”, que chamam a atenção do aluno para expressões típicas da linguagem falada ou escrita. Além disso, o “Lembra?”, como o nome assim sugere, almeja fazer o aluno relembrar tópicos importantes anteriormente estudados. As estruturas repetidas aparecem recicladas em outras unidades como parte essencial da construção de um conhecimento mais avançado (ver Figura 2.10).

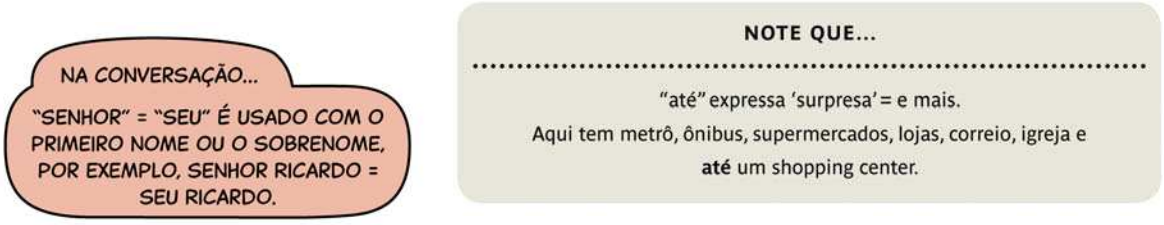


Figura 2.10: Chamadas “Na conversação” e “Note que” das unidades 1 e 7, respectivamente.

No entanto, essas chamadas não foram incluídas no *corpus* MD, como será especificado na seção 2.1.4 (Procedimentos de coleta do *corpus* de estudo MD), a seguir.

O sumário, mostrando o conteúdo do livro em cada uma das unidades, dividido por lição e por seção encontra-se no Anexo 2 desta dissertação.

2.1.4 Procedimentos de coleta do *corpus* de estudo MD

O MP contém, no total, 94.131 *tokens* (palavras), divididos entre 30 unidades (contando as unidades de revisão e pronúncia), apêndices (gramatical e lexical), respostas dos exercícios e transcrições de áudio, sumário, agradecimentos e apresentação.

Para esta pesquisa foram utilizados somente textos e diálogos sem os enunciados dos exercícios. Além disso, os quadros de gramática e os quadros “Note que...”, “Na

conversação...” e “Lembra?” não foram incluídos. Por exemplo, trechos como o apresentado na Figura 2.11 (i.e., Onde está o seu livro de português?) ficaram fora do *corpus* de estudo por representarem comentários explicativos das autoras. O restante do diálogo entrou no *corpus* (*subcorpus* MDNA – material didático não autêntico).

Foram preservadas as divisões entre as unidades, ou seja, cada lição foi armazenada em um arquivo diferente. Esse procedimento foi adotado para que, caso seja necessário, ou em um estudo futuro, seja possível identificar a lição da qual um determinado trigrama/pacote lexical foi retirado. As informações acerca de *tokens*, *types* e *type/token ratio* de cada arquivo (lição) encontram-se no Apêndice 1. Além disso, para a análise de convergência texto a texto, o *corpus* também sofreu uma separação por diálogo/texto, ou seja, um diálogo/texto por arquivo, com o intuito de fazer uma análise mais detalhada.

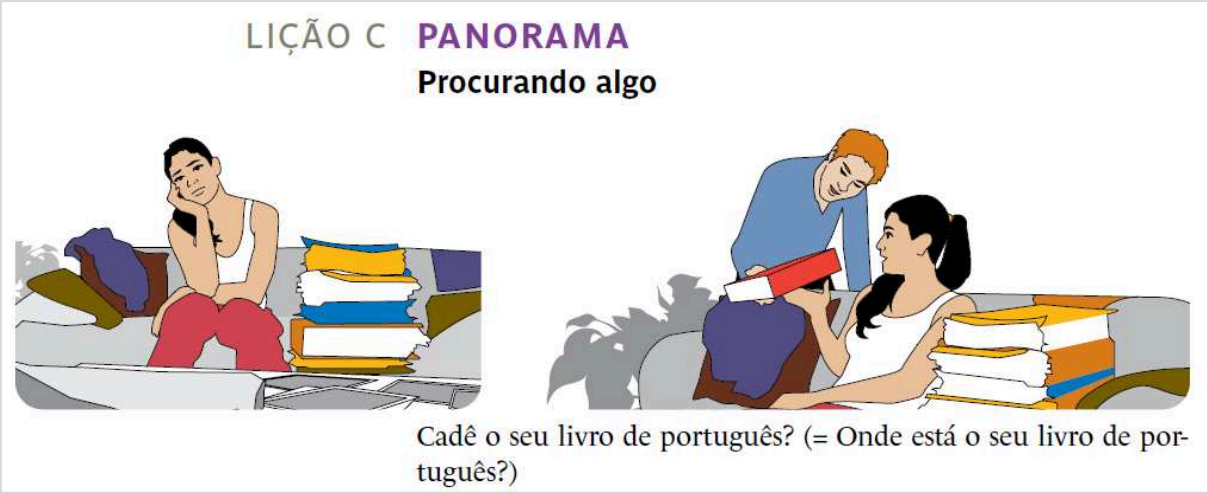


Figura 2.11: Unidade 5, Lição C – Panorama.

Após uma análise piloto dos *corpora* de referência e *corpus* MD (*corpus* MD completo com o Banco de Português – BP – falado), houve a necessidade de criar subdivisões no *corpus* MD com o intuito de separar o conteúdo elaborado pelas autoras do conteúdo retirado de fontes autênticas para a verificação dos elementos de autenticidade do conteúdo não autêntico. A proporção de cada um desses *subcorpora*, bem como a quantidade de *tokens* e *types* encontram-se na seção 2.2.

2.2 Corpora

A seguir, apresentamos detalhes sobre os *corpora* utilizados nesta pesquisa. Primeiramente, detalhamos nosso *corpus* de estudo (MD) e, em seguida, os *corpora* de referência, o Banco de Português (BP) e o *corpus* Baseline.

2.2.1 *Corpus* de estudo – Material didático

O *corpus* de estudo (MD – Muito prazer **–** fale o português do Brasil) possui aproximadamente 41.000 palavras, fora a metalinguagem, ou seja, foram mantidos somente os textos, diálogos e roteiro de gravação de áudio, sem os enunciados dos exercícios nem os comentários explicativos das autoras[2]. Além disso, foram retirados também os exercícios cujas respostas poderiam variar. Os apêndices, sumário, agradecimentos e apresentação também não fizeram parte do *corpus* de estudo.

De acordo Berber Sardinha (2004: 20 e 26), esse *corpus* pode ser considerado pequeno, de amostragem e estático (ver seção 1.1.2 – tipos de *corpora*). É considerado pequeno porque possui menos de 80.000 palavras e, como dito anteriormente, é de amostragem e estático porque é fechado e foi planejado para ser uma amostra finita das modalidades 'falada'[3] e escrita da língua como um todo.

Apresentamos na Tabela 2.1 as informações acerca de *tokens* (itens lexicais)*, types* (formas) e *type/token ratio* (razão forma/item) do *corpus* de estudo com o qual trabalhamos na presente pesquisa (*corpus* completo).

Tabela 2.1 – *Tokens*, *types* e *type/token ratio* do *corpus* de estudo

Tokens **(palavras)**	***Types*** **(formas)**	***Type/token ratio*** **(razão forma/item)**
40.815	1.672	4,1

2. Por exemplo, os quadros "Note que...", "Na conversação..." e "Lembra?".
3. Como dito anteriormente, em sentido estrito, o *corpus* do material didático não contém transcrições de conversação. Trata-se de uma simulação da representação da linguagem oral.

O *corpus* de estudo contém um total de 40.815 *tokens*, ou seja, palavras corridas (cada palavra conta como uma ocorrência, mesmo que repetida) e 1.672 *types* (ou formas, ou seja, vocábulos diferentes). A razão forma/item (*type/token ratio*, ou TTR) indica a riqueza lexical do texto. É obtida dividindo-se o total de formas pelo total de palavras dividido por cem (valor expresso em porcentagem). De acordo com Berber Sardinha (2004: 94), quanto maior seu valor, mais palavras diferentes o texto conterá. O autor afirma ainda que, em contraposição, um valor comparativamente baixo indicará um número alto de repetições, o que poderá indicar um texto menos rico do ponto de vista do seu vocabulário. A título ilustrativo, o nosso *corpus* de referência Baseline apresenta um TTR de 5,2, o que parece indicar um texto mais variado lexicalmente se compararmos esse valor ao do *corpus* MD, que apresentou TTR 4,1.

2.2.1.1 Divisão do corpus *MD em autêntico e não autêntico*

O *corpus* MD foi dividido em autêntico (MDA) e não autêntico (MDNA) com o intuito de verificarmos a proporção exata das duas partes e constatar se os trigramas encontrados especialmente na parte não autêntica serão encontrados nos *corpora* de referência e se podem ser considerados exemplos de linguagem autêntica.

O conteúdo do *subcorpus* MDNA representa grande parte do livro didático (Tabela 2.2) e é composto, em sua maior parte, por diálogos de início de unidade, diálogos criados para as atividades e *scripts* elaborados para os CDs de áudio. Os textos considerados autênticos (*subcorpus* MDA) foram retirados especialmente dos veículos jornais e revistas (internet). No material didático, eles correspondem aos textos da seção “Leitura”. Como o material não autêntico, o autêntico também foi comparado ao *corpus* de referência falado e escrito para obtenção de convergência e classificação dos achados a fim de respondermos as perguntas de pesquisa.

As estatísticas dos *subcorpora* autêntico e não autêntico são apresentadas na Tabela 2.2.

Tabela 2.2 – Estatísticas do material autêntico e não autêntico do *corpus* MD

Material autêntico (MDA)			Material não autêntico (MDNA)		
Tokens	***Types***	**% do *corpus***	***Tokens***	***Types***	**% do *corpus***
5.393	92	13,2%	35.429	1.512	86,8%

Na seção seguinte, 2.2.2, descrevemos os *corpora* de referência utilizados na presente pesquisa, o Banco de Português e o *corpus* Baseline.

2.2.2 *Corpora* de referência

Para a presente pesquisa, optamos por utilizar dois *corpora* de referência que proporcionaram os parâmetros de comparação com o *corpus* de estudo: o *corpus* Banco de Português (BP) e o *corpus* Baseline. O BP (versão 2.0) é um *corpus* de língua geral que possui cerca de 660 milhões de palavras, sendo, atualmente, o segundo maior *corpus* de português do Brasil[4]. O *corpus* Baseline, um *corpus* de amostra da língua, foi compilado para, juntamente com o BP, estipular uma faixa de representatividade, que será explicada em detalhes mais adiante.

A seguir, apresentamos a composição do *corpus* de referência BP, para em seguida detalhar a composição do *corpus* de referência Baseline.

2.2.2.1 Composição do Banco de Português

O BP, compilado e mantido pela Pontifícia Universidade Católica de São Paulo (Berber Sardinha, 2004), é composto de gêneros variados de textos completos, tanto

4. O maior é o *Corpus* Brasileiro, com 1 bilhão de *tokens.*

escritos como transcrições de fala. Os gêneros incluídos nos *subcorpora* escrito e falado do BP são apresentados na Tabela 2.3.

Tabela 2.3 – Composição do Banco de Português versão 2.0

***Subcorpus* escrito**		***Subcorpus* falado**	
Gênero	***Tokens***	**Gênero**	***Tokens***
Acadêmico	343.441.192	Congresso	77.330.504
Culinária	436.971	Conversação	21.430
Informática	874.087	Debate político	21.603
Jornalístico	226.128.749	Entrevista	3.371.725
Legal	246.437	Narração de futebol	74.604
Literatura	1.607.212	Negócios	5.355
Médico	148.256	Pronunciamentos	1.779.712
Negócios	275.817	Variados	3.296.319
Religioso	822.196		
Total escrito	**573.980.917**	**Total falado**	**85.901.252**
		Total geral (escrito + falado)	**659.882.169**

De acordo com a Tabela 2.3, o BP possui quase 660 milhões de *tokens*, e o *subcorpus* escrito do BP contém mais de 570 milhões de palavras (o que representa 87% do *corpus*), enquanto o *subcorpus* falado contém quase 86 milhões de palavras (13% do *corpus*), ou seja, como de costume nos grandes *corpora* eletrônicos gerais de uma língua, há mais textos provenientes da modalidade escrita do que da falada.

Em um primeiro momento, em um estudo piloto, foi utilizado somente o *subcorpus* falado do BP para a retirada das listas de trigramas como referência. A partir dessa análise, percebemos a necessidade de retirar também as listas de trigramas do *subcorpus* escrito e fazer as comparações entre os *corpora* de referência e o *corpus* de estudo, para uma análise mais detalhada.

2.2.2.2 Corpus *Baseline*

O objetivo principal desta pesquisa é verificar se o conteúdo do MD, em especial da parte não autêntica, pode ser considerado exemplo de linguagem autêntica com base na análise dos trigramas e dos pacotes lexicais. Para tanto, a primeira etapa da análise foi determinar uma 'faixa de representatividade', com base em dois *corpora* de referência, que indicaria, com certa segurança, se o conteúdo do *corpus* MD, bem como de seu *subcorpus* MDNA, pode ser considerado autêntico.

O Baseline foi compilado com o intuito de servir como uma amostra da língua autêntica que, ao ser comparado com o *corpus* de referência BP, nos retornaria essa 'faixa de representatividade', ou seja, porcentagens mínimas de convergência de trigramas para que os textos do MD possam ser considerados autênticos. Essa faixa de representatividade indica o que é de se esperar se compararmos os trigramas de dois *corpora* autênticos. Como não existe na literatura nenhuma medida pronta que mostre quantos trigramas existem em comum em dois *corpora* autênticos, precisamos introduzir essa etapa na metodologia. O resultado da comparação nos indica um *baseline,* isto é, uma 'base de correspondência' (*matching*) ou convergência dos dois *corpora* representativos.

O cálculo da convergência se deu do seguinte modo:

- Trigramas em comum entre o BP falado e o Baseline / Trigramas do Baseline × 100 = % de convergência de trigramas em textos da linguagem falada.

- Trigramas em comum entre o BP escrito e o Baseline / Trigramas do Baseline × 100 = % de convergência de trigramas em textos da linguagem escrita.

Em outras palavras, visto que ambos os *corpora* de referência (BP e Baseline) são compostos por textos autênticos escritos e falados, a convergência de trigramas entre esses dois *corpora* nos leva a esperar que um texto poderá ser considerado 'autêntico' se a porcentagem de convergência de seus trigramas estiver próxima ou acima da faixa de representatividade estipulada pelo cálculo acima.

Assim, se a convergência do *corpus* MD com o BP (nosso *corpus* de referência) estiver muito abaixo dos valores mínimos estipulados pela faixa de

representatividade podemos considerar os textos do MD 'não autênticos'; já se a convergência do *corpus* MD com o BP estiver próximo ou acima dos valores mínimos estipulados pela faixa de representatividade, então podemos considerar os textos do MD 'autênticos'.

Sendo assim, a primeira parte da pesquisa consistiu no cálculo da:

- porcentagem de convergência entre os dois *corpora* de referência (BP e Baseline) para obtermos uma estimativa dos valores mínimos (‘faixa de representatividade’) que os textos do material didático deveriam atingir para ser considerados ‘autênticos’; e
- porcentagem de convergência entre o *corpus* de referência BP e o de estudo com o intuito de verificar se o MD atingiu os valores mínimos de autenticidade estipulados na etapa anterior.

2.2.2.2.1 Critérios de coleta e composição do *corpus* Baseline

Como dito anteriormente, o *corpus* Baseline trata-se de uma amostra da língua real e foi compilado para, juntamente com o BP, estipular a faixa de representatividade como o primeiro passo da verificação da autenticidade do MD.

Na compilação, a princípio, pensamos em coletar o total de 100 (cem) textos para compor o *corpus*, sendo estes divididos entre os dois *subcorpora* (falado e escrito), com base nos gêneros existentes no BP (ou seja, acadêmico, culinária, jornalístico etc.). Após o início da coleta, percebemos que a metodologia de 100 textos não seria eficaz, porque havia textos/arquivos com mais *tokens* que outros, o que possivelmente comprometeria o equilíbrio do *corpus*.

Sendo assim, os textos foram coletados e incluídos no Baseline de acordo com o número de tokens, resultando na contagem de tokens e types apresentada na Tabela 2.4

Tabela 2.4 – Composição do *corpus* Baseline

***Subcorpus* escrito**				***Subcorpus* falado**			
Gênero	**Textos**	***Tokens***	***Types***	**Gênero**	**Textos**	***Tokens***	***Types***
Acadêmico	15	60.161	8.463	Congresso	20	143.741	9.190
Culinária	12	1.975	425	Pronunciamentos	5	98.237	8.892
Informática	3	18.358	2.448	Entrevistas	12	81.433	6.435
Jornalístico	65	59.700	11.487				
Jurídico	7	18.495	2.883				
Literatura	2	16.768	3.831				
Médico	18	22.862	4.063				
Negócios	7	103.233	6.098				
Religioso	6	16.651	3.675				
Total – escrito	**135**	**318.000**	**19.200**	**Total – falado**	**37**	**323.000**	**19.000**
				Total geral (escrito + falado)	**172**	**642.000**	**39.500**

Os textos escritos foram coletados da internet, sendo que os da parte falada foram baseados nas notas da Câmara Municipal de São Paulo, no Painel da Previdência e no Museu da Pessoa. As partes escrita e falada ficaram equilibradas em número de *tokens* e *types*, somando, o total de 318.000/19.200 e 323.000/19.000, respectivamente.

2.3 Análise dos *Corpora*

2.3.1. Preparação dos dados

A primeira etapa da pesquisa consistiu na retirada das listas de trigramas com frequência mínima de uma ocorrência, tanto do *corpus* de estudo como nos de referência (BP e Baseline). Para o *corpus* de estudo e Baseline, utilizamos a

ferramenta Listador de Palavras (‘*Wordlist*’) do programa WordSmith Tools 3.0 e para o *corpus* de referência BP, utilizamos *scripts*[5] em Shell e Python.

2.3.1.1 O programa WordSmith Tools e as ferramentas WordList e Concord

O programa WordSmith Tools, criado por Mike Scott por volta de 1996 e publicado atualmente pela Lexical Analysis Software Ltd. e distribuído pela Oxford University Press, é um programa destinado à análise linguística via computador, que disponibiliza uma série de recursos para preparação e análise de *corpora* eletrônicos. Ele apresenta, entre outras, três ferramentas principais: Listador de Palavras (‘*Wordlist*’), Concordanciador (‘*Concord*’) e Listador de Palavras-chave (‘*KeyWords*’). Nesta pesquisa, utilizamos as ferramentas Listador de Palavras e Concordanciador, sendo que a primeira produz três tipos de listas:

(A) Lista de palavras em ordem alfabética;

(F) Lista de palavras em ordem de frequência; e

(S) Lista com dados estatísticos.

Como *default*, são compiladas listas simples (uma palavra). Para que a ferramenta compile listas de três palavras, marcamos a opção ‘*clusters activated*’ em ‘*Settings*’, ‘*Min & Max Frequencies*’. Essa opção faz com que a lista seja montada com *clusters* em vez de palavras isoladas.

Na Figura 2.12 observamos uma das telas para configuração das listas de clusters do programa.

O processo de descobrir agrupamentos ou pacotes em *corpora* é uma tarefa relativamente fácil para um programa de computador, se compararmos como seria se essa tarefa fosse feita manualmente. De maneira simplificada, o computador abre uma janela com o número desejado de palavras (definidas pelo pesquisador, por exemplo, três palavras) e então pesquisa no *corpus* inteiro. Para uma janela de três

5. *Scripts* elaborados por Tony Berber Sardinha e José Lopes Moreira Filho.

palavras, o computador procura nas palavras 1, 2 e 3, depois nas 2, 3 e 4, e assim por diante. Para exemplificar o processo, incluímos uma frase retirada do *corpus* de estudo:

> Nas regiões que adotam a hora de verão, é normal se ter luz solar entre 18h30 e 20h15.

O computador juntará as palavras “nas regiões que” (palavras 1, 2 e 3 do texto), depois “regiões que adotam” (palavras 2, 3 e 4), “que adotam a” (3, 4 e 5), e assim por diante. Ao final, a ferramenta produz uma lista de *clusters* de três palavras, as quais ocorrem um determinado número de vezes, estipulado pelo pesquisador/usuário. No caso desta pesquisa, optamos por buscar até os trigramas que ocorressem uma única vez, visto que nosso *corpus* de estudo é pequeno (Berber Sardinha, 2004: 26).

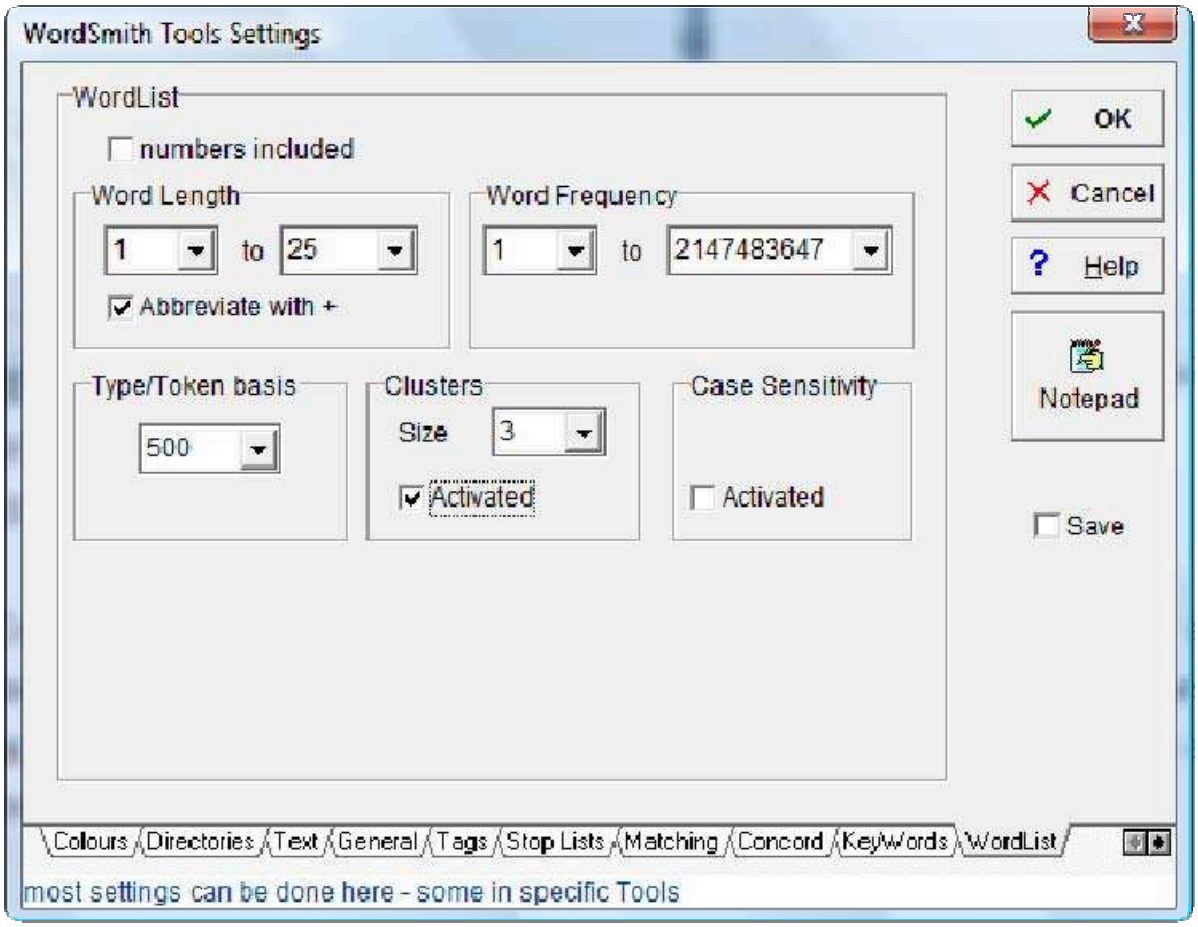


Figura 2.12: Tela do programa WordSmith Tools 3.0.

A escolha de trabalharmos com *clusters* de três palavras (trigramas) em vez de duas, quatro ou mais foi feita com base em Scott e Tribble (2006) e Berber Sardinha (2007). Em sua pesquisa, Scott e Tribble (2006) obtiveram melhores resultados na análise de listas de três e quatro palavras em vez de listas de palavras isoladas e de duas palavras. Além disso, os autores encontraram pouca diferença entre as listas de três ou quatro, visto que muitos dos *clusters* de quatro palavras contêm os de três em suas estruturas – por exemplo, “as a result of” (*cluster* de quatro palavras) contém “as a result” (*cluster* de três palavras).

Berber Sardinha (2007) afirma que para verificar até que ponto dois (ou mais) textos se comparam em termos de sua idiomaticidade, devemos primeiramente decompô-los em pacotes, normalmente de três palavras. A seguir, para cada um desses pacotes, é necessário buscarmos sua frequência em um *corpus* de referência.

Os *scripts* em Shell e Python seguiram a mesma metodologia acima descrita e foram utilizados pelo fato de o WordSmith Tools não ter sido capaz de processar a grande quantidade de dados do nosso *corpus* de referência BP.

A outra ferramenta do programa WordSmith Tools utilizada foi o Concordanciador, que permite ao pesquisador/usuário obter concordâncias de maneira rápida e prática. Em linhas gerais, as concordâncias são listagens das ocorrências de um item específico (que pode ser formado por uma ou mais palavras) acompanhado do texto ao seu redor (cf. Berber Sardinha, 2009).

A Figura 2.13 apresenta uma concordância com o termo de busca “que a gente”.

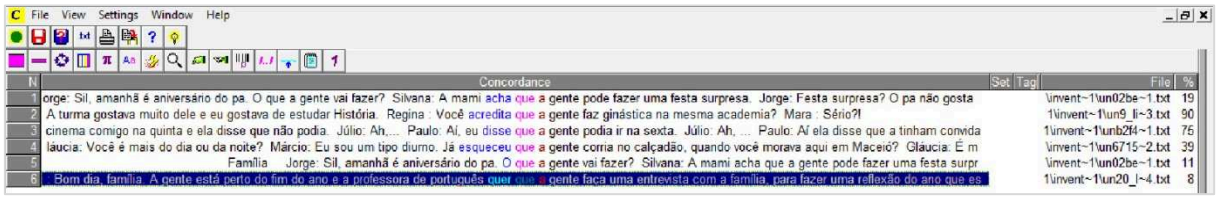


Figura 2.13: Tela do programa WordSmith Tools 3.0.

Segundo Tribble et al. (1990), as concordâncias são uma maneira de observar os padrões na linguagem em uso que permaneceriam “escondidos” sob outras

circunstâncias. Sem elas, a análise manual tanto de palavras isoladas como de *clusters* seria muito custosa e demandaria muito tempo, mesmo em um *corpus* considerado pequeno como o nosso.

2.3.2 Análise dos dados

A partir das listas de trigramas compiladas, novos *scripts*[6] foram elaborados e utilizados para a comparação entre as listas do *corpus* de estudo (MD e seus *subcorpora* MDA e MDNA) e do BP e a identificação dos trigramas convergentes entre os *corpora*. Além disso, o *script* incluiu os trigramas que apareceram somente no MD ('trigramas divergentes') quando comparados ao *corpus* de referência BP, com o intuito de analisá-los mais detalhadamente.

Os trigramas encontrados foram analisados quanto à convergência, ou seja, à quantidade de trigramas em comum entre o *corpus* de estudo e o de referência BP. Como dito anteriormente, essa primeira análise pode indicar uma maior ou menor porcentagem de autenticidade entre o MD e a linguagem em geral (conforme representada no *corpus* de referência), no que se refere à presença de trigramas, ou seja, quanto maior a convergência desses trigramas, maior o grau de autenticidade contido no MD.

A convergência também foi analisada texto a texto com o objetivo de avaliar a variação da convergência entre os textos dos *subcorpora* autêntico e não autêntico. Para tanto, antes da elaboração do *script* para essa análise, foi necessário que filtrássemos possíveis "sujeiras" nos trigramas, ou seja, fez-se necessário que desconsiderássemos os trigramas com características específicas, como nomes próprios, números, comentários das autoras e falta de pontuação, para evitar resultados enganosos. Após essa etapa de limpeza, finalmente foi elaborado o *script*[7] propriamente dito. O *script* nos retornou vários arquivos com a contagem de convergência texto a texto e, ao colocarmos esses dados em uma tabela em MS-Excel, foi possível observar com maior clareza a frequência de cada trigrama em cada texto e em cada unidade do *corpus* de estudo, bem como quais deles eram

6. *Scripts* elaborados por Tony Berber Sardinha e José Lopes Moreira Filho.
7. *Script* elaborado por Tony Berber Sardinha.

convergentes (e sua frequência no BP falado e escrito), além de sua porcentagem de convergência.

Além disso, foi feita a classificação e a análise dos trigramas convergentes. Essa classificação consistiu na divisão em trigramas subusados, de uso equivalente, e sobreusados no MD. Além deles, os divergentes também foram analisados.

Por fim, foram retirados os pacotes lexicais convergentes com base na normalização[8] dos trigramas e em ponto de corte. Após essa etapa, buscamos alguns dos mais representativos tanto nos *subcorpora* MDNA e MDA como no de referência BP. O objetivo desse procedimento é o estudo mais detalhado dos agrupamentos mais significativos nos dois *subcorpora*, visando comprovar qualitativamente a autenticidade observada nos procedimentos anteriores. Além disso, buscamos verificar se todos os pacotes lexicais divergentes encontrados são realmente divergentes, primeiramente por meio de uma amostra dos 100 (cem) mais frequentes e, depois, por análise qualitativa.

Sendo assim, a metodologia de pesquisa pode ser sintetizada pelas seguintes etapas:

1. coleta e divisão do *corpus* MD em MDA e MDNA;

2. divisão do *corpus* MD texto a texto;

3. coleta do *corpus* Baseline*;*

4. elaboração de listas de trigramas dos *corpora* de estudo e de referência;

5. cálculo da 'faixa de representatividade';

6. retirada dos trigramas convergentes e divergentes entre os *corpora* de estudo e referência*;*

7. cálculo da convergência entre os *corpora* MD e BP;

8. análise e classificação dos trigramas convergentes e divergentes;

8. Nome dado ao procedimento estatístico usado para ajustar a contagem da frequência bruta de *corpora* de tamanhos diferentes para conduzir uma comparação confiável (Biber et al., 1998 apud Cortes, 2006).

9. retirada dos pacotes lexicais convergentes e divergentes;

10. retirada dos trigramas realmente divergentes;

11. análise da convergência texto a texto.

A seguir, apresentamos os resultados da análise e as considerações finais da pesquisa.

CAPÍTULO 3

APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Neste capítulo apresentamos as estatísticas gerais dos *corpora* e os resultados das análises quantitativa e qualitativa, bem como as descobertas em relação aos trigramas e pacotes lexicais convergentes e divergentes encontrados no MD em comparação ao *corpus* de referência BP.

Inicialmente, apresentamos os dados estatísticos encontrados nos *corpora* pesquisados e a convergência dos trigramas entre os *corpora* de estudo e de referência.

3.1 Faixa de representatividade

O objetivo principal da nossa pesquisa é verificar se o conteúdo do MD, em especial a parte não autêntica (MDNA), pode ser considerado exemplo de linguagem autêntica com base na análise dos trigramas e dos pacotes lexicais encontrados. Para isso, na primeira parte da análise, utilizamos dois *corpora*, o BP e o Baseline, para calcularmos o que chamamos de 'faixa de representatividade', isto é, valores de referência mínimos em que poderíamos nos embasar para o primeiro passo da verificação do grau de autenticidade do MD.

Como dito anteriormente, essa faixa de representatividade indica o que é de se esperar se compararmos os trigramas de dois *corpora* autênticos e o resultado da comparação nos indica um *baseline,* isto é, uma base de correspondência ou convergência dos dois *corpora* representativos.

Os *scripts* utilizados para retirada dos trigramas e a ferramenta Listador de Palavras ('Wordlist') do programa WordSmith Tools nos retornaram as seguintes quantidades de trigramas apresentadas na Tabela 3.1.

Tabela 3.1 – Número de trigramas dos *corpora* de referência BP e Baseline

Corpus	**Número de trigramas (formas)**
BP falado	25.374.300
BP escrito	176.995.186
Baseline	417.159

A partir dos trigramas apresentados na Tabela 3.1 foi possível calcular a convergência entre nossos *corpora* de referência, o BP e o Baseline. Os resultados dessa comparação são apresentados na Tabela 3.2.

A convergência foi calculada com base nos números de trigramas em comum entre o BP Falado e Escrito e o Baseline (109.715 e 217.559, respectivamente) divididos pelo número de trigramas do Baseline (417.159). Assim, temos 26,30% (109.715 / 417.159 × 100) de convergência entre o *corpus* Baseline e o BP falado e 52,15% de convergência entre o Baseline e o BP escrito. Isso nos leva a crer que seria de se esperar que textos autênticos da linguagem falada estejam acima ou na faixa de 26,30% e textos autênticos da linguagem escrita estejam acima ou na faixa de 52,15% de trigramas em comum com os *corpora* de referência. Se a comparação do *corpus* MD (MDA e MDNA) com o BP falado e escrito estiver muito abaixo desses valores mínimos, então podemos considerar os textos 'não autênticos' e se estiverem próximo ou acima da faixa 'autênticos'.

Verificamos a necessidade dessa faixa de representatividade devido ao fato de que somente os valores de convergência entre o MD e o BP não nos deram suporte suficiente para avaliar o que podemos considerar autêntico ou não autêntico para o MD. Os valores da faixa de representatividade nos deram um ponto de referência mínima para nos apoiarmos, visto que os *corpora* comparados (BP e Baseline) se tratam de textos autênticos da linguagem falada e escrita do português do Brasil.

Tabela 3.2 – Convergência entre os *corpora* Baseline e BP falado e escrito

Baseline e BP Falado				Baseline e BP Escrito			
Baseline	BP Falado	Em comum	Convergência	Baseline	BP Escrito	Em comum	Convergência
417.159	25.374.300	109.715	26,30%	417.159	176.995.186	217.559	52,15%

Tabela 3.3 – Convergência dos trigramas no *subcorpora* MDNA com o BP falado e escrito

MDNA e BP Falado				MDNA e BP Escrito			
MDNA	BP Falado	Em comum	Convergência	MDNA	BP Escrito	Em comum	Convergência
21.500	25.374.300	6.165	28,60%	21.500	176.995.186	12.426	57,80%

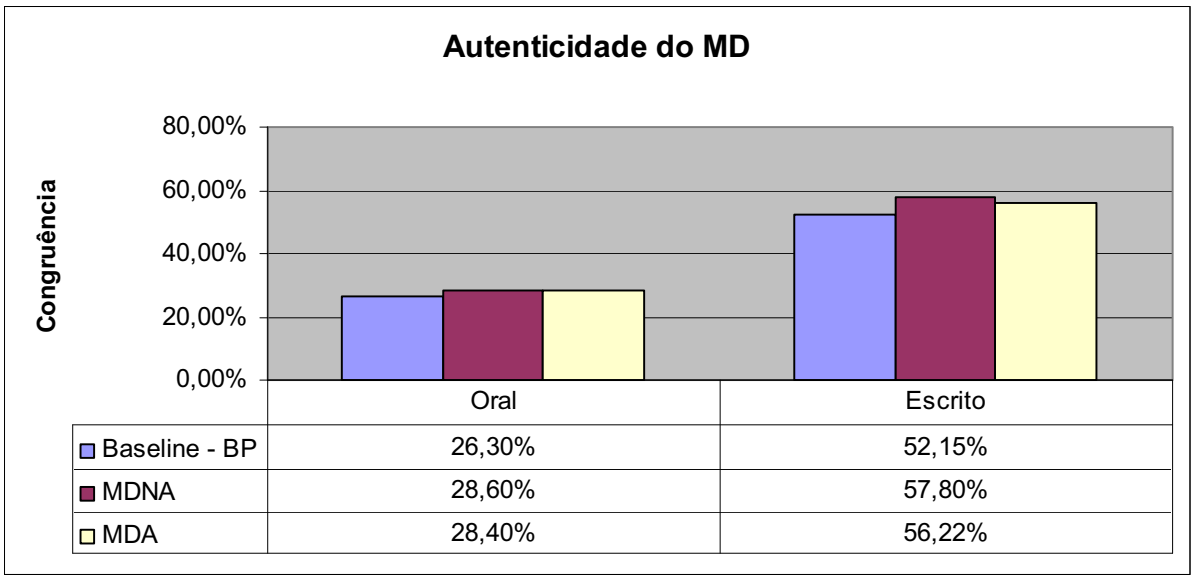
Tabela 3.4 – Convergência dos trigramas no *subcorpora* MDA com o BP falado e escrito

MDA e BP Falado				MDA e BP Escrito			
MDA	BP Falado	Em comum	Convergência	MDA	BP Escrito	Em comum	Convergência
4.589	25.374.300	1.304	28,40%	4.589	176.995.186	2.580	56,22%

3.1.1 Convergência entre o MD e o BP

Foram feitos cálculos de convergência entre o *corpus* de estudo MD e o *corpus* de referência BP (falado e escrito, separadamente), sendo que eles seguiram a mesma metodologia utilizada para o cálculo da convergência entre o BP e o Baseline, ou seja, a convergência foi calculada com base nos números de trigramas em comum entre o *corpus* MD e o de referência BP divididos pelo número de trigramas do *corpus* MD. Por esta ser uma pesquisa que visa verificar a autenticidade do MD, sentimos necessidade em dividir o *corpus* em MDA e MDNA (autêntico e não autêntico, respectivamente) para melhor análise do material elaborado pelas autoras e buscamos conhecer a convergência entre esses dois *subcorpora* com o *corpus* de referência BP. Os resultados são apresentados nas tabelas 3.3 e 3.4.

Assim, os valores de 28,60% e 28,40% para o MD com BP falado e 57,80% e 56,22% para o MD com BP escrito nos levam a acreditar que o livro é tão 'autêntico' quanto uma coletânea de textos autênticos, visto que nossa faixa de representatividade apresentou os valores de convergência de 26,30% e 52,15%, ou seja, os valores de convergência do MD estão acima do mínimo esperado para serem considerados representativos de textos 'autênticos'.



	Oral	Escrito
Baseline - BP	26,30%	52,15%
MDNA	28,60%	57,80%
MDA	28,40%	56,22%

Gráfico 3.1: Valores de convergência entre o Baseline-BP (Faixa de representatividade), MD não autêntico e autêntico vs. BP falado e escrito.

Com base nos dados apresentados no Gráfico 3.1, a diferença entre o número de trigramas convergentes dos textos autênticos e dos não autênticos não nos parece estatisticamente relevante, visto que o MDNA e o MDA vs. BP falado apresentaram os valores de convergência de 28,60% e 28,40% e o MDNA e o MDA vs. BP escrito, 57,80% e 56,22%, respectivamente. Isso, mais uma vez, nos leva a acreditar que o *corpus* do material não autêntico parece ser tão 'autêntico' quanto o do autêntico. Para termos certeza de que há empate técnico entre o autêntico e o não autêntico, utilizamos a calculadora de qui-quadrado[1].

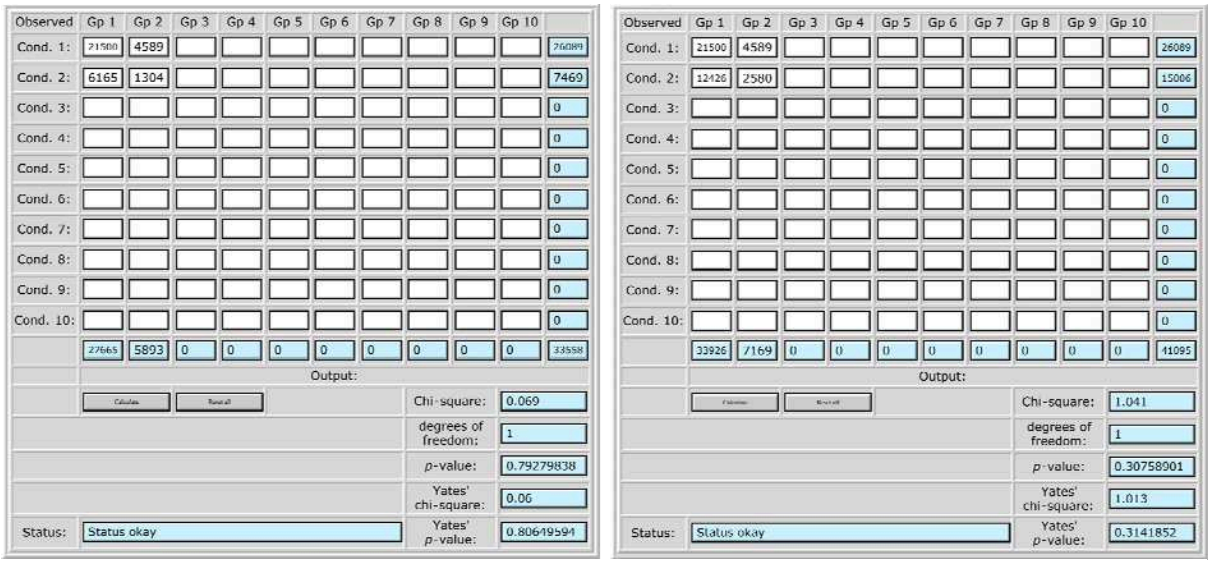


Figura 3.1: Telas com os resultados da calculadora de qui-quadrado.

O qui-quadrado é uma medida estatística de comparação (Berber Sardinha, 2004: 104) que testa a associação significativa entre variáveis e, para haver diferença significativa entre os valores obtidos o resultado do cálculo (valor de significância – 'p-value') precisa ser inferior a 0,05. Em ambos os cálculos, os resultados foram maiores a esse valor (0,79279 para o MD vs. BP falado e 0,30758 para o MD vs. BP escrito), portanto, não há diferença estatisticamente relevante entre os trigramas convergentes dos *subcorpora*.

Outro dado que nos chamou a atenção com relação à convergência foi que os valores são mais altos nas comparações entre o MDNA com o BP escrito (57,80%) do que do MDNA com o BP falado (28,60%). Nossa intuição inicial seria que o

1. Disponível em: <http://people.ku.edu/~preacher/chisq/chisq.htm>. Acesso em: 24 jul. 2010.

subcorpus MDNA, que faz uma tentativa de representar a linguagem oral com diálogos e roteiros de áudio para compreensão auditiva, teria mais trigramas em comum com o BP falado e não com o BP escrito. No entanto, contrariando nossas expectativas como coautora do material didático, o MDNA apresentou convergência maior com o BP escrito.

Uma das possíveis explicações para isso é que, ainda que tenha havido uma tentativa de representar a fala no MDNA com marcas de oralidade (ao incluir o léxico do dia a dia e alguns elementos próprios da linguagem falada, como repetições, marcadores conversacionais e interjeições), os textos possuem estruturas da modalidade escrita, uma vez que foram primeiramente escritos para posteriormente serem gravados e interpretados. Leite (2009) analisa a oralidade em textos escritos e cita Urbano (1999: 104), que afirma que "por mais realista que o autor [de uma obra] pretenda ser, ele esbarra nos limites da escrita, da economia e funcionalidade do texto literário e da aceitabilidade do leitor". Além disso, outro fator que pode ter contribuído para tal resultado é a preocupação das autoras em relação, principalmente, à aceitabilidade do leitor, que no caso do material didático em questão é um aluno iniciante e com conhecimento superficial da língua.

Uma vez estabelecida a convergência entre os trigramas, a próxima etapa da pesquisa é analisar como os trigramas estão representados no MD. Para isso, os classificamos em subusados, de uso equivalente, sobreusados e únicos no MD.

3.1.2 Análise e classificação dos trigramas

3.1.2.1 Trigramas convergentes

Como dito anteriormente, os trigramas encontrados na comparação entre os *corpora* foram divididos em convergentes (trigramas em comum) e divergentes (trigramas únicos) e manteve-se a separação entre MDA e MDNA.

Em uma primeira análise, verificamos a frequência dos trigramas no MD absoluta e normalizada por 1.000.000 (representadas na Tabela 3.5, respectivamente, por 'Freq MD completo absoluta' e 'Freq MD normalizada'), a frequência no BP (falado e

escrito, absoluta, ‘Freq BP falado absoluta’, e normalizada por 1.000.000, ‘Freq BP falado normalizada’) e a razão, isto é, a porcentagem de uso do trigrama no MD quando comparado ao BP. Essa razão foi calculada por meio da divisão do número de ocorrências normalizado no MD pelo número de ocorrências normalizado no BP.

Tabela 3.5 – Amostra de dados dos trigramas convergentes do MD completo vs. BP falado

Trigrama	Freq MD completo absoluta	Freq MD normalizada	Freq BP falado absoluta	Freq BP falado normalizada	Razão MD/BP
O_QUE_TINHA	1	24,51	17891	208,03	0,12
DIA_#_DE	1	24,51	14966	174,02	0,14
DO_RIO_DE	1	24,51	13459	156,50	0,16
GRANDE_DO_SUL	1	24,51	12546	145,88	0,17

Sendo assim, o trigrama ‘o que tinha’, por exemplo, possui frequência normalizada de 24,51 no *corpus* MD, ou seja, esse trigrama ocorre 24,51 vezes a cada 1.000.000 de palavras. No BP falado, o mesmo trigrama aparece 208,03 vezes a cada 1.000.000 de palavras. A partir dessa pequena amostra de dados, já é possível depreender que há trigramas no MD em subuso quando comparados ao *corpus* de referência.

Além desses dados, foram calculadas as médias da frequência total no *corpus* MD e no BP e a razão total (Tabela 3.6).

Tabela 3.6 – Razão total: *corpus* MD e BP

MD e BP falado			MD e BP escrito		
Média MD	Média BP falado	Razão	Média MD	Média BP escrito	Razão
0,30%	0,02%	15	0,31%	0,01%	31

As médias 0,30% e 0,31% e 0,02% e 0,01% apresentadas na Tabela 3.6 representam a frequência média dos trigramas, em porcentagem. Os trigramas têm em média a frequência de 0,30% e 0,31% no *corpus* MD e de 0,02% e 0,01% no BP falado e escrito, respectivamente. Dividindo o valor da média MD pela do BP falado e escrito nos retornam os valores de 15 e 31, ou seja, os trigramas do material didático aparentemente ocorrem 15 vezes mais do que no BP falado e 31 vezes mais do que no BP escrito.

Aparentemente, esses valores podem ser considerados altos, porque se partirmos do pressuposto de que a amostra de língua do livro didático deveria ser parecida com a da língua 'real', então a razão deveria ser 1, ou seja, não deveria haver diferença entre o *corpus* do MD e o de referência. No entanto, isso é somente hipotético, porque, como já vimos, o *corpus* Baseline, que é uma amostra de textos 'reais', não se iguala com o *corpus* de referência (ver Tabela 3.2). Contudo, esses valores podem indicar um possível sobreuso dos trigramas no MD. Sendo assim, a próxima seção trata da classificação da frequência dos trigramas em subuso, uso equivalente e sobreuso no MD de acordo com o *corpus* de referência BP.

3.1.2.1.1 Subuso, uso equivalente e sobreuso

A partir da razão MD/BP (ver Tabela 3.5, coluna 'razão MD/BP' e Apêndice 4, colunas 'razão' e 'classificação'), os trigramas foram classificados em subusados, de uso equivalente e sobreusados. Eles foram classificados de acordo com os critérios elencados na Tabela 3.7.

Para a elaboração destes critérios, nos baseamos inicialmente em Lúcio (2006) que, em sua dissertação de Mestrado, classifica o sobreuso de adjetivos em inglês pela porcentagem de pelo menos duas vezes mais do que no *corpus* de referência.

Tabela 3.7 – Classificação dos trigramas convergentes quanto ao subuso, uso equivalente e sobreuso

Subuso	Uso equivalente	Sobreuso
Até 0,99 (ou seja, os trigramas no MD precisam ocorrer até 0,99 vezes quando comparados ao BP).	De 1,00 a 1,99 (ou seja, os trigramas no MD precisam ocorrer de 1 a 1,99 vezes mais do que no BP).	De 2,00 a ... (ou seja, os trigramas no MD precisam ocorrer de 2 a ....vezes mais do que no BP).

Com base nos critérios apresentados na Tabela 3.7, chegamos aos dados apresentados nas tabelas 3.8 e 3.9.

Ao verificar a grande quantidade de trigramas sobreusados em todas as comparações (ver Tabelas 3.8 e 3.9), tanto no material não autêntico (97,94% e 99,15%) como no autêntico (93,79% e 96,86%), acreditamos que o MD tende à repetição dos trigramas, que gera o sobreuso. Isso talvez se deva ao fato de as autoras sentirem a necessidade de incluir os mesmos trigramas muitas vezes para que a reiteração sirva como uma oportunidade de rever os tópicos de diferentes níveis de profundidade e solidificar o conhecimento. Do ponto de vista didático, isso nos parece adequado, visto que, de acordo com a teoria denominada 'lexical priming' de Hoey (2005 apud Alambert, 2008), o contato reiterado com sequências e estruturas recorrentes faz com que elas se fixem na memória e sejam ativadas ('primed') quando o estímulo é apresentado.

Na seção seguinte, 3.1.2.2, analisaremos os trigramas divergentes, ou seja, aqueles que somente apareceram no *corpus* MD.

Tabela 3.8 – Resultado da classificação dos trigramas convergentes no *subcorpus* MDNA

MDNA vs. BP falado					MDNA vs. BP escrito				
Total de trigramas (MDNA)	**Total de trigramas convergentes**	**Subuso**	**Uso equivalente**	**Sobreuso**	**Total de trigramas (MDNA)**	**Total de trigramas convergentes**	**Subuso**	**Uso equivalente**	**Sobreuso**
21.500	6.165	51 trigramas (0,83%)	76 trigramas (1,23%)	6.038 trigramas (97,94%)	21.500	12.426	35 trigramas (0,28%)	69 trigramas (0,55%)	12.321 trigramas (99,15%)

Tabela 3.9 – Resultado da classificação dos trigramas convergentes no *subcorpus* MDA

MDA vs. BP falado					MDA vs. BP escrito				
Total de trigramas (MDA)	**Total de trigramas convergentes**	**Subuso**	**Uso equivalente**	**Sobreuso**	**Total de trigramas (MDA)**	**Total de trigramas convergentes**	**Subuso**	**Uso equivalente**	**Sobreuso**
4.589	1.304	39 trigramas (2,99%)	42 trigramas (3,22%)	1.223 trigramas (93,79%)	4.589	2.580	38 trigramas (1,47%)	43 trigramas (1,67%)	2.499 trigramas (96,86%)

3.1.2.2 Trigramas divergentes

Foram elaborados *scripts* em Shell e Python para obtermos as listas de trigramas que são únicos no MD. Apresentamos na Tabela 3.10 as estatísticas de trigramas que ocorrem somente no MD quando comparados aos *subcorpora* falado e escrito do BP.

Tabela 3.10 – Trigramas divergentes no MD

MD vs. BP falado		**MD vs. BP escrito**	
Trigramas que ocorrem no MD e no BP falado	Trigramas que ocorrem somente no MD	Trigramas que ocorrem no MD e no BP escrito	Trigramas que ocorrem somente no MD
7.393	6.437	13.797	4.668

Como visto anteriormente, a convergência é maior entre o BP escrito e o MD, consequentemente, há menos trigramas divergentes no MD nessa comparação (4.668).

Em uma análise inicial, verificamos que os trigramas divergentes com frequências mais altas foram retirados do *subcorpus* MDNA, o que nos levaria a pensar que estes seriam trigramas não autênticos. No entanto, como veremos mais adiante, nem todos esses trigramas são realmente não autênticos, pois algumas características específicas, tais como nomes próprios, pontuação retirada pelo *script* de programação e numerais, influenciaram os valores apresentados na Tabela 3.10, levando-nos a falsos trigramas divergentes.

Devido à grande quantidade de trigramas convergentes e divergentes e o fato de que muitos deles possivelmente não possuem frequência representativa, a próxima etapa foi retirar os trigramas de alta frequência (‘pacotes lexicais’). Para isso, utilizamos a nota de corte tradicional de pelo menos 20 vezes por milhão de palavras. Os resultados da análise estão nas seções seguintes: pacotes lexicais convergentes e divergentes.

3.2 Pacotes lexicais convergentes e divergentes

3.2.1 Pacotes lexicais convergentes

Como dito anteriormente (ver seção 1.3 – 'Pacotes lexicais' (*lexical bundles*)), somente podemos considerar pacotes lexicais os trigramas que mostram uma tendência estatística de co-ocorrerem juntos em um determinado tipo de texto. Como eles são definidos por sua frequência, a combinação de palavras tem de ocorrer, pelo menos, dez vezes por milhão de palavras (Biber et al., 1999: 990). No entanto, para a presente análise, optamos por um valor mais conservador, por segurança, qual seja, trabalhar com o ponto de corte (PC) de vinte vezes por milhão de palavras.

Sendo assim, verificamos entre os trigramas convergentes do MD se estes possuíam a frequência adequada para ser considerados pacotes lexicais. O intuito dessa classificação é nos concentrarmos na análise dos trigramas convergentes mais importantes/representativos do MD. A retirada dos pacotes lexicais dos *subcorpora* MDA e MDNA foi feita com base em sua frequência no *corpus* de referência BP e ficamos com os valores para análise apresentados nas Tabelas 3.11 e 3.13.

Tabela 3.11 – Total de pacotes lexicais convergentes no MDA

MDA vs. BP falado	**MDA vs. BP escrito**
Pacotes lexicais encontrados (PC = 20x por milhão)	**Pacotes lexicais encontrados (PC = 20x por milhão)**
Total de pacotes lexicais convergentes = 62	Total de pacotes lexicais convergentes = 55

Dos 1.304 trigramas encontrados na comparação MDA vs. BP falado, somente 62 foram considerados pacotes lexicais (ver a lista de pacotes lexicais encontrados no Apêndice 4). Além disso, nenhum dos trigramas subusados ou de uso equivalente foi considerado pacote lexical (eram 39 trigramas subusados e 42 de uso equivalente antes dessa análise).

Na comparação MDA vs. BP escrito, dos 2.580 trigramas convergentes, somente 55 foram considerados pacotes lexicais (ver a lista de pacotes lexicais encontrados no Apêndice 4), sendo que desses, todos os trigramas subusados eram pacotes lexicais (38), dos 43 de uso equivalente 16 foram considerados pacotes lexicais e dos 2.499 sobreusados somente 1 foi considerado pacote lexical.

Com base no total de pacotes lexicais (62 e 55) e nos trigramas convergentes (1.304 e 2.580) encontrados, podemos concluir que a maior parte dos trigramas convergentes no MDA é de baixa frequência no BP.

Desses pacotes, na Tabela 3.12 selecionamos os vinte mais frequentes do MDA (coluna à esquerda) e os vinte mais frequentes do BP escrito (coluna à direita). Vale lembrar que o material autêntico do MD é representado por textos escritos, por isso a comparação da Tabela 3.12 foi feita somente com o *subcorpus* escrito do BP.

Tabela 3.12 – Pacotes lexicais mais frequentes do MDA e do BP escrito

N	Pacote lexical (MDA)	Freq MD normalizada por 1.000.000	N	Pacote lexical (BP escrito)	Freq BP escrito normalizada por 1.000.000
1	A PARTIR DE	171,57	1	DE SÃO PAULO	479,79
2	DE ACORDO COM	122,55	2	RIO DE JANEIRO	398,83
3	DE SÃO PAULO	73,53	3	DE ACORDO COM	269,17
4	ACORDO COM A	73,53	4	A PARTIR DE	212,05
5	A FALTA DE	73,53	5	AO MESMO TEMPO	93,33
6	PARA O BRASIL	73,53	6	A PARTIR DO	89,69
7	RIO DE JANEIRO	49,02	7	EM QUE O	74,14

8	AO MESMO TEMPO	49,02	8	O USO DE	73,42
9	O USO DE	49,02	9	MAIS DO QUE	72,36
10	PARA QUE O	49,02	10	O QUE É	71,28
11	UOL COM BR	49,02	11	ACORDO COM A	70,76
12	QUE O BRASIL	49,02	12	PARA A FOLHA	69,64
13	A PARTIR DO	24,51	13	DE TODOS OS	68,47
14	EM QUE O	24,51	14	DA UNIVERSIDADE DE	66,09
15	MAIS DO QUE	24,51	15	A FIM DE	53,85
16	O QUE É	24,51	16	SÃO PAULO E	52,36
17	PARA A FOLHA	24,51	17	O QUE SE	52,19
18	DE TODOS OS	24,51	18	A FALTA DE	51,02
19	DA UNIVERSIDADE DE	24,51	19	AO LONGO DO	44,72
20	A FIM DE	24,51	20	DA DÉCADA DE	43,46

Desses, os pacotes encontrados no MD ‘a partir de’, de acordo com’, de São Paulo’, ‘acordo com a’, ‘Rio de Janeiro’, ‘ao mesmo tempo’, ‘o uso de’, ‘a partir do’, ‘em que o’ e ‘mais do que’ também são os mais frequentes no BP escrito, sendo que os pacotes ‘a fim de’, ‘a partir de’, ‘acordo com a’, ‘em que o’, ‘o uso de’ parecem ser característicos da linguagem escrita. O pacote ‘de acordo com’ também nos parece ser característico da linguagem escrita, no entanto, ele aparece na lista de pacotes mais frequentes do BP falado, apesar de apresentar uma frequência mais baixa no BP falado do que no BP escrito (136,19 no BP falado vs. 269,17 no BP escrito por 1.000.000 de palavras), como podemos observar na Tabela 3.14.

Na Tabela 3.13, dos 6.165 trigramas encontrados na comparação MDNA vs. BP falado, somente 101 foram considerados pacotes lexicais (ver Apêndice 4), sendo que desses, 51 foram subusados, 32 tiveram uso equivalente e dos 6.038 somente 18 eram pacotes lexicais sobreusados.

Tabela 3.13 – Total de pacotes lexicais convergentes no MDNA

MDNA vs. BP falado	MDNA vs. BP escrito
Pacotes lexicais encontrados (NC = 20x por milhão)	**Pacotes lexicais encontrados (NC= 20x por milhão)**
Total de pacotes lexicais convergentes = 101	Total de pacotes lexicais convergentes = 68

Na comparação MDNA vs. BP escrito, dos 12.426 trigramas convergentes, somente 68 foram considerados pacotes lexicais (ver Apêndice 4), sendo que desses, todos os trigramas subusados eram pacotes lexicais (36), dos 69 de uso equivalente 20 foram considerados pacotes e dos 12.321 somente 12 foram considerados pacotes lexicais sobreusados.

Desses pacotes, na Tabela 3.14 selecionamos os vinte mais frequentes do MDNA (coluna à esquerda) e os vinte mais do BP falado (coluna à direita). Vale lembrar que o material não autêntico do MD é representado por textos 'falados', por isso a comparação da tabela abaixo foi feita somente com o *subcorpus* falado do BP.

Tabela 3.14 – Pacotes lexicais mais frequentes do MDNA e do BP falado

N	Pacote lexical (MDNA)	Freq MD normalizada por 1.000.000	N	Pacote lexical (BP falado)	Freq BP falado normalizada por 1.000.000
1	RIO_DE_JANEIRO	196,08	1	MAIS_DE_#	291,43
2	QUE_A_GENTE	147,06	2	RIO_DE_JANEIRO	268,30
3	A_FIM_DE	122,55	3	O_QUE_TINHA	208,03
4	EU_ACHO_QUE	122,55	4	QUE_O_GOVERNO	205,35
5	EM_#_DE	98,04	5	RIO_GRANDE_DO	176,60
6	CADA_VEZ_MAIS	98,04	6	A_FIM_DE	175,62
7	DE_R_#	73,53	7	DIA_#_DE	174,02

8	MERCADO_DE_TRABALHO	73,53	8	DO_RIO_DE	156,50
9	MAIS_DE_#	49,02	9	GRANDE_DO_SUL	145,88
10	QUE_O_GOVERNO	49,02	10	DE_TODOS_OS	140,47
11	RIO_GRANDE_DO	49,02	11	DE_ACORDO_COM	136,19
12	NO_RIO_DE	49,02	12	DE_QUE_A	115,80
13	NO_ANO_PASSADO	49,02	13	MAIS_DO_QUE	114,16
14	A_OPORTUNIDADE_DE	49,02	14	EM_#_DE	111,41
15	EM_#_O	49,02	15	CADA_VEZ_MAIS	108,15
16	O_QUE_O	49,02	16	DE_R_#	98,48
17	DO_ANO_PASSADO	49,02	17	DE_#_ANOS	93,43
18	EM_#_A	49,02	18	EU_ACHO_QUE	81,24
19	TUDO_O_QUE	49,02	19	AO_MESMO_TEMPO	79,71
20	SOBRE_A_MESA	49,02	20	QUE_A_GENTE	74,71

Muitos deles parecem ser característicos da linguagem oral, como ‘eu acho que’ e ‘que a gente’, sendo que esses, no BP falado, são parte de um pacote maior (eu acho que a gente).

Além disso, alguns deles aparecem no material didático com frequência semelhante ao *corpus* de referência BP falado (p. ex., o pacote ‘cada vez mais’ tem frequência normalizada de 98,04 no MD vs. 108,15 no BP falado).

Um item que nos chamou a atenção por estar presente nas comparações tanto com o *corpus* de referência falado como escrito foi ‘a fim de’. À primeira vista, pensamos que se tratava de um pacote típico da linguagem escrita. Por isso, utilizamos a ferramenta ‘concordanciador’ do programa WordSmith Tools que nos retornou as seguintes linhas de concordância para o *corpus* MD:

1. Estou **a fim de** uma moqueca Tipos de comida
2. e comida servem? Só massas. Você está **a fim de** ir? Claro, mas é baratinho?
3. ceterias. Marina: Calma Alba. Você não está **a fim de** dançar? Então, tem que esperar. Al
4. Francisco: Hoje não quero ir ao quilo. Estou **a fim de** uma moqueca. Fernando: Então, v
5. eiro. Vamos ao teatro hoje? Vamos! Estou **a fim de** ver aquela peça "Trair e Coçar é só c
6. ara participar de uma entrevista de trabalho. **A fim de** orientar quem está à procura de um

Nessas linhas de concordância observamos que há, no MD, dois usos diferentes da expressão 'a fim de': um mais informal e utilizado na linguagem falada para expressar vontade ou disposição de fazer algo já mencionado (linhas 1 a 5) e outro mais formal (linha 6), para expressar propósito ou intenção de algo (Ferreira, 2004).

3.2.2 Pacotes lexicais divergentes

Devido à grande quantidade de trigramas divergentes no MD, foi necessário, em uma primeira análise, escolhermos uma amostra dos cem mais frequentes com o intuito de verificar se todos eles eram, de fato, divergentes. Essa etapa foi necessária por conta de haver trigramas com 'sujeira', ou seja, com características específicas que impossibilitavam a verificação exata da convergência com o *corpus* de referência. Entre essas características, temos:

1. Numeração

Pacote lexical – somente MD (completo)	Frequência no MD	Freq normalizada por 1.000.000
# # #	52	1275
ÀS # #	11	270
É # #	11	270
# DA NOITE	9	221
# ANOS EU	7	172
# DIAS E	7	172
# # E	6	147

TENHO # ANOS	3	74
TRABALHO ÀS #	3	74

Ao retirar as listas de trigramas, o *script* de programação inclui o símbolo # para representar um numeral.

2. Pontuação

Pacote lexical – somente MD (completo)	Frequência no MD	Freq normalizada por 1.000.000
B ACHO QUE	6	147
B EU SEI	5	123
ATENDENTE QUAL É	4	98
REPÓRTER O QUE	4	98
PEDRO A SENHORA	3	74
TAXISTA MUITO OBRIGADO	3	74
B AINDA NÃO	3	74
B BOM DIA	3	74
MUITOPRAZER COM BR	5	123

O *script* também eliminou a pontuação, o que acabou juntando, por exemplo, as marcações de fala (‘B: Acho que’ e ‘Taxista: Muito obrigado’).

3. Comentários / observações das autoras

Pacote lexical – Somente MD (completo)	Frequência no MD	Freq normalizada por 1.000.000
ADAPTADO DE HTTP	5	123
ADAPTADO DO SITE	3	74

4. Nomes próprios

Pacote lexical – Somente MD (completo)	Frequência no MD	Freq normalizada por 1.000.000
VIAGENS MUITO PRAZER	3	74
DA VEJA RIO	3	74
É O FERNANDO	3	74

Sendo assim, da amostra dos 100 mais frequentes, desconsideramos os pacotes com as características acima mencionadas. Na comparação do MD com o BP falado, restaram 57 pacotes realmente divergentes (o que representa 57% da amostra) sendo que desses, 55 foram retirados do MDNA. Na comparação com o BP escrito restaram 30 (30% da amostra é realmente divergente) sendo que os 30 foram encontrados no MDNA (ver Tabela 3.15).

Tabela 3.15 – Distribuição dos pacotes lexicais divergentes na comparação com BP falado e escrito

BP falado com MDNA	BP falado com MDA	BP escrito com MDNA	BP escrito com MDA
55	10	30	9

Como podemos observar, a maior parte dos pacotes lexicais divergentes foi retirada do material não autêntico (MDNA). No entanto, visto que a recontagem de pacotes nos levou a uma queda considerável de pacotes realmente divergentes, isso nos leva a crer que a convergência entre o MD e o *corpus* de referência BP parece ser maior do que pensávamos, apresentando, assim, mais uma possível evidência da autenticidade do MD.

Os treze pacotes lexicais mais frequentes do MD[2] realmente divergentes são apresentados na Tabela 3.16.

Tabela 3.16 – Pacotes lexicais (amostra) realmente divergentes

BP falado	BP escrito
COM CAFÉ DA	CINEMA COMIGO NA
ANOS EU TERIA	AO CINEMA COMIGO
DE DEIXAR RECADO	HORA OFICIAL DE
ELA ME LIGAR	AJUDAR COM ESTA
LIGAR MAIS TARDE	COM ESTA LIÇÃO
CINEMA COMIGO NA	COMIGO NA QUINTA
VOCÊ TERIA FEITO	ENQUANTO VOCÊ COMPRA
AO CINEMA COMIGO	ESTA LIÇÃO DE
FAZER A CARTEIRINHA	MANDAR UM TORPEDO
FEITO ALGO DIFERENTE	PAPO PELO MSN
HORA OFICIAL DE	PRECISA TOMAR CAFÉ
LIÇÃO DE PORTUGUÊS	QUER DEIXAR RECADO
MAS NINGUÉM ATENDE	VOCÊ MORAVA AQUI

Dos pacotes apresentados na Tabela 3.16, os que nos chamaram a atenção foram os divergentes comparados ao BP falado ‘ligar mais tarde’ e ‘mas ninguém atende’ que parecem ser autênticos e característicos da linguagem oral (conversação/conversas telefônicas). Como o *corpus* de referência contém uma porcentagem menor de linguagem oral (por exemplo, os textos de conversas telefônicas somam aproximadamente 21.500 *tokens*), fizemos uma busca no Google[3] para verificar as ocorrências desses dois pacotes. Encontramos aproximadamente 402.000 páginas com uma ou mais menções do pacote ‘ligar mais

2 . Classificados por ordem de frequência no MD.
3. Disponível em: <www.google.com.br>. Acesso em: 23 jun. 2010.

tarde' e aproximadamente 127.000 páginas com uma ou mais menções de 'mas ninguém atende'. Assim, é possível que a divergência encontrada (pelo menos com relação a esses dois pacotes) não seja real, visto que uma rápida checagem no Google mostrou muitas ocorrências para os itens pesquisados, o que indica que os pacotes aparentemente são comuns.

Com relação à lista de pacotes lexicais divergentes comparados ao BP escrito (Tabela 3.16, à direita), muitos deles parecem ser mais característicos da linguagem falada (p. ex., 'quer deixar recado') e, consequentemente, não encontrados no *corpus* de referência escrito.

Sendo assim, acreditamos que precisaríamos de um estudo mais aprofundado dos pacotes lexicais divergentes para verificar o grau de autenticidade ou inautenticidade de todos os pacotes possivelmente divergentes e utilizarmos um *corpus* de referência falado maior[4].

A seguir, a última etapa da pesquisa: a análise de convergência cada texto/diálogo dos *subcorpora* MDNA e MDA.

3.3 Análise de convergência texto a texto

A convergência entre os textos do MD e o *corpus* de referência BP (falado e escrito) também foi analisada texto a texto com o objetivo de avaliar a variação da convergência entre os textos dos *subcorpora* de estudo autêntico e não autêntico. Em um primeiro momento, imaginamos que os textos do início do material didático poderiam conter menos trigramas convergentes (consequentemente, com um grau de autenticidade menor) do que aqueles do final. O *script* elaborado especialmente para esta análise nos retornou a quantidade de trigramas em cada texto (coluna 'trigramas – MD' na Tabela 3.17), quais destes eram convergentes ('trigramas convergentes (BP falado)') e sua porcentagem de convergência[5].

4. O maior *corpus* falado de português hoje é o *Corpus* Brasileiro, com 1 bilhão de *tokens.*
5. As listas completas de convergência de todos os textos estão disponíveis no Apêndice 5.

Tabela 3.17 – Porcentagem de convergência texto a texto da Unidade 6 do MD comparado ao BP falado

arquivo	trigramas – MD	trigramas convergentes (BP falado)	% de convergência
MDNA/Unidade 6_Lição A	79	25	31,6
MDNA/Unidade 6_Lição A_1	88	27	30,6
MDNA/Unidade 6_Lição A_2	64	18	28,1
MDNA/Unidade 6_Lição A_3	97	35	36
MDNA/Unidade 6_Lição B	72	30	41,6
MDNA/Unidade 6_Lição B_1	44	20	45,4
MDNA/Unidade 6_Lição B_2	27	5	18,5
MDNA/Unidade 6_Lição B_3	69	27	39,1
MDNA/Unidade 6_Lição C	40	5	12,5
MDNA/Unidade 6_Lição C_1	34	10	29,4
MDNA/Unidade 6_Lição C_2	39	9	23
MDNA/Unidade 6_Lição C_3	25	5	20
MDNA/Unidade 6_Lição C_4	71	9	12,6
MDNA/Unidade 6_Lição C_5	57	18	31,5
MDNA/Unidade 6_Lição ABC	111	32	28,8
MDA/Unidade 6_Lição ABC_L	42	15	35,7
	Unidade 6	**média**	**29,0**

Como dito anteriormente (seção 2.2.1.1 Divisão do *corpus* MD em autêntico e não autêntico), o conteúdo do *subcorpus* MDNA representa grande parte do livro didático e é composto por diálogos e roteiros elaborados para os CDs de áudio. Os textos considerados autênticos (MDA) correspondem aos textos da seção “Leitura” (na Tabela 3.17 representados pelo arquivo ‘MDA/Unidade 6_Lição ABC_L’). Somente por essa unidade, é possível observar que a grande maioria dos textos não autênticos apresenta percentual de convergência superior a 26,30% que, de acordo com a Faixa de Representatividade (ver seção 3.1), indica textos ‘autênticos’. Assim,

a partir dos dados fornecidos pelo *script*, foi possível calcularmos a média de convergência por unidade (ver Tabela 3.18).

Tabela 3.18 – Média de Convergência das Unidades do MD com o BP falado e escrito

Unidade	Média – % de convergência (BP falado)	Média – % de convergência (BP escrito)
1	26,6	28,8
2	23,9	32,3
3	25,2	28,4
4	21,2	24,1
Rev e Pron 1	29,9	20,1
5	22	30,1
6	29,0	29,5
7	21,6	29,3
8	28,8	24,6
Rev e Pron 2	23,5	26,5
9	25,8	20,9
10	21,3	27,7
11	29,4	23,8
12	31,2	22
Rev e Pron 3	26,6	29,9
13	29,2	18,8
14	31,6	23,3
15	33	20,2
16	27	21,8
Rev e Pron 4	38,3	19,9
17	26,3	23,3
18	36	25,2
19	31,3	21,2
20	35,1	20,9
Rev e Pron 5	35,5	22,4

Colocando os dados da Tabela 3.18 em gráficos, temos o seguinte:

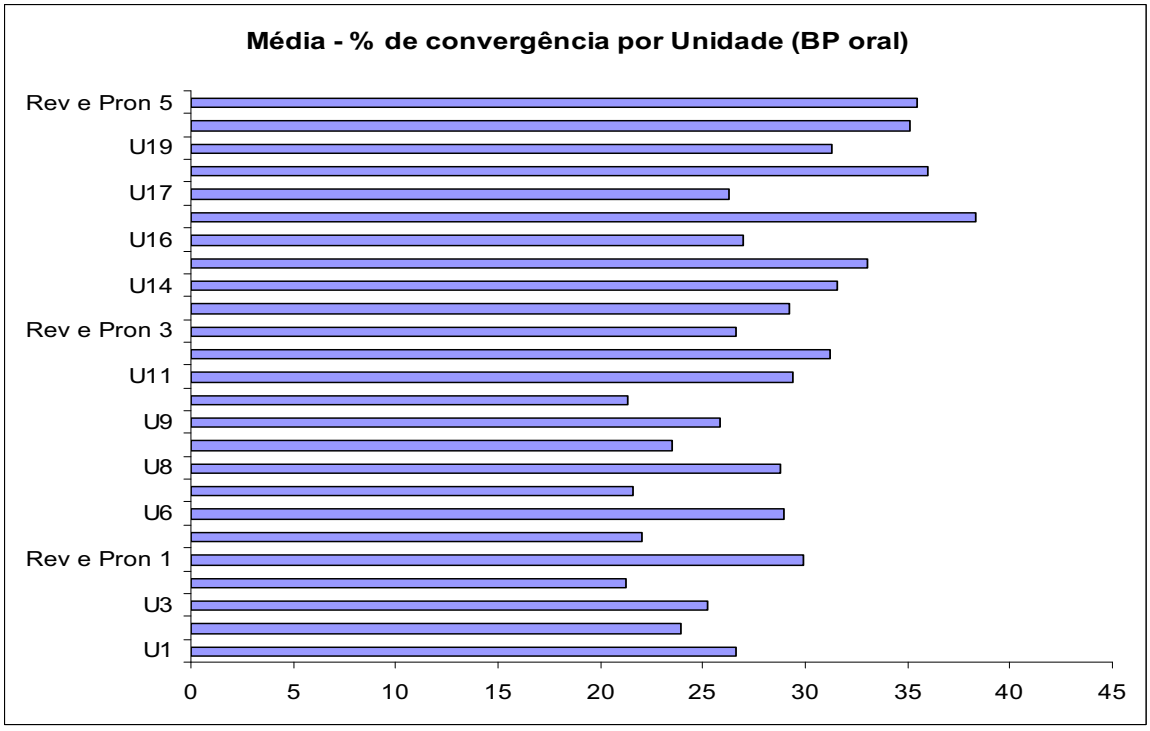


Gráfico 3.2: Média de convergência por unidade do MD comparado ao BP falado.

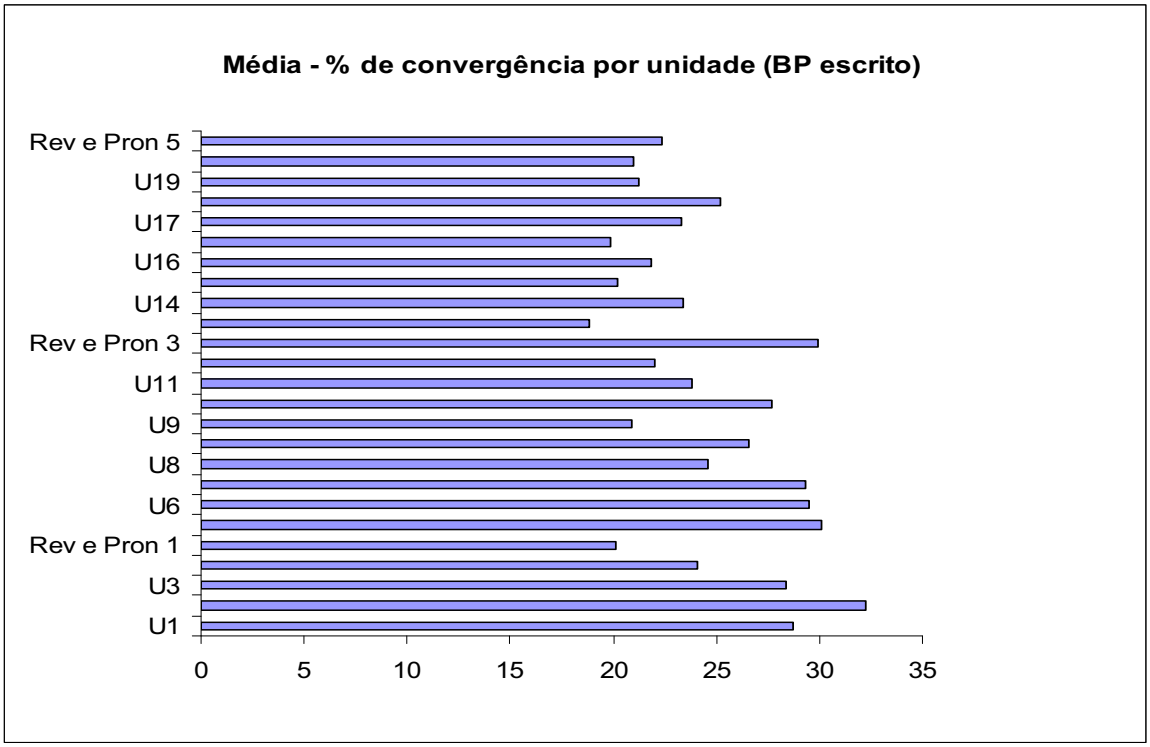


Gráfico 3.3: Média de convergência por unidade do MD comparado ao BP escrito.

Como é possível perceber pela Tabela 3.18 e pelos gráficos 3.2 e 3.3, e contrariando nossa intuição de que as unidades mais iniciais teriam um grau de autenticidade menor do que as unidades mais avançadas, há um equilíbrio de convergência em todas as unidades do MD, tanto quando comparamos a convergência com o BP falado e com o escrito, sendo que as unidades 4 e 10 (comparadas ao BP falado, 21,20% e 21,30%, respectivamente) tiveram convergência mais baixa e com convergência mais alta encontramos a unidade de revisão e pronúncia 4 e a unidade 18 (38,30% e 36,00%, respectivamente). Na comparação com o BP escrito, encontramos as unidades 13 e de revisão e pronúncia 4 (18,80% e 19,90%, respectivamente) com convergência mais baixa e as unidades 5 e 2 (30,10% e 32,30%, respectivamente) com convergência mais alta (Tabela 3.19, classificados por ordem decrescente de convergência).

Tabela 3.19 – Média de convergência das unidades do MD

Unidade	Média – % conv (BP falado)	Unidade	Média – % conv (BP escrito)
Revisão e Pronúncia 4	38,3	2	32,3
18	36	5	30,1
Revisão e Pronúncia 5	35,5	Revisão e Pronúncia 3	29,9
20	35,1	6	29,5
15	33	7	29,3
14	31,6	1	28,8
19	31,3	3	28,4
12	31,2	10	27,7
Revisão e Pronúncia 1	29,9	Revisão e Pronúncia 2	26,6
11	29,4	18	25,2
13	29,2	8	24,6
6	29	4	24,1
8	28,8	11	23,8

16	27	14	23,4
1	26,6	17	23,3
Revisão e Pronúncia 3	26,6	Revisão e Pronúncia 5	22,4
17	26,3	12	22,0
9	25,8	16	21,8
3	25,2	19	21,2
Revisão e Pronúncia 2	23,5	9	20,9
2	22,9	20	20,9
5	22	15	20,2
7	21,6	Revisão e Pronúncia 1	20,1
10	21,3	Revisão e Pronúncia 4	19,9
4	21,2	13	18,8

Outra surpresa foi a alta porcentagem de convergência das unidades de revisão e pronúncia (1, 5 e 4 com 29,90%, 35,50% e 38,20% de convergência, respectivamente) por se tratarem de unidades somente com textos não autênticos.

Dessa forma, parece ser correto considerarmos que os textos do MD, de maneira geral, possuem um alto grau de autenticidade conforme medido pela convergência de pacotes lexicais (mesmo os das unidades mais iniciais).

A seguir, verificaremos o grau de convergência/autenticidade de cada texto do *corpus* MD.

3.3.1 Grau de autenticidade dos textos

Como dito anteriormente, os valores estipulados pela faixa de representatividade (de 26,30% para textos da linguagem falada e 52,15% para textos da linguagem escrita), levam-nos a crer que o *corpus* MD (e seus *subcorpora* MDA e MDNA) é tão ‘autêntico’ quanto uma coletânea de textos autênticos. Verificamos também, ao

analisar unidade por unidade do material, que, diferente do que imaginamos, mesmo as unidades mais iniciais possuem uma porcentagem de convergência semelhante à porcentagem de convergência de unidades mais avançadas. Sendo assim, como etapa final da análise, verificaremos a seguir o grau de autenticidade texto por texto. Para tanto, como ponto inicial, nos baseamos na faixa de representatividade para elaborar graus de autenticidade para a classificação da convergência dos textos, o que nos levou aos seguintes números apresentados na Tabela 3.20.

Tabela 3.20 – Classificação da autenticidade

Grau de autenticidade			
Muito baixo	**Baixo**	**Bom**	**Alto**
0 a 10	11 a 20	21 a 30	31 em diante

Sendo assim, classificamos todos os textos[6] do MD com base na Tabela 3.20 e chegamos aos seguintes resultados apresentados na Tabela 3.21.

Tabela 3.21 – Classificação dos textos do MD de acordo com o grau de autenticidade

Grau de autenticidade	**Número de textos**	**% do total (MD vs. BP falado)**	**Número de textos**	**% do total (MD vs. BP escrito)**
Muito baixo	37	7,50%	45	9,20%
Baixo	82	16,80%	153	31,30%
Bom	200	40,90%	168	34,35%
Alto	170	34,80%	123	25,15%

6. As listas completas de classificação dos textos por grau de autenticidade estão disponíveis no Apêndice 6.

De acordo com a Tabela 3.21, a maior parte dos textos do MD encontra-se na faixa de autenticidade considerada ‘boa’, com 40,90% dos textos do MD comparados ao BP falado e 34,35% com o BP escrito e na faixa de autenticidade considerada ‘alta’, obtivemos 34,80% e 25,15% do total de textos (nas comparações com o BP falado e escrito, respectivamente). Somente 7,50% e 9,20% dos textos tiveram porcentagem de convergência considerada muito baixa e 16,80% e 31,30% deles obtiveram porcentagem baixa. Outro dado interessante é que vários dos textos autênticos do MD (comparados ao BP escrito) ficaram na faixa de autenticidade considerada baixa (aproximadamente 40%).

Sendo assim, acreditamos que esses achados corroboram os anteriores e podemos considerar que os textos não autênticos do material didático estão próximos da linguagem autêntica, representada pelo *corpus* de referência, visto que, na última análise elaborada, obtiveram um grau de autenticidade classificado como ‘bom-alto’ quando comparado ao *corpus* de referência BP (falado e escrito).

A seguir apresentamos nossas considerações finais.

CAPÍTULO 4

CONSIDERAÇÕES FINAIS

Após a apresentação de nossa análise, podemos tecer algumas considerações a respeito do material didático, baseadas nos resultados obtidos. Vale reiterar as questões por nós investigadas:

1. Quantos trigramas e pacotes lexicais existem nos textos (falados e escritos) do material didático?
2. Quais desses são convergentes (i.e., existem no *corpus* de referência) e divergentes (i.e., não existem no *corpus* de referência)?
3. A proporção de uso dos convergentes é equivalente nos *corpora*?
4. Com base nas respostas às perguntas acima, qual é o grau de autenticidade dos textos do material didático?

As perguntas 1 e 2 nos serviram como base inicial para a análise e seus resultados foram apresentados em detalhes no capítulo 3, mais especificamente nas seções 3.1.1 e 3.2. Tendo em vista a pergunta 3, a proporção de uso dos trigramas e pacotes lexicais convergentes, de acordo com as análises elaboradas na seção 3.1.2.1.1, indica que a maior parte dos trigramas convergentes foi sobreusada no MD, ou seja, eles aparecem muito mais vezes no material didático do que no *corpus* de referência. Contudo, a maior parte dos pacotes lexicais do MD de alta frequência também é altamente frequente no *corpus* de referência BP, o que aparentemente indica que os alunos estão sendo expostos a alguns dos pacotes comuns da linguagem autêntica.

Quanto à resposta da quarta pergunta, que constitui um resumo de toda a pesquisa, em linhas gerais, podemos sugerir que muitos textos utilizados no material didático analisado parecem possuir lexicogramática semelhante à de textos autênticos,.de acordo com os seguintes achados:

- os valores de convergência, i.e., os valores referentes à comparação do número de trigramas em comum entre o *corpus* MD e o de referência BP, ficaram acima dos valores estipulados pela faixa de representatividade (valores de referência mínimos para que um texto possa ser considerado 'autêntico');
- a recontagem dos pacotes lexicais realmente divergentes (para a retirada de trigramas com características específicas que impossibilitavam a verificação exata da convergência com o *corpus* de referência) nos leva a crer que a convergência entre o *corpus* MD e o *corpus* de referência BP parece ser maior do que havíamos estipulado nas análises anteriores;
- muitos dos pacotes lexicais mais frequentes do *corpus* MD também o são no *corpus* de referência BP (ver seção 3.2.1);
- não há variações muito grandes de convergência entre as unidades e os textos do MD, ou seja, tanto as unidades mais iniciais como as mais avançadas possuem um alto grau de semelhança com textos autênticos conforme medido pela convergência de pacotes lexicais;
- de acordo com a classificação do grau de autenticidade (seção 3.3.1), o MD apresenta a maior parte de seus textos na faixa considerada 'boa-alta' quando comparada ao *corpus* de referência BP (falado e escrito).

Ao mesmo tempo, obtivemos também os seguintes achados:

- convergência maior dos trigramas do *subcorpus* material didático não autêntico (MDNA) com o *subcorpus* escrito do *corpus* de referência BP, ou seja, mais trigramas em comum com o BP escrito e não com o BP falado.
- vários dos textos autênticos do MD (comparados ao BP escrito) ficaram na faixa de autenticidade considerada baixa (aproximadamente 40%).

Com base nos resultados expostos acima, podemos concluir que nem todo texto não autêntico é um mau exemplo de lexicogramática. Há textos não autênticos que trazem muitas ocorrências de pacotes lexicais recorrentes na linguagem autêntica e esse resultado corrobora os achados de Contrera (2010) com a língua inglesa. Antes desta pesquisa, tinha-se a crença de que por não ser autêntico o texto invariavelmente seria questionável para o ensino de língua, aos olhos dos proponentes do uso da linguagem autêntica. Mas os resultados desta pesquisa

indicam que alguns textos não autênticos podem ser bons veículos para o contato com a lexicogramática autêntica. Contudo, os textos não autênticos analisados que tentam representar a linguagem falada apresentam mais pacotes característicos da linguagem escrita. Isso revela suas condições de produção, visto que os textos não autênticos falados foram primeiramente escritos para serem lidos e interpretados por atores.

Ao mesmo tempo, com relação aos textos autênticos, o fato de ser autêntico por si só não garante que os pacotes lexicais que ele contenha sejam típicos da linguagem falada ou escrita. Há textos autênticos mais e menos típicos, mais e menos usuais. A metodologia de convergência aqui desenvolvida pode ser um instrumento para o professor mensurar esse grau de tipicidade da lexicogramática de um texto. De posse dos resultados dessa análise, o professor e/ou autor de materiais didáticos pode fazer, possivelmente, melhores escolhas do que faria apenas impressionisticamente por meio da leitura dos textos candidatos a figurar nos materiais ou nas aulas. No final das contas, pode-se dizer que o que vale é encontrar textos que satisfaçam as necessidades de contextos variados de ensino. Se a necessidade for encontrar textos mais próximos da fala, então uma maneira de fazer isso pode ser por meio do cálculo da convergência de pacotes em contraste com um *corpus* de linguagem falada autêntica; se for preciso encontrar textos mais próximos da linguagem escrita, então o cálculo deve ser feito com um *corpus* de linguagem escrita autêntica. Concordamos que, no momento, a aplicação da metodologia aqui desenvolvida pode ser complexa demais para a maior parte dos professores. Para popularizar nossa metodologia como um instrumento para auxílio do professor na seleção de textos, seria necessário desenvolver um software que automatizasse e integrasse as várias comparações e demais tipos de processamento de *corpora* envolvidos. No entanto, devido às limitações inerentes a um estudo de mestrado, essa etapa permanece como proposta de futura pesquisa.

Sendo assim, o trabalho aqui descrito espera ter contribuído para um melhor entendimento da complexidade da questão da autenticidade de textos na esfera do ensino de língua estrangeira. De modo mais específico, esperamos ter avançado na discussão de algumas questões no âmbito da área de Linguística de *Corpus*

aplicada ao ensino de língua estrangeira ao desenvolver uma metodologia de identificação de autenticidade em *corpora* de textos autênticos e não autênticos.

REFERÊNCIAS BIBLIOGRÁFICAS

AIJIMER, K. (Ed.). *Corpora and Language Teaching*. Amsterdam: John Benjamins, 2009.

ALAMBERT, E. *Uma tradução premiada sob a perspectiva da Linguística de Corpus*. Dissertação de Mestrado. São Paulo: PUC-SP, 2008.

ALENCAR, R. A. *E aí?* Uma proposta descritiva das expressões formulaicas para português L2 para estrangeiros. Tese de Doutorado. Rio de Janeiro: PUC-RJ, 2004.

ARAUJO, L. D. *Brasil brasileiro*: o léxico e a identidade nacional. Tese de Doutorado. Rio de Janeiro: UERJ, 2010.

ALLAN, R. Can a graded reader *corpus* provide 'authentic' input? *ELT Journal*, v. 63(1), p. 23-32, 2009.

AMADO, R. S. O ensino e a pesquisa de português para falantes de outras línguas. *Guavira Letras*, v. 6, p. 67-75, 2008.

BEARZOTI FILHO, P. A palavra que não para de crescer. *Discutindo Língua Portuguesa*, ano I, n. 1, p. 30, 2008.

BEAUGRANDE, R. de. Reconnecting real language with real texts: text linguistics and corpus linguistics. *International Journal of Corpus Linguistics*, 4(2), 243-260, 1999.

BERBER SARDINHA, T. 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*. Pelotas: ALAB/EDUCAT, 2000, p. 45-72.

_____. Beginning Portuguese Corpus Linguistics: exploring a *corpus* to teach Portuguese as a Foreign Language. *D.E.L.T.A.*, v. 15, n. 2, p. 289-299, 1999.

_____. Concordancing Portuguese. Apresentação em PowerPoint. Birmingham: University of Birmingham, 1997.

_____. *Linguística de Corpus*. Barueri: Manole, 2004.

_____. Preparação de material didático para Aprendizagem Baseada em Tarefas com WordSmith Tools e *corpora*. *Calidoscópio*, v. 4, n. 3, p. 148-155, 2006.

_____. The book is not on the table: autenticidade e idiomaticidade do texto para ensino de inglês na perspectiva da Linguística de Corpus. In: DAMIANOVIC, M. C. (Org.). *Material didático*: elaboração e avaliação. Taubaté: Cabral, 2007.

_____. *Pesquisa em Linguística de Corpus com Wordsmith Tools*. Mercado de Letras, 2009.

_____; SHEPHERD, T. An online system for error identification in Brazilian learner English. *Anais do 8th Teaching and Language Corpora Conference.* Lisboa: Associação de Estudos e de Investigação Científica do ISLA-Lisboa. p. 257-262, 2008.

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

_____. *University Language*: A Corpus-Based Study of Spoken and Written Registers. Amsterdam: John Benjamins, 2006.

_____; JOHANSSON, S.; LEECH, G. et al. *Longman Grammar of Spoken and Written English*. London: Longman, 1999.

_____; CONRAD, S.; CORTES, V. *If You Look At…*: Lexical Bundles in University Teaching and Textbooks. Oxford: Oxford University Press, 2004.

_____; _____; REPPEN, R. *Corpus Linguistics*. Investigating Structure and Use. Cambridge: Cambridge University Press, 1998.

BRAUN, S.; KOHN, K; MUKHERJEE, J. (Eds.). *Corpus Technology and Language Pedagogy*. New York: Peter Lang, 2006.

BREEN, M. P. Authenticity in the Language Classroom. *Applied Linguistics*, v. 6 n. 1. p. 60-70, 1985.

BROWN, S.; MENASCHE, L. Defining Authenticity. Disponível em: <http://www.as.ysu.edu/~english/faculty/brown/personal/BrownMenasche.doc>. Acesso em: 8 jun. 2010.

CARVALHO, O. L. S. Aspectos da identidade brasileira em livros didáticos de português para estrangeiros: um estudo lexical. Disponível em: <http://www.onda.eti.br/revistaintercambio/conteudo/arquivos/1771.pdf>. Acesso em: 20 maio 2010.

CAVALCANTE, C. *Formas verbais em um livro didático de português para estrangeiros*: uma análise baseada em *corpus*. Dissertação de Mestrado. São Paulo: PUC-SP, 2006.

CONRAD, S. Corpus Linguistics and L2 teaching. In: HINKEL, E. *Handbook of Research in Second Language Teaching and Learning*. New Jersey: Lawrence Erlbaum, 2005, p. 393-409.

CONTRERA, S. *Autenticidade em livros didáticos para o ensino de inglês como língua estrangeira*: um estudo diacrônico sob a perspectiva da linguística de *corpus*. Dissertação de Mestrado. São Paulo: PUC-SP, 2010.

COOK, G. Discourse. In: CANDLIN, C. N. WIDDOWSON, H. G. (Eds.). *Language Teaching*: A Scheme for Teacher Education. Oxford: Oxford University Press, 1989.

COMET – Corpus Multilíngue para Ensino e Tradução. Disponível em: <http://www.fflch.usp.br/dlm/comet/>. Acesso em: 30 set. 2009.

CORTES, V. Teaching lexical bundles in the disciplines: an example from a writing intensive history class. *Science Direct – Linguistics and Education*, v. 17, p. 391-406, 2006.

COWIE, A. P. Introduction. In: _____ (Org.). Phraseology – Theory, Analysis, and Application. Oxford: Clarendon Press, 1998, p. 1-22.

DAY, R. R. A critical look at authentic materials. *The journal of Asia TEFL*. v. 1, n. 1, p. 101-114, 2004.

DELL'SOLA, R. L. A multimídia aplicada ao ensino do Português-Língua Estrangeira. In: JÚDICE, N. *Português para estrangeiros* – perspectivas de quem ensina. Niterói: Intertexto, 2002, p. 9-27.

FERREIRA, A. B. H. *Novo Aurélio século XXI*. 3. ed. Curitiba: Positivo, 2004.

FERNANDES, G.; SÃO BENTO FERREIRA, T.; RAMOS, V. *Muito prazer* – fale o português do Brasil. São Paulo: Disal, 2008.

FOX, G. Using *corpus* data in the classroom. In: TOMLINSON, B. *Materials Development in Language Teaching*. Cambridge: Cambridge University Press, 1998.

GABRIELATOS, C. Corpora and language teaching: just a fling or wedding bells? *TESL-EJ*. v. 8, n. 4, p. 1-37, 2005.

_____. Corpus-based evaluation of pedagogical materials: if-conditionals in ELT coursebooks and the BNC. 7th Teaching and Language Corpora Conference, 1º-4 jul. 2006, França (trabalho não publicado).

_____. Grammar, grammars and intuitions in ELT: A second opinion. *IATEFL Issues*, dez. 2002/jan. 2003.

GAVIOLI, L.; ASTON, G. Enriching reality: language corpora in language pedagogy. *ELT Journal*. v. 55(3), p. 238-246, 2001.

GILLMORE, A. A comparison of textbook and authentic interactions. *ELT Journal*, v. 58(4), p. 363-374, 2004.

GOMES DE MATTOS, F. Quando a prática precede a teoria: a criação do PBE. In: ALMEIDA FILHO, J. C. P de; LOMBELLO, L. C. (Orgs.). *O ensino de português para*

estrangeiros: pressupostos para o planejamento de cursos e elaboração de materiais. 2. ed. Campinas: Pontes, 1997, p. 11-17.

GUARIENTO, W.; MORLEY, J. Text and task authenticity in the EFL classroom. *ELT Journal*. v. 55(4), p. 347-353, 2001.

HADLEY, G. An introduction to data-driven learning. *RELC Journal*, 33(2), p. 99-124, 2002.

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

_____. Language as system and language as instance: the *corpus* as a theoretical construct. In: SVARTVIK, J. (Org.). *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, 1992, p. 61-78.

HARWOOD, N. Taking a lexical approach to teaching: principles and problems. *International Journal of applied linguistics*, v. 12, n. 2, 2002.

HOEY, M. *Lexical Priming*: A New Theory of Words and Language. London: Routledge, 2005.

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

_____; FRANCIS, G. Verbs observed: a corpus-driven pedagogic grammar. *Applied Linguistics*, 19 (1), p. 45-72, 1998.

HUTCHINSON, A.; LLOYD, J. *Portuguese*: An Essential Grammar. 2. ed. London: Routledge, 2003.

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

_____. As can be seen: lexical bundles and disciplinary variation. *Science Direct – English for Specific Purposes*, v. 27, p. 4-21, 2008.

ILLÉS, E. What makes a coursebook series stand the test of time? *ELT Journal*, v. 63(2), p. 145-153, abr. 2009.

JÚDICE, N. Representações do Brasil dos anos 40 e 90 em textos de materiais didáticos para o ensino de português para estrangeiros de português para estrangeiros. Disponível em: <http://www.letras.puc-rio.br/Publicacoes/ccci/artigos.html>. Acesso em: 10 dez. 2008.

KENNEDY, G. *An Introduction to Corpus Linguistics*. London: Longman, 1998.

KOPROWSKI, M. Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal*. v. 59(4), p. 322-332, 2005.

LEECH, G. Corpora and theories of linguistic performance. In: SVARTVIK, J. *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, 1992, p. 105-122.

LEITE, M. Do falado ao escrito e vice-versa. In: PRETI, D. *Oralidade em textos escritos*. São Paulo: Humanitas, 2009.

LEWIS, M. *Implementing the Lexical Approach* – Putting Theory into Practice. São Paulo: LTP, 1997.

_____. There is nothing as practical as a good theory. In: _____ (Org.). *Teaching Collocation* – Further Developments in the Lexical Approach. Hove: LTP, 2000, p. 10-27.

LÚCIO, D.D. A *relexicalização de adjetivos nas redações de alunos de inglês*: um estudo baseado em *corpus* de aprendiz. Dissertação de Mestrado. São Paulo: PUC-SP, 2006.

MACDONALD, M. N.; BADGER, R.; DASLI, M. Authenticity, culture and language learning. *Language and Intercultural Communication*, v. 6, n. 3 & 4, 2006.

MEDEIROS, A. A. D. de. O português no mundo. Disponível em: <http://www.linguaportuguesa.ufrn.br/pt_3.php>. Acesso em: 7 jul. 2010.

MINDT, D. English *corpus* linguistics and the foreign language teaching syllabus. In: THOMAS, J.; SHORT, M. (Eds.). *Using Corpora for Language Research*. London: Longman, 1996, p. 232-47.

MIRA MATEUS, M. H. Difusão da língua portuguesa no mundo. Disponível em: <http://www.fflch.usp.br/dlcv/lport/pdf/mes/01.pdf>. Acesso em: 13 jul. 2010.

MISHAN, F. Authenticating corpora for language learning: a problem and its resolution. *ELT Journal*, v. 58(3), 2004.

_____. *Designing Authenticity into Language Learning Materials*. Bristol: Intellect, 2004.

MORITA, M. K. (Re)pensando sobre o material didático de PLE. In: SILVEIRA, R. C. P. da (Org.). *Português língua estrangeira*: perspectivas. São Paulo: Cortez, 1998, p. 59-72.

MORROW, K. Authentic texts in ESP. In: HOLDEN, S. (Ed.). *English for Specific Purposes.* London: Modern English Publications, 1977.

MURPHY, J. Task-based learning: the interaction between tasks and learners. *ELT Journal*. v. 57(4), p. 352-360, 2003.

NEKRASOVA, T. English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning*, 59(3), p. 647-686, 2009.

NUNAN, D. *Designing Tasks for the Communicative Classroom*. Cambridge: Cambridge University Press, 1989.

O'KEEFFE, A.; MCCARTHY, M.; CARTER, R. *From Corpus to Classroom. Language Use and Language Teaching*. Cambridge: Cambridge University Press, 2007.

PAES DE ALMEIDA FILHO, J. C. Índices nacionais de desenvolvimento do ensino de português língua estrangeira. In: _____; CAVALCANTI CUNHA, M. J. *Projetos iniciais em português para falantes de outras línguas*, Campinas: Pontes, 2007, p. 39-55.

PICASSO, R. A. *Uma contribuição da linguística de corpus para a fonologia*: um estudo de colocações e aspectos segmentais das vogais da língua inglesa. Dissertação de Mestrado. São Paulo: PUC-SP, 2005.

PONCE, M. H. *Tudo bem?* Português para a nova geração. São Paulo: SBS, 2002, v. 2.

PREACHER, K. J. *Calculation for the chi-square test*: an interactive calculation tool for chi-square tests of goodness of fit and independence. Programa de computador. Disponível em: <http://www.quantpsy.org>. Acesso em: 7 jul. 2010.

RICHARDS, J. C.; RODGERS, T. S. *Approaches and Methods in Language Teaching*. 2. ed. Cambridge: Cambridge University Press, 2001.

SCOTT, M. WordSmith Tools. Programa de computador. Oxford: Oxford University Press, 1997.

_____; TRIBBLE, C. *Textual Patterns*: Keywords and Corpus Analysis in Language Education. Amsterdam: John Benjamins, 2006.

SHEPHERD, T. *Corpora* de aprendiz de língua estrangeira: um estudo contrastivo de n-gramas. *Veredas On-Line – Linguística de Corpus e Computacional*, v. 13, p. 100-116, 2/2009.

_____; ZYNGIER, S Identidades sociais e linguística de corpus: um estudo de três contextos sociais. *Revista da Abralin*, s.d. (no prelo).

_____; _____; VIANA, V. A tale of two cities: lexical bundles as indicators of linguistic choices and socio-cultural traces. In: JEFFRIES, L.; MCINTIRE, D.; BOUSFIELD, D. (Eds.). *Stylistics and Social Cognition*. Amsterdam: Rodopi, 2007.

_____; _____; _____. Feixes lexicais e visões de mundo: um estudo sobre *corpus*. *Matraga*, v. 13, p. 125-140, 2006.

_____; VIANA, V. A Linguística de corpus e a sala de aula de língua estrangeira: interfaces. *Cadernos do CNFL*, v. IX, n. 3, 2006. Disponível em: <http://www.filologia.org.br/xcnlf/5/02.htm>. Acesso em: 13 jul. 2010.

SHOMOOSSI, N; KETABI, S. A critical look at the concept of authenticity. *Electronic Journal of Foreign Language Teaching*, v. 4, n.1, p. 149-155, 2007.

SHORTALL, T. The L2 syllabus: *corpus* or contrivance? *Corpora*, v. 2, p. 157-185, 2007.

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

_____. *Corpus* evidence in language description. In: WICHMANN, A.; FLIGELSTONE, S.; MCENERY, T.; KNOWLES, G. (Eds.). *Teaching and Language Corpora*. New York: Addison Wesley Longman, 1997, p. 27-39.

_____. *Reading Concordances*. London: Pearson Longman, 2003.

SOUZA, A. P. K.; GAMA, E. M. P. O ensino de língua portuguesa foi além do limite: uma discussão teórica e metodológica sobre o ensino de PLE. In: MEYER, R. M. B.; REBELO, I. M. M. (Orgs.). *Português para estrangeiros*: entre a área de estudos e pesquisa e a prática profissional cotidiana. Disponível em: <http://www.letras.puc-rio.br/publicacoes/ccci/artigos.html>. Acesso em: 13 jul. 2010.

TAGNIN, S. *O jeito que a gente diz*: expressões convencionais e idiomáticas inglês-português. São Paulo: Disal, 2005.

TATSUKI, D. What is authenticity? *Authentic Communication*: Proceedings of the 5th Annual JALT Pan-SIG Conference. 13-14 maio 2006. Shizuoka, Japan: Tokai University College of Marine Science, p.1-15.

TAYLOR, D. Inauthentic authenticity or authentic inauthenticity? *TESL-EJ*, v. 1, n. 2, ago. 1994.

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

TOMLINSON, B. *Developing Materials for Language Teaching*. London: Continuum, 2002.

TRIBBLE, C.; JONES, G. *Concordances in the Classroom* – A Resource Book for Teachers. London: Longman, 1990.

WILKINS, D. *Notional Syllabuses*. Oxford: Oxford University Press, 1976.

APÊNDICES E ANEXOS

A. APÊNDICES

Apêndice 1

Tabela com o número de palavras, formas e razão palavras/formas do *corpus* de estudo

Nome do arquivo	*Tokens*	*Types*	*Type/ Token Ratio*
TOTAL	**40.815**	**1.672**	**4,1**
UN1_LIA.TXT	246	75	30,49
UN1_LIB.TXT	280	43	15,36
UN1_LIC.TXT	84	16	19,05
UN1_LIABC.TXT	227	49	21,59
UN2_LIA.TXT	194	45	23,2
UN2_LIB.TXT	221	72	32,58
UN2_LIC.TXT	106	28	26,42
UN2_LIABC.TXT	242	76	31,4
UN3_LIA.TXT	173	49	28,32
UN3_LIB.TXT	264	88	33,33
UN3_LIC.TXT	358	118	32,96
UN3_LIABC.TXT	274	121	44,16
UN4_LIA.TXT	543	177	32,6
UN4_LIB.TXT	336	161	47,92
UN4_LIC.TXT	411	198	48,18
UN4_LIABC.TXT	524	258	49,24
PR_1.TXT	55	20	36,36
REV_1.TXT	401	188	46,88
UN5_LIA.TXT	177	60	33,9
UN5_LIB.TXT	328	131	39,94
UN5_LIC.TXT	517	138	26,69
UN5_LIABC.TXT	420	166	39,52
UN6_LIA.TXT	434	206	47,47
UN6_LIB.TXT	272	124	45,59
UN6_LIC.TXT	459	207	45,1
UN6_LIABC.TXT	231	116	50,22
UN7_LIA.TXT	459	201	43,79
UN7_LIB.TXT	94	39	41,49
UN7_LIC.TXT	586	283	48,29
UN7_LIABC.TXT	377	187	49,6
UN8_LIA.TXT	413	177	42,86
UN8_LIB.TXT	511	232	45,4
UN8_LIC.TXT	123	63	51,22
UN8_LIABC.TXT	380	158	41,58
PR_2.TXT	61	24	39,34
REV_2.TXT	439	197	44,87
UN9_LIA.TXT	412	197	47,82
UN9_LIB.TXT	351	176	50,14

UN9_LIC.TXT	413	194	46,97
UN9_LIABC.TXT	426	244	57,28
UN10_LIA.TXT	543	249	45,86
UN10_LIB.TXT	428	211	49,3
UN10_LIC.TXT	309	167	54,05
UN10_LIABC.TXT	499	269	53,91
UN11_LIA.TXT	352	198	56,25
UN11_LIB.TXT	487	267	54,83
UN11_LIC.TXT	674	405	60,09
UN11_LIABC.TXT	516	284	55,04
UN12_LIA.TXT	547	260	47,53
UN12_LIB.TXT	699	418	59,8
UN12_LIC.TXT	604	363	60,1
UN12_LIABC.TXT	845	507	60
PR_3.TXT	103	56	54,37
REV_3.TXT	465	218	46,88
UN13_LIA.TXT	482	279	57,88
UN13_LIB.TXT	658	383	58,21
UN13_LIC.TXT	578	350	60,55
UN13_LIABC.TXT	812	460	56,65
UN14_LIA.TXT	408	202	49,51
UN14_LIB.TXT	519	244	47,01
UN14_LIC.TXT	403	180	44,67
UN14_LIABC.TXT	422	200	47,39
UN15_LIA.TXT	609	288	47,29
UN15_LIB.TXT	539	313	58,07
UN15_LIC.TXT	750	418	55,73
UN15_LIABC.TXT	752	493	65,56
UN16_LIA.TXT	486	265	54,53
UN16_LIB.TXT	518	276	53,28
UN16_LIC.TXT	416	192	46,15
UN16_LIABC.TXT	800	540	67,5
PR_4.TXT	107	63	58,88
REV_4.TXT	811	433	53,39
UN17_LIA.TXT	463	463	61,34
UN17_LIB.TXT	712	466	65,45
UN17_LIC.TXT	422	254	60,19
UN17_LIABC.TXT	723	436	60,3
UN18_LIA.TXT	386	242	62,69
UN18_LIB.TXT	411	258	62,77
UN18_LIC.TXT	662	434	65,56
UN18_LIABC.TXT	779	501	64,31
UN19_LIA.TXT	417	239	57,31
UN19_LIB.TXT	585	371	63,42
UN19_LIC.TXT	524	308	58,78
UN19_LIABC.TXT	1.125	704	62,58
UN20_LIA.TXT	527	305	57,87
UN20_LIB.TXT	948	607	64,03
UN20_LIC.TXT	695	421	60,58
UN20_LIABC.TXT	790	481	60,89
PR_5.TXT	126	63	50
REV_5.TXT	557	323	57,99

Apêndice 2

Lista de estatísticas – Material autêntico no *corpus* MD

N	Nome do arquivo	*Tokens*	*Types*	*Type/ Token Ratio*
	OVERALL	5.393	92	1,71
1	UN3LIA~1	105	64	60,95
2	UN3LIC_P	47	27	57,45
3	UN4LIA~1	208	109	52,4
4	UN6LIA~1	79	51	64,56
5	UN7LIA~1	205	121	59,02
6	UN8_LI~1	157	49	31,21
7	UN9_LI~1	212	155	73,11
8	UN10LI~1	202	134	66,34
9	UN11LI~1	180	105	58,33
10	UN11LI~2	216	140	64,81
11	UN12LI~1	387	267	68,99
12	UN13LI~1	447	273	61,07
13	UN14LI~1	120	74	61,67
14	UN15LI~1	431	320	74,25
15	UN16LI~1	459	340	74,07
16	UN16LI~2	158	84	53,16
17	UN16LI~3	112	47	41,96
18	UN17LI~1	264	165	62,5
19	UN17LI~2	200	158	79
20	UN18LI~1	297	223	75,08
21	UN19LI~1	330	237	71,82
22	UN20LIB	175	125	71,43
23	UN20LI~1	402	252	62,69

Apêndice 3

Lista de estatísticas – Material não autêntico no *corpus* MD

N	Nome do arquivo	*Tokens*	*Types*	*Type/Token Ratio*
	OVERALL	35.429	1.512	4,27
1	UN1_LIA	246	75	30,49
2	UN1_LIB	280	43	15,36
3	UN1_LIC	84	16	19,05
4	UN1_LI~1	227	49	21,59
5	UN2_LIA	194	45	23,2
6	UN2_LIB	221	72	32,58
7	UN2_LIC	106	28	26,42
8	UN2_LI~1	242	76	31,4
9	UN3_LIA	173	49	28,32
10	UN3_LIB	264	88	33,33
12	UN3_LIC	311	91	29,26
13	UN3_LI~1	169	57	33,73
14	UN4_LIA	543	177	32,6
15	UN4_LIB	336	161	47,92
16	UN4_LIC	411	198	48,18
17	UN4_LI~1	316	149	47,15
18	UN5_LIA	177	60	33,9
19	UN5_LIB	328	131	39,94
20	UN5_LIC	517	138	26,69
21	UN5_LI~1	420	166	39,52
22	UN6_LIA	434	206	47,47
23	UN6_LIB	272	124	45,59
24	UN6_LIC	459	207	45,1
25	UN6_LI~1	152	65	42,76
26	UN7_LIA	459	201	43,79
27	UN7_LIB	94	39	41,49
28	UN7_LIC	586	283	48,29
29	UN7_LI~1	172	66	38,37
30	UN8_LIA	413	177	42,86
31	UN8_LIB	511	232	45,4
32	UN8_LIC	123	63	51,22
33	UN8_LI~1	223	109	48,88
34	UN9_LIA	412	197	47,82
35	UN9_LIB	351	176	50,14
36	UN9_LIC	413	194	46,97
37	UN9_LI~1	214	89	41,59
38	UN10_LIA	543	249	45,86
39	UN10_LIB	428	211	49,3
40	UN10_LIC	309	167	54,05
41	UN10_L~1	297	136	45,79
42	UN11_LIA	352	198	56,25
43	UN11_LIB	487	267	54,83
44	UN11_LIC	458	265	57,86
45	UN11_L~1	336	179	53,27
46	UN12_LIA	547	260	47,53

47	UN12_LIB	699	418	59,8
48	UN12_LIC	604	363	60,1
49	UN12_L~1	458	240	52,4
50	UN13_LIA	482	279	57,88
51	UN13_LIB	658	383	58,21
52	UN13_LIC	584	354	60,62
53	UN13_L~1	365	187	51,23
54	UN14_LIA	408	202	49,51
55	UN14_LIB	519	244	47,01
56	UN14_LIC	403	180	44,67
57	UN14_L~1	302	126	41,72
58	UN15_LIA	609	288	47,29
59	UN15_LIB	539	313	58,07
60	UN15_LIC	750	418	55,73
61	UN15_L~1	321	174	54,21
62	UN16_LIA	329	180	54,71
63	UN16_LIB	406	231	56,9
64	UN16_LIC	416	192	46,15
65	UN16_L~1	341	200	58,65
66	UN17_LIA	463	284	61,34
67	UN17_LIB	512	313	61,13
68	UN17_LIC	422	254	60,19
69	UN17_L~1	459	271	59,04
70	UN18_LIA	386	242	62,69
71	UN18_LIB	411	258	62,77
72	UN18_LIC	662	434	65,56
73	UN18_L~1	482	278	57,68
74	UN19_LIA	417	239	57,31
75	UN19_LIB	585	371	63,42
76	UN19_LIC	524	308	58,78
77	UN19_L~1	795	467	58,74
78	UN20_LIA	527	305	57,87
79	UN20_LIB	773	482	62,35
80	UN20_LIC	695	421	60,58
81	UN20_L~1	388	231	59,54
82	PR_1	55	20	36,36
83	PR_2	61	24	39,34
84	PR_3	103	56	54,37
85	PR_4	107	63	58,88
86	PR_5	126	63	50
87	REV_1	401	188	46,88
88	REV_2	439	197	44,87
89	REV_3	465	218	46,88
90	REV_4	811	433	53,39
91	REV_5	557	323	57,99

Apêndice 4

Listas de pacotes lexicais

MDA vs. BP oral **MD autêntico vs. BP oral – pacotes lexicais encontrados (NC= 20x por milhão)** Total de pacotes em comum = 62, Subusados = 0, de Uso equivalente = 0, Sobreusados = 62							
Trigrama	**Freq MDA absoluta**	**Freq MDA normalizada**	**Freq MD total normalizada**	**Freq BP oral absoluta**	**Freq BP oral normalizada**	**Razão**	**Classificação**
A_FIM_DE	1	185,19	4538,85	15103	175,62	25,85	Sobreuso
MAIS_DE_#	2	370,37	9077,71	25063	291,43	31,15	
DE_TODOS_OS	1	185,19	4538,85	12080	140,47	32,31	
RIO_DE_JANEIRO	2	370,37	9077,71	23074	268,30	33,83	
DE_#_DE	1	185,19	4538,85	11529	134,06	33,86	
MAIS_DO_QUE	1	185,19	4538,85	9818	114,16	39,76	
DE_#_E	1	185,19	4538,85	7953	92,48	49,08	
DE_#_QUE	1	185,19	4538,85	7384	85,86	52,86	
CERCA_DE_#	2	370,37	9077,71	14259	165,80	54,75	
EM_QUE_O	1	185,19	4538,85	6699	77,90	58,27	
DE_#_#	4	740,74	18155,41	26728	310,79	58,42	
PARA_QUE_A	1	185,19	4538,85	6654	77,37	58,66	
O_QUE_SE	1	185,19	4538,85	5579	64,87	69,97	
DO_MEIO_AMBIENTE	1	185,19	4538,85	5566	64,72	70,13	
DE_#_MIL	3	555,56	13616,56	15902	184,91	73,64	
DE_#_O	1	185,19	4538,85	4935	57,38	79,10	
QUE_O_BRASIL	2	370,37	9077,71	8791	102,22	88,80	
DEZEMBRO_DE_#	1	185,19	4538,85	4366	50,77	89,40	
PARA_QUE_O	2	370,37	9077,71	8685	100,99	89,89	

DE_#_A	3	555,56	13616,56	12115	140,87	96,66	
PARA_#_#	1	185,19	4538,85	3825	44,48	102,05	
A_PARTIR_DO	1	185,19	4538,85	3686	42,86	105,90	
FOLHA_DE_S	1	185,19	4538,85	3503	40,73	111,43	
AO_MESMO_TEMPO	2	370,37	9077,71	6855	79,71	113,89	
MERCADO_DE_TRABALHO	1	185,19	4538,85	3408	39,63	114,54	
JANEIRO_DE_#	1	185,19	4538,85	3181	36,99	122,71	
POR_FALTA_DE	1	185,19	4538,85	3170	36,86	123,14	
UM_DOS_MAIS	1	185,19	4538,85	3004	34,93	129,94	
DE_JANEIRO_E	1	185,19	4538,85	2997	34,85	130,24	
AO_LONGO_DE	1	185,19	4538,85	2932	34,09	133,13	
DE_#_MINUTOS	1	185,19	4538,85	2737	31,83	142,62	
DE_JANEIRO_DE	1	185,19	4538,85	2520	29,30	154,90	
DIA_A_DIA	1	185,19	4538,85	2506	29,14	155,76	
JUNHO_DE_#	1	185,19	4538,85	2483	28,87	157,21	
AOS_#_ANOS	1	185,19	4538,85	2396	27,86	162,91	
NOVEMBRO_DE_#	1	185,19	4538,85	2354	27,37	165,82	
DE_ACORDO_COM	5	925,93	22694,26	11712	136,19	166,64	
A_PROPOSTA_DE	1	185,19	4538,85	2336	27,16	167,10	
A_CIDADE_DE	1	185,19	4538,85	2246	26,12	173,79	
NO_MERCADO_DE	1	185,19	4538,85	2093	24,34	186,50	
A_CAPACIDADE_DE	1	185,19	4538,85	2025	23,55	192,76	
HOMENS_E_MULHERES	1	185,19	4538,85	1941	22,57	201,10	
E_AO_MESMO	1	185,19	4538,85	1931	22,45	202,14	
ENTRE_#_E	2	370,37	9077,71	3847	44,73	202,93	
A_FALTA_DE	3	555,56	13616,56	5691	66,17	205,77	
O_MEIO_AMBIENTE	1	185,19	4538,85	1887	21,94	206,86	
DE_JUNHO_DE	1	185,19	4538,85	1867	21,71	209,07	
PARA_O_BRASIL	3	555,56	13616,56	5517	64,15	212,26	
AO_LONGO_DO	1	185,19	4538,85	1836	21,35	212,60	

A_VER_COM	1	185,19	4538,85	1798	20,91	217,10	
A_#_#	2	370,37	9077,71	3528	41,02	221,28	
CONSELHO_NACIONAL_DE	1	185,19	4538,85	1725	20,06	226,28	
DE_R_#	5	925,93	22694,26	8469	98,48	230,45	
E_#_#	2	370,37	9077,71	3317	38,57	235,36	
ACORDO_COM_A	3	555,56	13616,56	4656	54,14	251,51	
A_#_ANOS	2	370,37	9077,71	2341	27,22	333,48	
A_PARTIR_DE	7	1296,30	31771,97	8170	95,00	334,44	
PARTIR_DE_#	2	370,37	9077,71	2200	25,58	354,86	
O_BRASIL_TEM	2	370,37	9077,71	2185	25,41	357,29	
MEIO_AMBIENTE_E	3	555,56	13616,56	2722	31,65	430,21	
EM_#_#	5	925,93	22694,26	4507	52,41	433,04	
R_#_#	18	3333,33	81699,35	12506	145,42	561,82	

MDA *vs.* BP escrito
MD autêntico *vs.* BP escrito – Pacotes lexicais encontrados (NC= 20x por milhão)

Total de pacotes em comum = 55, Subuso = 38, Uso equivalente = 16, Sobreuso = 1

Trigrama	Freq MDA absoluta	Freq MDA normalizada	Freq MD total normalizada	Freq BP escrito absoluta	Freq BP escito normalizada	Razão	Classificação
RIO DE JANEIRO	2	370,37	49,02	228926	398,83	0,12	Subuso
DE SÃO PAULO	3	555,56	73,53	275399	479,79	0,15	
A PARTIR DO	1	185,19	24,51	51481	89,69	0,27	
EM QUE O	1	185,19	24,51	42558	74,14	0,33	
MAIS DO QUE	1	185,19	24,51	41533	72,36	0,34	
O QUE É	1	185,19	24,51	40915	71,28	0,34	
PARA A FOLHA	1	185,19	24,51	39971	69,64	0,35	

DE TODOS OS	1	185,19	24,51	39300	68,47	0,36	
DA UNIVERSIDADE DE	1	185,19	24,51	37936	66,09	0,37	
A FIM DE	1	185,19	24,51	30909	53,85	0,46	
DE ACORDO COM	5	925,93	122,55	154501	269,17	0,46	
SÃO PAULO E	1	185,19	24,51	30053	52,36	0,47	
O QUE SE	1	185,19	24,51	29955	52,19	0,47	
AO MESMO TEMPO	2	370,37	49,02	53572	93,33	0,53	
AO LONGO DO	1	185,19	24,51	25671	44,72	0,55	
DA DÉCADA DE	1	185,19	24,51	24944	43,46	0,56	
NA DÉCADA DE	1	185,19	24,51	22992	40,06	0,61	
O NÍVEL DE	1	185,19	24,51	22359	38,95	0,63	
O PERÍODO DE	1	185,19	24,51	22213	38,70	0,63	
A FORMAÇÃO DE	1	185,19	24,51	21897	38,15	0,64	
QUE PODE SER	1	185,19	24,51	21714	37,83	0,65	
SÃO PAULO O	1	185,19	24,51	21217	36,96	0,66	
O USO DE	2	370,37	49,02	42142	73,42	0,67	
A CAPACIDADE DE	1	185,19	24,51	20475	35,67	0,69	
DE NOVA YORK	1	185,19	24,51	19347	33,71	0,73	
QUE NÃO É	1	185,19	24,51	18633	32,46	0,76	
A PARTIR DE	7	1296,30	171,57	121717	212,05	0,81	
A REALIZAÇÃO DE	1	185,19	24,51	17041	29,69	0,83	
PARA QUE A	1	185,19	24,51	16836	29,33	0,84	
UMA FORMA DE	1	185,19	24,51	16517	28,78	0,85	
DURANTE O PERÍODO	1	185,19	24,51	16461	28,68	0,85	
MERCADO DE TRABALHO	1	185,19	24,51	16289	28,38	0,86	
QUE É O	1	185,19	24,51	16226	28,27	0,87	
NO MERCADO DE	1	185,19	24,51	16012	27,90	0,88	
CIDADE DE SÃO	1	185,19	24,51	15628	27,23	0,90	
QUE É A	1	185,19	24,51	15403	26,83	0,91	
POR CAUSA DA	1	185,19	24,51	14627	25,48	0,96	

NÃO SE PODE	1	185,19	24,51	14462	25,20	0,97	
DE JANEIRO DE	1	185,19	24,51	14088	24,54	1,00	Uso equivalente
ACORDO COM A	3	555,56	73,53	40617	70,76	1,04	
PARA A REALIZAÇÃO	1	185,19	24,51	13304	23,18	1,06	
AO LONGO DE	1	185,19	24,51	13290	23,15	1,06	
ACORDO COM AS	1	185,19	24,51	12865	22,41	1,09	
DA AMÉRICA LATINA	1	185,19	24,51	12567	21,89	1,12	
DE CADA UM	1	185,19	24,51	12380	21,57	1,14	
UM DOS MAIS	1	185,19	24,51	12231	21,31	1,15	
O QUE ESTÁ	1	185,19	24,51	12055	21,00	1,17	
DO MEIO AMBIENTE	1	185,19	24,51	11868	20,68	1,19	
UM PROGRAMA DE	1	185,19	24,51	11776	20,52	1,19	
DO CENTRO DE	1	185,19	24,51	11498	20,03	1,22	
PARA QUE O	2	370,37	49,02	21592	37,62	1,30	
UOL COM BR	2	370,37	49,02	20968	36,53	1,34	
A FALTA DE	3	555,56	73,53	29283	51,02	1,44	
QUE O BRASIL	2	370,37	49,02	14107	24,58	1,99	
PARA O BRASIL	3	555,56	73,53	15050	26,22	2,80	Sobreuso

MDNA *vs.* BP oral
MD não autêntico *vs.* BP oral – Pacotes lexicais encontrados (NC= 20x por milhão)

Total de pacotes em comum = 101, Subuso = 51, Uso equivalente = 32, Sobreuso = 18

Trigrama	Freq MDNA absoluta	Freq MDNA normalizada	Freq MD total normalizada	Freq BP oral absoluta	Freq BP oral normalizada	Razão	Classificação
O_QUE_TINHA	1	28,22	24,51	17891	208,03	0,12	Subuso
DIA_#_DE	1	28,22	24,51	14966	174,02	0,14	
DO_RIO_DE	1	28,22	24,51	13459	156,50	0,16	

GRANDE_DO_SUL	1	28,22	24,51	12546	145,88	0,17	
MAIS_DE_#	2	56,45	49,02	25063	291,43	0,17	
DE_TODOS_OS	1	28,22	24,51	12080	140,47	0,17	
DE_ACORDO_COM	1	28,22	24,51	11712	136,19	0,18	
DE_QUE_A	1	28,22	24,51	9959	115,80	0,21	
MAIS_DO_QUE	1	28,22	24,51	9818	114,16	0,21	
QUE_O_GOVERNO	2	56,45	49,02	17660	205,35	0,24	
DE_#_ANOS	1	28,22	24,51	8035	93,43	0,26	
RIO_GRANDE_DO	2	56,45	49,02	15188	176,60	0,28	
AO_MESMO_TEMPO	1	28,22	24,51	6855	79,71	0,31	
DE_TODAS_AS	1	28,22	24,51	5711	66,41	0,37	
DO_MEIO_AMBIENTE	1	28,22	24,51	5566	64,72	0,38	
QUE_O_PRESIDENTE	1	28,22	24,51	5503	63,99	0,38	
ESSE_TIPO_DE	1	28,22	24,51	4492	52,23	0,47	
OS_ESTADOS_UNIDOS	1	28,22	24,51	4313	50,15	0,49	
NA_CIDADE_DE	1	28,22	24,51	4151	48,27	0,51	
DO_DIA_#	1	28,22	24,51	3926	45,65	0,54	
EM_#_E	1	28,22	24,51	3798	44,16	0,55	
DOS_ESTADOS_UNIDOS	1	28,22	24,51	3700	43,02	0,57	
PARA_QUE_OS	1	28,22	24,51	3638	42,30	0,58	
O_PRESIDENTE_DO	1	28,22	24,51	3484	40,51	0,61	
POR_FALTA_DE	1	28,22	24,51	3170	36,86	0,66	
DE_DEZEMBRO_DE	1	28,22	24,51	3102	36,07	0,68	
A_FIM_DE	5	141,12	122,55	15103	175,62	0,70	
RIO_DE_JANEIRO	8	225,80	196,08	23074	268,30	0,73	
DE_R_#	3	84,67	73,53	8469	98,48	0,75	
QUE_O_SENHOR	1	28,22	24,51	2759	32,08	0,76	
DE_#_MINUTOS	1	28,22	24,51	2737	31,83	0,77	
DA_CIDADE_DE	1	28,22	24,51	2686	31,23	0,78	
GRANDE_DO_NORTE	1	28,22	24,51	2663	30,97	0,79	

NO_RIO_DE	2	56,45	49,02	5208	60,56	0,81	
POR_MEIO_DA	1	28,22	24,51	2554	29,70	0,83	
NO_ANO_PASSADO	2	56,45	49,02	5014	58,30	0,84	
QUE_TODOS_OS	1	28,22	24,51	2476	28,79	0,85	
ABRIL_DE_#	1	28,22	24,51	2407	27,99	0,88	
EM_#_DE	4	112,90	98,04	9581	111,41	0,88	
QUE_EM_#	1	28,22	24,51	2382	27,70	0,88	
A_OPORTUNIDADE_DE	2	56,45	49,02	4729	54,99	0,89	
NOVEMBRO_DE_#	1	28,22	24,51	2354	27,37	0,90	
COM_O_QUE	1	28,22	24,51	2350	27,33	0,90	
A_PROPOSTA_DE	1	28,22	24,51	2336	27,16	0,90	
CADA_VEZ_MAIS	4	112,90	98,04	9301	108,15	0,91	
SETEMBRO_DE_#	1	28,22	24,51	2308	26,84	0,91	
DO_IMPOSTO_DE	1	28,22	24,51	2276	26,47	0,93	
COM_#_#	1	28,22	24,51	2232	25,95	0,94	
O_AUMENTO_DA	1	28,22	24,51	2214	25,74	0,95	
SOBRE_O_ASSUNTO	1	28,22	24,51	2157	25,08	0,98	
A_MAIORIA_DOS	1	28,22	24,51	2142	24,91	0,98	
UM_GRUPO_DE	1	28,22	24,51	2091	24,31	1,01	Uso equivalente
NO_BRASIL_O	1	28,22	24,51	2067	24,03	1,02	
E_POR_ISSO	1	28,22	24,51	2061	23,97	1,02	
O_RIO_GRANDE	1	28,22	24,51	2009	23,36	1,05	
POR_TODOS_OS	1	28,22	24,51	2009	23,36	1,05	
ESTADOS_UNIDOS_E	1	28,22	24,51	2007	23,34	1,05	
EM_#_O	2	56,45	49,02	3991	46,41	1,06	
COM_TODOS_OS	1	28,22	24,51	1949	22,66	1,08	
DE_ABRIL_DE	1	28,22	24,51	1926	22,40	1,09	
O_DIA_#	1	28,22	24,51	1922	22,35	1,10	
O_QUE_O	2	56,45	49,02	3787	44,03	1,11	
DO_ANO_PASSADO	2	56,45	49,02	3776	43,91	1,12	

E_QUE_O	1	28,22	24,51	1887	21,94	1,12	
DE_NOVEMBRO_DE	1	28,22	24,51	1870	21,74	1,13	
EM_BUSCA_DE	1	28,22	24,51	1841	21,41	1,14	
AO_LONGO_DO	1	28,22	24,51	1836	21,35	1,15	
NO_FINAL_DO	1	28,22	24,51	1827	21,24	1,15	
A_VER_COM	1	28,22	24,51	1798	20,91	1,17	
DE_SETEMBRO_DE	1	28,22	24,51	1779	20,69	1,18	
PELO_MENOS_#	1	28,22	24,51	1760	20,47	1,20	
DIZER_QUE_A	1	28,22	24,51	1723	20,03	1,22	
EM_#_A	2	56,45	49,02	3207	37,29	1,31	
TUDO_O_QUE	2	56,45	49,02	3061	35,59	1,38	
SOBRE_A_MESA	2	56,45	49,02	3053	35,50	1,38	
EU_ACHO_QUE	5	141,12	122,55	6987	81,24	1,51	
DE_TRABALHO_E	2	56,45	49,02	2708	31,49	1,56	
POR_ISSO_QUE	2	56,45	49,02	2574	29,93	1,64	
DIA_A_DIA	2	56,45	49,02	2506	29,14	1,68	
A_TAXA_DE	2	56,45	49,02	2479	28,83	1,70	
TEM_QUE_SER	2	56,45	49,02	2347	27,29	1,80	
MERCADO_DE_TRABALHO	3	84,67	73,53	3408	39,63	1,86	
QUE_A_GENTE	6	169,35	147,06	6425	74,71	1,97	
POR_EXEMPLO_A	2	0,01	49,02	2099	24,41	2,01	Sobreuso
E_O_QUE	5	0,01	122,55	4985	57,97	2,11	
NA_SEMANA_PASSADA	4	0,01	98,04	3153	36,66	2,67	
IMPOSTO_DE_RENDA	5	0,01	122,55	3924	45,63	2,69	
O_QUE_A	3	0,01	73,53	2205	25,64	2,87	
AS_PESSOAS_QUE	3	0,01	73,53	2045	23,78	3,09	
O_RIO_DE	3	0,01	73,53	1982	23,05	3,19	
O_MEIO_AMBIENTE	3	0,01	73,53	1887	21,94	3,35	
R_#_#	20	0,06	490,20	12506	145,42	3,37	
O_QUE_EU	5	0,01	122,55	3010	35,00	3,50	

UM_POUCO_MAIS	5	0,01	122,55	1939	22,55	5,44	
MAIS_OU_MENOS	8	0,02	196,08	2965	34,48	5,69	
COMO_POR_EXEMPLO	6	0,02	147,06	1798	20,91	7,03	
QUE_AS_PESSOAS	8	0,02	196,08	1942	22,58	8,68	
O_QUE_FOI	9	0,03	220,59	2040	23,72	9,30	
FINAL_DE_SEMANA	8	0,02	196,08	1730	20,12	9,75	
O_QUE_ACONTECEU	10	0,03	245,10	1837	21,36	11,47	
TODOS_OS_DIAS	14	0,04	343,14	1751	20,36	16,85	

MDNA *vs.* BP escrito
MD não autêntico *vs.* BP escrito – Pacotes lexicais encontrados (NC= 20x por milhão)

Total de pacotes em comum = 68, Subuso = 36, Uso equivalente = 20, Sobreuso = 12

Trigrama	Freq MDNA absoluta	Freq MDNA normalizada	Freq MD total normalizada	Freq BP escrito absoluta	Freq BP escrito normalizada	Razão	Classificação
DE ACORDO COM	1	28,22	24,51	154501	269,17	0,09	Subuso
O NÚMERO DE	1	28,22	24,51	79875	139,16	0,18	
GRANDE DO SUL	1	28,22	24,51	56273	98,04	0,25	
AO MESMO TEMPO	1	28,22	24,51	53572	93,33	0,26	
DE QUE A	1	28,22	24,51	52215	90,97	0,27	
DO RIO DE	1	28,22	24,51	50032	87,16	0,28	
DE SÃO PAULO	6	169,35	147,06	275399	479,79	0,31	
UNIVERSIDADE DE SÃO	1	28,22	24,51	43815	76,33	0,32	
O USO DE	1	28,22	24,51	42142	73,42	0,33	
MAIS DO QUE	1	28,22	24,51	41533	72,36	0,34	
DE TODOS OS	1	28,22	24,51	39300	68,47	0,36	
A MAIORIA DOS	1	28,22	24,51	32043	55,82	0,44	
RIO GRANDE DO	2	56,45	49,02	62867	109,52	0,45	

SÃO PAULO E	1	28,22	24,51	30053	52,36	0,47	
POR MEIO DA	1	28,22	24,51	29219	50,90	0,48	
O PRESIDENTE DO	1	28,22	24,51	28837	50,24	0,49	
RIO DE JANEIRO	8	225,80	196,08	228926	398,83	0,49	
AO LONGO DO	1	28,22	24,51	25671	44,72	0,55	
DE TODAS AS	1	28,22	24,51	25519	44,46	0,55	
UM GRUPO DE	1	28,22	24,51	24659	42,96	0,57	
O AUMENTO DA	1	28,22	24,51	24302	42,34	0,58	
ESSE TIPO DE	1	28,22	24,51	20804	36,24	0,68	
NA CIDADE DE	1	28,22	24,51	20540	35,78	0,68	
NO FINAL DO	1	28,22	24,51	19148	33,36	0,73	
DOS ESTADOS UNIDOS	1	28,22	24,51	17850	31,10	0,79	
NO ANO PASSADO	2	56,45	49,02	33238	57,91	0,85	
UMA FORMA DE	1	28,22	24,51	16517	28,78	0,85	
A FORMA DE	1	28,22	24,51	16466	28,69	0,85	
A TAXA DE	2	56,45	49,02	31373	54,66	0,90	
SERVIÇOS DE SAÚDE	1	28,22	24,51	15604	27,18	0,90	
DA CIDADE DE	1	28,22	24,51	14845	25,86	0,95	
OU SEJA A	1	28,22	24,51	14674	25,56	0,96	
QUE O GOVERNO	2	56,45	49,02	29121	50,73	0,97	
DE DEZEMBRO DE	1	28,22	24,51	14538	25,33	0,97	
DE SAÚDE E	1	28,22	24,51	14525	25,30	0,97	
DO ANO PASSADO	2	56,45	49,02	28540	49,72	0,99	
NÃO É O	1	28,22	24,51	13881	24,18	1,01	Uso equivalente
E POR ISSO	1	28,22	24,51	13531	23,57	1,04	
EM BELO HORIZONTE	1	28,22	24,51	12609	21,97	1,12	
O QUE ESTÁ	1	28,22	24,51	12055	21,00	1,17	
FINAL DO ANO	1	28,22	24,51	11964	20,84	1,18	
NO RIO DE	2	56,45	49,02	23891	41,62	1,18	
DO MEIO AMBIENTE	1	28,22	24,51	11868	20,68	1,19	

DE NOVEMBRO DE	1	28,22	24,51	11837	20,62	1,19	
PARA QUE OS	1	28,22	24,51	11715	20,41	1,20	
FAZ COM QUE	1	28,22	24,51	11572	20,16	1,22	
NO BRASIL O	1	28,22	24,51	11528	20,08	1,22	
EM SÃO PAULO	9	254,02	220,59	99118	172,68	1,28	
O QUE NÃO	2	56,45	49,02	20410	35,56	1,38	
CADA VEZ MAIS	4	112,90	98,04	37861	65,96	1,49	
TUDO O QUE	2	56,45	49,02	16277	28,36	1,73	
QUE É O	2	56,45	49,02	16226	28,27	1,73	
CIDADE DE SÃO	2	56,45	49,02	15628	27,23	1,80	
O QUE O	2	56,45	49,02	15413	26,85	1,83	
POR EXEMPLO A	2	56,45	49,02	14918	25,99	1,89	
O TRABALHO DE	2	56,45	49,02	14660	25,54	1,92	
SÃO PAULO A	3	0,01	73,53	20048	34,93	2,11	Sobreuso
DISSE QUE A	3	0,01	73,53	19674	34,28	2,15	
DE TRABALHO E	2	0,01	49,02	12524	21,82	2,25	
QUE NÃO É	3	0,01	73,53	18633	32,46	2,27	
A FIM DE	5	0,01	122,55	30909	53,85	2,28	
NA SEMANA PASSADA	4	0,01	98,04	23056	40,17	2,44	
MERCADO DE TRABALHO	3	0,01	73,53	16289	28,38	2,59	
NO FINAL DE	4	0,01	98,04	12750	22,21	4,41	
COMO POR EXEMPLO	6	0,02	147,06	18326	31,93	4,61	
E O QUE	5	0,01	122,55	14652	25,53	4,80	
O QUE É	20	0,06	490,20	40915	71,28	6,88	
MAIS OU MENOS	8	0,02	196,08	16100	28,05	6,99	

Apêndice 5

Listas de Convergência texto a texto (por unidade)

MD vs. BP oral			
arquivo	**trigramas - texto**	**trigramas convergentes (BP oral)**	**% de convergência**
MDNA/UN1_LIA	16	6	37.500
MDNA/UN1_LIA_1	11	4	36.300
MDNA/UN1_LIA_2	9	3	33.300
MDNA/UN1_LIA_3	13	6	46.100
MDNA/UN1_LIB	36	5	13.800
MDNA/UN1_LIB_1	28	9	32.100
MDNA/UN1_LIB_2	48	15	31.200
MDNA/UN1_LIB_3	19	4	21.000
MDNA/UN1_LIC	31	5	16.100
MDNA/UN1_LIC_1	17	1	5.800
MDNA/UN1_LIABC	26	7	26.900
MDNA/UN1_LIABC_1	15	3	20.000
MDNA/UN1_LIABC_2	17	5	29.400
MDNA/UN1_LIABC_3	43	10	23.200
	U1	**Média**	**26.621**
MDNA/UN2_LIA	10	0	0
MDNA/UN2_LIA_1	25	5	20.000
MDNA/UN2_LIA_2	4	0	0
MDNA/UN2_LIA_4	10	1	10.000
MDNA/UN2_LIA_5	15	2	13.300
MDNA/UN2_LIA_7	17	5	29.400
MDNA/UN2_LIB	27	7	25.900
MDNA/UN2_LIB_1	10	1	10.000
MDNA/UN2_LIB_2	10	0	0
MDNA/UN2_LIB_3	13	8	61.500
MDNA/UN2_LIB_4	32	10	31.200
MDNA/UN2_LIB_5	21	8	38.000
MDNA/UN2_LIC	10	4	40.000
MDNA/UN2_LIC_1	17	1	5.800
MDNA/UN2_LIC_2	37	5	13.500
MDNA/UN2_LIABC	16	9	56.200
MDNA/UN2_LIABC_1	21	6	28.500
MDNA/UN2_LIABC_2	45	8	17.700
MDNA/UN2_LIABC_3	23	8	34.700
	U2	**Média**	**22.932**
MDNA/UN3_LIA	17	2	11.700
MDNA/UN3_LIA_1	34	0	0

MDNA/UN3_LIA_2	15	5	33.300
MDNA/UN3_LIA_3	26	9	34.600
MDNA/UN3_LIB	50	16	32.000
MDNA/UN3_LIB_1	38	7	18.400
MDNA/UN3_LIB_2	47	11	23.400
MDNA/UN3_LIC	51	14	27.400
MDNA/UN3_LIC_1	51	13	25.400
MDNA/UN3_LIC_2	50	16	32.000
MDA/UN3LIC_P	18	8	44.400
MDNA/UN3_LIABC	40	11	27.500
MDNA/UN3_LIABC_1	39	13	33.300
MDNA/UN3_LIABC_2	22	4	18.100
MDA/UN3LIABC_L	61	10	16.300
	U3	**Média**	**25.187**
MDNA/UN4_LIA	79	18	22.700
MDNA/UN4_LIA_1	11	1	9.000
MDNA/UN4_LIA_2	17	4	23.500
MDNA/UN4_LIA_3	18	4	22.200
MDNA/UN4_LIA_4	24	3	12.500
MDNA/UN4_LIA_5	39	17	43.500
MDNA/UN4_LIA_6	15	4	26.600
MDNA/UN4_LIA_7	105	25	23.800
MDNA/UN4_LIA_8	18	4	22.200
MDNA/UN4_LIB	119	28	23.500
MDNA/UN4_LIB_1	26	1	3.800
MDNA/UN4_LIB_2	84	8	9.500
MDNA/UN4_LIC	73	22	30.100
MDNA/UN4_LIC_1	39	4	10.200
MDNA/UN4_LIC_2	81	25	30.800
MDNA/UN4_LIC_3	86	24	27.900
MDNA/UN4_LIC_4	43	12	27.900
MDNA/UN4_LIABC	93	15	16.100
MDNA/UN4_LIABC_1	72	17	23.600
MDNA/UN4_LIABC_2	81	23	28.300
MDA/UN4LIABC_L	106	9	8.400
	U4	**Média**	**21.243**
MDNA/REV_1	176	57	32.300
MDNA/PR_1	29	8	27.500
	Rev e Pron 1	**Média**	**29.900**
MDNA/UN5_LIA	38	4	10.500
MDNA/UN5_LIA_1	33	2	6.000
MDNA/UN5_LIA_2	55	17	30.900
MDNA/UN5_LIB	17	1	5.800
MDNA/UN5_LIB_1	97	24	24.700
MDNA/UN5_LIB_2	12	3	25.000
MDNA/UN5_LIB_3	23	2	8.600
MDNA/UN5_LIB_4	33	5	15.100

MDNA/UN5_LIB_5	50	16	32.000
MDNA/UN5_LIC	22	5	22.700
MDNA/UN5_LIC_1	54	16	29.600
MDNA/UN5_LIC_2	31	11	35.400
MDNA/UN5_LIC_3	22	5	22.700
MDNA/UN5_LIC_4	54	16	29.600
MDNA/UN5_LIC_5	25	11	44.000
MDNA/UN5_LIC_6	54	13	24.000
MDNA/UN5_LIC_7	17	3	17.600
MDNA/UN5_LIC_8	43	13	30.200
MDNA/UN5_LIABC	52	7	13.400
MDNA/UN5_LIABC_1	51	5	9.800
MDNA/UN5_LIABC_2	31	5	16.100
MDNA/UN5_LIABC_3	27	6	22.200
MDNA/UN5_LIABC_4	92	28	30.400
	U5	**Média**	**22.013**
MDNA/UN6_LIA	79	25	31.600
MDNA/UN6_LIA_1	88	27	30.600
MDNA/UN6_LIA_2	64	18	28.100
MDNA/UN6_LIA_3	97	35	36.000
MDNA/UN6_LIB	72	30	41.600
MDNA/UN6_LIB_1	44	20	45.400
MDNA/UN6_LIB_2	27	5	18.500
MDNA/UN6_LIB_3	69	27	39.100
MDNA/UN6_LIC	40	5	12.500
MDNA/UN6_LIC_1	34	10	29.400
MDNA/UN6_LIC_2	39	9	23.000
MDNA/UN6_LIC_3	25	5	20.000
MDNA/UN6_LIC_4	71	9	12.600
MDNA/UN6_LIC_5	57	18	31.500
MDNA/UN6_LIABC	111	32	28.800
MDA/UN6LIABC_L	42	15	35.700
	U6	**Média**	**29.025**
MDNA/UN7_LIA	85	18	21.100
MDNA/UN7_LIA_1	13	2	15.300
MDNA/UN7_LIA_2	112	21	18.700
MDNA/UN7_LIA_3	9	0	0
MDNA/UN7_LIA_4	8	2	25.000
MDNA/UN7_LIA_5	7	0	0
MDNA/UN7_LIA_6	8	0	0
MDNA/UN7_LIA_7	79	18	22.700
MDNA/UN7_LIB	76	14	18.400
MDNA/UN7_LIB_1	38	8	21.000
MDNA/UN7_LIB_2	62	15	24.100
MDNA/UN7_LIB_3	68	30	44.100
MDNA/UN7_LIB_4	39	15	38.400
MDNA/UN7_LIB_5	15	3	20.000

MDNA/UN7_LIB_6	24	5	20.800
MDNA/UN7_LIB_7	25	16	64.000
MDNA/UN7_LIB_8	64	17	26.500
MDNA/UN7_LIB_9	34	4	11.700
MDNA/UN7_LIC	53	7	13.200
MDNA/UN7_LIABC	83	17	20.400
MDNA/UN7_LIABC_1	48	13	27.000
MDA/UN7LIABC_L	155	37	23.800
	U7	**Média**	**21.645**
MDNA/UN8_LIA	38	11	28.900
MDNA/UN8_LIA_1	15	1	6.600
MDNA/UN8_LIA_2	13	4	30.700
MDNA/UN8_LIA_3	52	15	28.800
MDNA/UN8_LIA_4	53	12	22.600
MDNA/UN8_LIA_5	30	10	33.300
MDNA/UN8_LIA_6	83	23	27.700
MDNA/UN8_LIB	37	11	29.700
MDNA/UN8_LIB_1	65	25	38.400
MDNA/UN8_LIB_2	52	16	30.700
MDNA/UN8_LIB_3	39	6	15.300
MDNA/UN8_LIB_4	173	53	30.600
MDNA/UN8_LIC	96	21	21.800
MDNA/UN8_LIABC	51	9	17.600
MDNA/UN8_LIABC_1	76	38	50.000
MDNA/UN8_LIABC_2	52	25	48.000
	U8	**Média**	**28.794**
MDNA/REV_2	46	8	17.300
MDNA/REV_2_1	24	4	16.600
MDNA/REV_2_2	17	4	23.500
MDNA/REV_2_3	3	0	0
MDNA/REV_2_4	32	10	31.200
MDNA/REV_2_5	17	6	35.200
MDNA/REV_2_6	14	5	35.700
MDNA/REV_2_7	10	2	20.000
MDNA/REV_2_8	17	3	17.600
MDNA/REV_2_9	32	7	21.800
MDNA/REV_2_10	13	5	38.400
MDNA/PR_2	4	1	25.000
	Rev e Pron 2	**Média**	**23.525**
MDNA/UN9_LIA	183	42	22.900
MDNA/UN9_LIA_1	63	10	15.800
MDNA/UN9_LIA_2	70	24	34.200
MDNA/UN9_LIB	38	11	28.900
MDNA/UN9_LIB_1	50	13	26.000
MDNA/UN9_LIB_2	48	16	33.300
MDNA/UN9_LIB_3	77	27	35.000
MDNA/UN9_LIB_4	69	19	27.500

MDNA/UN9_LIC	127	40	31.400
MDNA/UN9_LIC_1	90	28	31.100
MDNA/UN9_LIC_2	102	19	18.600
MDA/UN9_LIABC_L	208	69	33.100
MDNA/UN9_LIABC	54	9	16.600
MDNA/UN9_LIABC_1	81	14	17.200
MDNA/UN9_LIABC_2	32	5	15.600
	U9	**Média**	**25.813**
MDNA/UN10_LIA	30	6	20.000
MDNA/UN10_LIA_1	22	9	40.900
MDNA/UN10_LIA_2	26	2	7.600
MDNA/UN10_LIA_3	28	3	10.700
MDNA/UN10_LIA_4	23	1	4.300
MDNA/UN10_LIA_5	41	13	31.700
MDNA/UN10_LIA_6	15	3	20.000
MDNA/UN10_LIA_7	15	4	26.600
MDNA/UN10_LIA_8	10	4	40.000
MDNA/UN10_LIA_9	15	5	33.300
MDNA/UN10_LIA_10	13	0	0
MDNA/UN10_LIA_11	63	17	26.900
MDNA/UN10_LIA_12	113	25	22.100
MDNA/UN10_LIB	23	2	8.600
MDNA/UN10_LIB_1	63	18	28.500
MDNA/UN10_LIB_2	19	2	10.500
MDNA/UN10_LIB_3	66	12	18.100
MDNA/UN10_LIB_4	95	17	17.800
MDNA/UN10_LIB_5	90	17	18.800
MDNA/UN10_LIC	66	18	27.200
MDNA/UN10_LIC_1	109	27	24.700
MDNA/UN10_LIC_2	26	6	23.000
MDNA/UN10_LIC_3	53	11	20.700
MDNA/UN10_LIABC	95	25	26.300
MDNA/UN10_LIABC_1	77	14	18.100
MDNA/UN10_LIABC_2	60	14	23.300
MDA/UN10LIABC_L	167	44	26.300
	U10	**Média**	**21.333**
MDNA/UN11_LIA	57	23	40.300
MDNA/UN11_LIA_1	57	13	22.800
MDNA/UN11_LIA_2	78	31	39.700
MDNA/UN11_LIA_3	48	21	43.700
MDNA/UN11_LIB	20	3	15.000
MDNA/UN11_LIB_1	32	9	28.100
MDNA/UN11_LIB_2	38	11	28.900
MDNA/UN11_LIB_3	70	25	35.700
MDNA/UN11_LIB_4	150	68	45.300
MDNA/UN11_LIB_5	53	10	18.800
MDNA/UN11_LIC	87	42	48.200

MDNA/UN11_LIC_1	51	10	19.600
MDNA/UN11_LIC_2	168	29	17.200
MDNA/UN11_LIC_3	68	18	26.400
MDA/UN11LIC_V	150	36	24.000
MDA/UN11LIABC_L	150	38	25.300
MDNA/UN11_LIABC	77	21	27.200
MDNA/UN11_LIABC_1	87	22	25.200
MDNA/UN11_LIABC_2	104	29	27.800
	U11	**Média**	**29.432**
MDNA/UN12_LIA	37	14	37.800
MDNA/UN12_LIA_1	42	5	11.900
MDNA/UN12_LIA_2	39	9	23.000
MDNA/UN12_LIA_3	59	13	22.000
MDNA/UN12_LIA_4	73	17	23.200
MDNA/UN12_LIA_5	127	31	24.400
MDNA/UN12_LIA_6	48	18	37.500
MDNA/UN12_LIB	43	14	32.500
MDNA/UN12_LIB_1	18	7	38.800
MDNA/UN12_LIB_2	41	7	17.000
MDNA/UN12_LIB_3	58	30	51.700
MDNA/UN12_LIB_4	38	9	23.600
MDNA/UN12_LIB_5	101	44	43.500
MDNA/UN12_LIB_6	80	30	37.500
MDNA/UN12_LIB_7	144	56	38.800
MDNA/UN12_LIB_8	41	13	31.700
MDNA/UN12_LIC	37	10	27.000
MDNA/UN12_LIC_1	20	6	30.000
MDNA/UN12_LIC_2	69	28	40.500
MDNA/UN12_LIC_3	110	50	45.400
MDNA/UN12_LIC_4	176	49	27.800
MDNA/UN12_LIC_5	58	20	34.400
MDNA/UN12_LIABC	107	24	22.400
MDNA/UN12_LIABC_1	105	28	26.600
MDNA/UN12_LIABC_2	139	49	35.200
MDNA/UN12_LIABC_3	27	6	22.200
MDA/UN12LIABC_L	283	104	36.700
	U12	**Média**	**31.226**
MDNA/REV_3	55	14	25.400
MDNA/REV_3_1	14	4	28.500
MDNA/REV_3_2	23	7	30.400
MDNA/REV_3_3	38	10	26.300
MDNA/REV_3_4	21	5	23.800
MDNA/REV_3_5	35	15	42.800
MDNA/REV_3_6	12	4	33.300
MDNA/REV_3_7	31	7	22.500
MDNA/REV_3_8	14	0	0
MDNA/REV_3_9	22	9	40.900

MDNA/REV_3_10	15	4	26.600
MDNA/REV_3_11	22	5	22.700
MDNA/REV_3_12	27	7	25.900
MDNA/REV_3_13	10	3	30.000
MDNA/REV_3_14	19	3	15.700
MDNA/REV_3_15	117	29	24.700
MDNA/PR_3	49	16	32.600
	Rev e Pron 3	**Média**	**26.594**
MDNA/UN13_LIA	40	10	25.000
MDNA/UN13_LIA_1	16	4	25.000
MDNA/UN13_LIA_2	63	9	14.200
MDNA/UN13_LIA_3	55	18	32.700
MDNA/UN13_LIA_4	119	34	28.500
MDNA/UN13_LIA_5	109	23	21.100
MDNA/UN13_LIB	60	12	20.000
MDNA/UN13_LIB_1	77	18	23.300
MDNA/UN13_LIB_2	81	28	34.500
MDNA/UN13_LIB_3	215	69	32.000
MDNA/UN13_LIB_4	126	35	27.700
MDNA/UN13_LIC	107	46	42.900
MDNA/UN13_LIC_1	98	24	24.400
MDNA/UN13_LIC_2	80	25	31.200
MDNA/UN13_LIC_3	129	36	27.900
MDNA/UN13_LIC_4	69	36	52.100
MDNA/UN13_LIABC	103	29	28.100
MDNA/UN13_LIABC_1	99	31	31.300
MDNA/UN13_LIABC_2	77	23	29.800
MDNA/UN13_LIABC_3	12	4	33.300
MDA/UN13LIABC_L	226	65	28.700
	U13	**Média**	**29.224**
MDNA/UN14_LIA	52	14	26.900
MDNA/UN14_LIA_1	6	1	16.600
MDNA/UN14_LIA_2	77	23	29.800
MDNA/UN14_LIA_3	66	12	18.100
MDNA/UN14_LIA_4	7	6	85.700
MDNA/UN14_LIA_5	53	7	13.200
MDNA/UN14_LIA_6	52	17	32.600
MDNA/UN14_LIB	22	8	36.300
MDNA/UN14_LIB_1	33	2	6.000
MDNA/UN14_LIB_2	57	21	36.800
MDNA/UN14_LIB_3	34	6	17.600
MDNA/UN14_LIB_4	19	8	42.100
MDNA/UN14_LIB_5	76	25	32.800
MDNA/UN14_LIB_6	56	14	25.000
MDNA/UN14_LIB_7	10	4	40.000
MDNA/UN14_LIB_8	50	24	48.000
MDNA/UN14_LIC	18	5	27.700

MDNA/UN14_LIC_1	56	24	42.800
MDNA/UN14_LIC_2	5	2	40.000
MDNA/UN14_LIC_3	14	1	7.100
MDNA/UN14_LIC_4	32	7	21.800
MDNA/UN14_LIC_5	22	9	40.900
MDNA/UN14_LIC_6	87	31	35.600
MDNA/UN14_LIC_7	57	20	35.000
MDNA/UN14_LIABC	90	31	34.400
MDNA/UN14_LIABC_1	59	12	20.300
MDNA/UN14_LIABC_2	67	20	29.800
MDA/UN14LIABC_L	78	32	41.000
	U14	**Média**	**31.568**
MDNA/UN15_LIA	24	5	20.800
MDNA/UN15_LIA_1	36	8	22.200
MDNA/UN15_LIA_2	19	6	31.500
MDNA/UN15_LIA_3	43	16	37.200
MDNA/UN15_LIA_4	4	3	75.000
MDNA/UN15_LIA_5	20	9	45.000
MDNA/UN15_LIA_6	14	4	28.500
MDNA/UN15_LIA_7	27	7	25.900
MDNA/UN15_LIA_8	18	3	16.600
MDNA/UN15_LIA_9	32	12	37.500
MDNA/UN15_LIA_10	149	49	32.800
MDNA/UN15_LIB	52	20	38.400
MDNA/UN15_LIB_1	84	26	30.900
MDNA/UN15_LIB_2	86	33	38.300
MDNA/UN15_LIB_3	96	28	29.100
MDNA/UN15_LIB_4	36	10	27.700
MDNA/UN15_LIC	64	20	31.200
MDNA/UN15_LIC_1	47	19	40.400
MDNA/UN15_LIC_2	34	16	47.000
MDNA/UN15_LIC_3	22	4	18.100
MDNA/UN15_LIC_4	154	60	38.900
MDNA/UN15_LIC_5	53	17	32.000
MDNA/UN15_LIC_6	86	21	24.400
MDNA/UN15_LIC_7	105	43	40.900
MDNA/UN15_LIC_8	45	14	31.100
MDNA/UN15_LIABC	79	22	27.800
MDNA/UN15_LIABC_1	60	14	23.300
MDNA/UN15_LIABC_2	105	32	30.400
MDA/UN15LIABC_L	352	119	33.800
	U15	**Média**	**32.990**
MDNA/UN16_LIA	95	34	35.700
MDNA/UN16_LIA_1	77	17	22.000
MDNA/UN16_LIA_2	99	24	24.200
MDNA/UN16_LIB	18	1	5.500
MDNA/UN16_LIB_1	30	10	33.300

MDNA/UN16_LIB_2	46	24	52.100
MDNA/UN16_LIB_3	109	31	28.400
MDNA/UN16_LIB_4	101	12	11.800
MDA/UN16LIB_V	57	12	21.000
MDNA/UN16_LIC	35	7	20.000
MDNA/UN16_LIC_1	10	2	20.000
MDNA/UN16_LIC_2	47	16	34.000
MDNA/UN16_LIC_3	98	28	28.500
MDNA/UN16_LIC_4	36	12	33.300
MDNA/UN16_LIC_5	8	3	37.500
MDNA/UN16_LIC_6	57	8	14.000
MDNA/UN16_LIC_7	38	11	28.900
MDNA/UN16_LIABC	279	81	29.000
MDA/UN16LIABC_L	332	114	34.300
	U16	**Média**	**27.026**
MDNA/REV_4	25	6	24.000
MDNA/REV_4_1	117	29	24.700
MDNA/REV_4_2	64	30	46.800
MDNA/REV_4_3	64	30	46.800
MDNA/REV_4_4	45	13	28.800
MDNA/REV_4_5	16	12	75.000
MDNA/REV_4_6	23	8	34.700
MDNA/REV_4_7	24	6	25.000
MDNA/REV_4_8	52	16	30.700
MDNA/REV_4_9	35	15	42.800
MDNA/REV_4_10	36	14	38.800
MDNA/REV_4_11	10	2	20.000
MDNA/REV_4_12	41	15	36.500
MDNA/REV_4_13	8	5	62.500
MDNA/PR_4	91	34	37.300
	Rev e Pron 4	**Média**	**38.293**
MDNA/UN17_LIA	55	26	47.200
MDNA/UN17_LIA_1	25	8	32.000
MDNA/UN17_LIA_2	75	25	33.300
MDNA/UN17_LIA_3	112	11	9.800
MDNA/UN17_LIA_4	59	24	40.600
MDNA/UN17_LIA_5	14	2	14.200
MDNA/UN17_LIA_6	12	5	41.600
MDNA/UN17_LIB	17	4	23.500
MDNA/UN17_LIB_1	12	1	8.300
MDNA/UN17_LIB_2	25	4	16.000
MDNA/UN17_LIB_3	31	8	25.800
MDNA/UN17_LIB_4	91	28	30.700
MDNA/UN17_LIB_5	31	11	35.400
MDNA/UN17_LIB_6	27	6	22.200
MDNA/UN17_LIB_7	16	2	12.500
MDNA/UN17_LIB_8	92	20	21.700

MDNA/UN17_LIB_9	47	8	17.000
MDA/UN17LIB_V	186	43	23.100
MDNA/UN17_LIC	66	12	18.100
MDNA/UN17_LIC_1	43	23	53.400
MDNA/UN17_LIC_2	65	16	24.600
MDNA/UN17_LIC_3	119	29	24.300
MDNA/UN17_LIC_4	53	13	24.500
MDNA/UN17_LIABC	117	23	19.600
MDNA/UN17_LIABC_1	108	19	17.500
MDNA/UN17_LIABC_2	78	19	24.300
MDNA/UN17_LIABC_3	97	45	46.300
MDA/UN17LIABC_L	192	58	30.200
	U17	**Média**	**26.346**
MDNA/UN18_LIA	60	26	43.300
MDNA/UN18_LIA_1	66	15	22.700
MDNA/UN18_LIA_2	116	39	33.600
MDNA/UN18_LIA_3	58	27	46.500
MDNA/UN18_LIB	20	7	35.000
MDNA/UN18_LIB_1	20	11	55.000
MDNA/UN18_LIB_2	11	0	0
MDNA/UN18_LIB_3	27	17	62.900
MDNA/UN18_LIB_4	10	5	50.000
MDNA/UN18_LIB_5	148	41	27.700
MDNA/UN18_LIB_6	81	30	37.000
MDNA/UN18_LIC	19	9	47.300
MDNA/UN18_LIC_1	15	7	46.600
MDNA/UN18_LIC_2	15	8	53.300
MDNA/UN18_LIC_3	19	6	31.500
MDNA/UN18_LIC_4	65	26	40.000
MDNA/UN18_LIC_5	26	6	23.000
MDNA/UN18_LIC_6	39	16	41.000
MDNA/UN18_LIC_7	22	6	27.200
MDNA/UN18_LIC_8	16	3	18.700
MDNA/UN18_LIC_9	10	1	10.000
MDNA/UN18_LIC_10	71	31	43.600
MDNA/UN18_LIC_11	157	63	40.100
MDNA/UN18_LIC_12	76	29	38.100
MDNA/UN18_LIABC	173	44	25.400
MDNA/UN18_LIABC_1	125	40	32.000
MDNA/UN18_LIABC_2	101	38	37.600
MDA/UN18LIABC_L	275	104	37.800
	U18	**Média**	**35.961**
MDNA/UN19_LIA	33	16	48.400
MDNA/UN19_LIA_1	19	5	26.300
MDNA/UN19_LIA_2	25	11	44.000
MDNA/UN19_LIA_3	84	30	35.700
MDNA/UN19_LIA_4	100	26	26.000

MDNA/UN19_LIA_5	65	20	30.700
MDNA/UN19_LIB	88	34	38.600
MDNA/UN19_LIB_1	62	18	29.000
MDNA/UN19_LIB_2	69	30	43.400
MDNA/UN19_LIB_3	176	42	23.800
MDNA/UN19_LIB_4	72	15	20.800
MDNA/UN19_LIC	54	15	27.700
MDNA/UN19_LIC_1	55	24	43.600
MDNA/UN19_LIC_2	85	19	22.300
MDNA/UN19_LIC_3	74	31	41.800
MDNA/UN19_LIC_4	72	18	25.000
MDNA/UN19_LIC_5	70	20	28.500
MDNA/UN19_LIABC	196	57	29.000
MDNA/UN19_LIABC_1	172	34	19.700
MDNA/UN19_LIABC_2	191	49	25.600
MDNA/UN19_LIABC_3	99	26	26.200
MDA/UN19LIABC_L	246	82	33.300
	U19	**Média**	**31.336**
MDNA/UN20_LIA	5	3	60.000
MDNA/UN20_LIA_1	118	35	29.600
MDNA/UN20_LIA_2	85	42	49.400
MDNA/UN20_LIA_3	145	39	26.800
MDNA/UN20_LIA_4	53	26	49.000
MDA/UN20LIB	28	13	46.400
MDNA/UN20_LIB	137	42	30.600
MDNA/UN20_LIB_1	70	17	24.200
MDNA/UN20_LIB_2	65	11	16.900
MDNA/UN20_LIB_3	202	69	34.100
MDNA/UN20_LIB_4	27	10	37.000
MDNA/UN20_LIB_5	11	4	36.300
MDNA/UN20_LIB_6	20	9	45.000
MDNA/UN20_LIB_7	41	9	21.900
MDNA/UN20_LIB_8	8	1	12.500
MDNA/UN20_LIB_9	21	9	42.800
MDNA/UN20_LIC_9	321	113	35.200
MDNA/UN20_LIABC	338	118	34.900
MDA/UN20LIABC_L	140	47	33.500
	U20	**Média**	**35.058**
MDNA/REV_5	38	10	26.300
MDNA/REV_5_1	15	6	40.000
MDNA/REV_5_2	14	3	21.400
MDNA/REV_5_3	9	2	22.200
MDNA/REV_5_4	27	14	51.800
MDNA/REV_5_5	14	6	42.800
MDNA/REV_5_6	115	32	27.800
MDNA/REV_5_7	20	8	40.000
MDNA/REV_5_8	33	22	66.600

MDNA/REV_5_9	20	8	40.000
MDNA/REV_5_10	26	8	30.700
MDNA/REV_5_11	23	11	47.800
MDNA/REV_5_12	38	11	28.900
MDNA/REV_5_13	36	10	27.700
MDNA/PR_5	104	20	19.200
	Rev e Pron 5	**Média**	**35.547**

MD vs. BP escrito			
arquivo	**trigramas - texto**	**trigramas convergentes (BP escrito)**	**% de convergência**
MDNA/UN1_LIA	16	3	18.700
MDNA/UN1_LIA_1	11	4	36.300
MDNA/UN1_LIA_2	9	3	33.300
MDNA/UN1_LIA_3	13	5	38.400
MDNA/UN1_LIB	36	5	13.800
MDNA/UN1_LIB_1	28	10	35.700
MDNA/UN1_LIB_2	48	11	22.900
MDNA/UN1_LIB_3	19	2	10.500
MDNA/UN1_LIC	31	11	35.400
MDNA/UN1_LIC_1	17	3	17.600
MDNA/UN1_LIABC	26	9	34.600
MDNA/UN1_LIABC_1	15	6	40.000
MDNA/UN1_LIABC_2	17	6	35.200
MDNA/UN1_LIABC_3	43	13	30.200
	U1	**Média**	**28.757**
MDNA/UN2_LIA	10	7	70.000
MDNA/UN2_LIA_1	25	7	28.000
MDNA/UN2_LIA_2	4	2	50.000
MDNA/UN2_LIA_4	10	5	50.000
MDNA/UN2_LIA_5	15	7	46.600
MDNA/UN2_LIA_7	17	8	47.000
MDNA/UN2_LIB	27	3	11.100
MDNA/UN2_LIB_1	10	3	30.000
MDNA/UN2_LIB_2	10	1	10.000
MDNA/UN2_LIB_3	13	3	23.000
MDNA/UN2_LIB_4	32	0	0
MDNA/UN2_LIB_5	21	7	33.300
MDNA/UN2_LIC	10	2	20.000
MDNA/UN2_LIC_1	17	8	47.000
MDNA/UN2_LIC_2	37	17	45.900
MDNA/UN2_LIABC	16	2	12.500
MDNA/UN2_LIABC_1	21	2	9.500
MDNA/UN2_LIABC_2	45	14	31.100

MDNA/UN2_LIABC_3	23	11	47.800
	U2	**Média**	**32.253**
MDNA/UN3_LIA	17	7	41.100
MDNA/UN3_LIA_1	35	18	51.400
MDNA/UN3_LIA_2	15	5	33.300
MDNA/UN3_LIA_3	26	8	30.700
MDNA/UN3_LIB	50	4	8.000
MDNA/UN3_LIB_1	38	12	31.500
MDNA/UN3_LIB_2	47	6	12.700
MDNA/UN3_LIC	51	10	19.600
MDNA/UN3_LIC_1	51	17	33.300
MDNA/UN3_LIC_2	50	4	8.000
MDA/UN3LIC_P	18	5	27.700
MDNA/UN3_LIABC	40	7	17.500
MDNA/UN3_LIABC_1	39	10	25.600
MDNA/UN3_LIABC_2	22	9	40.900
MDA/UN3LIABC_L	61	27	44.200
	U3	**Média**	**28.367**
MDNA/UN4_LIA	79	13	16.400
MDNA/UN4_LIA_1	11	5	45.400
MDNA/UN4_LIA_2	17	3	17.600
MDNA/UN4_LIA_3	18	3	16.600
MDNA/UN4_LIA_4	24	4	16.600
MDNA/UN4_LIA_5	39	11	28.200
MDNA/UN4_LIA_6	15	5	33.300
MDNA/UN4_LIA_7	107	38	35.500
MDNA/UN4_LIA_8	18	3	16.600
MDNA/UN4_LIB	120	33	27.500
MDNA/UN4_LIB_1	26	12	46.100
MDNA/UN4_LIB_2	84	15	17.800
MDNA/UN4_LIC	73	16	21.900
MDNA/UN4_LIC_1	39	7	17.900
MDNA/UN4_LIC_2	81	7	8.600
MDNA/UN4_LIC_3	86	26	30.200
MDNA/UN4_LIC_4	43	9	20.900
MDNA/UN4_LIABC	93	27	29.000
MDNA/UN4_LIABC_1	72	17	23.600
MDNA/UN4_LIABC_2	81	13	16.000
MDA/UN4LIABC_L	106	21	19.800
	U4	**Média**	**24.071**
MDNA/PR_1	29	5	17.200
MDNA/REV_1	178	41	23.000
	Rev e Pron 1	**Média**	**20.100**
MDNA/UN5_LIA	38	17	44.700
MDNA/UN5_LIA_1	33	18	54.500
MDNA/UN5_LIA_2	55	12	21.800
MDNA/UN5_LIB	17	5	29.400

MDNA/UN5_LIB_1	97	13	13.400
MDNA/UN5_LIB_2	12	5	41.600
MDNA/UN5_LIB_3	24	9	37.500
MDNA/UN5_LIB_4	34	10	29.400
MDNA/UN5_LIB_5	50	15	30.000
MDNA/UN5_LIC	22	6	27.200
MDNA/UN5_LIC_1	54	15	27.700
MDNA/UN5_LIC_2	31	4	12.900
MDNA/UN5_LIC_3	22	6	27.200
MDNA/UN5_LIC_4	54	15	27.700
MDNA/UN5_LIC_5	25	3	12.000
MDNA/UN5_LIC_6	54	18	33.300
MDNA/UN5_LIC_7	17	9	52.900
MDNA/UN5_LIC_8	43	10	23.200
MDNA/UN5_LIABC	52	10	19.200
MDNA/UN5_LIABC_1	51	20	39.200
MDNA/UN5_LIABC_2	31	9	29.000
MDNA/UN5_LIABC_3	27	9	33.300
MDNA/UN5_LIABC_4	92	23	25.000
	U5	**Média**	**30.091**
MDNA/UN6_LIA	79	14	17.700
MDNA/UN6_LIA_1	88	13	14.700
MDNA/UN6_LIA_2	64	18	28.100
MDNA/UN6_LIA_3	99	29	29.200
MDNA/UN6_LIB	74	22	29.700
MDNA/UN6_LIB_1	44	7	15.900
MDNA/UN6_LIB_2	28	12	42.800
MDNA/UN6_LIB_3	70	12	17.100
MDNA/UN6_LIC	40	14	35.000
MDNA/UN6_LIC_1	34	9	26.400
MDNA/UN6_LIC_2	39	19	48.700
MDNA/UN6_LIC_3	25	6	24.000
MDNA/UN6_LIC_4	71	34	47.800
MDNA/UN6_LIC_5	57	20	35.000
MDNA/UN6_LIABC	113	47	41.500
MDA/UN6LIABC_L	42	8	19.000
	U6	**Média**	**29.538**
MDNA/UN7_LIA	85	18	21.100
MDNA/UN7_LIA_1	14	6	42.800
MDNA/UN7_LIA_2	113	37	32.700
MDNA/UN7_LIA_3	9	3	33.300
MDNA/UN7_LIA_4	8	2	25.000
MDNA/UN7_LIA_5	7	2	28.500
MDNA/UN7_LIA_6	8	5	62.500
MDNA/UN7_LIA_7	79	33	41.700
MDNA/UN7_LIB	76	23	30.200
MDNA/UN7_LIB_1	38	8	21.000

MDNA/UN7_LIB_2	62	16	25.800
MDNA/UN7_LIB_3	69	21	30.400
MDNA/UN7_LIB_4	39	9	23.000
MDNA/UN7_LIB_5	15	1	6.600
MDNA/UN7_LIB_6	24	10	41.600
MDNA/UN7_LIB_7	25	3	12.000
MDNA/UN7_LIB_8	64	15	23.400
MDNA/UN7_LIB_9	34	11	32.300
MDNA/UN7_LIC	53	18	33.900
MDNA/UN7_LIABC	83	16	19.200
MDNA/UN7_LIABC_1	48	12	25.000
MDA/UN7LIABC_L	155	51	32.900
	U7	**Média**	**29.314**
MDNA/UN8_LIA	38	12	31.500
MDNA/UN8_LIA_1	15	5	33.300
MDNA/UN8_LIA_2	13	5	38.400
MDNA/UN8_LIA_3	52	8	15.300
MDNA/UN8_LIA_4	53	11	20.700
MDNA/UN8_LIA_5	30	10	33.300
MDNA/UN8_LIA_6	83	17	20.400
MDNA/UN8_LIB	37	2	5.400
MDNA/UN8_LIB_1	65	10	15.300
MDNA/UN8_LIB_2	52	4	7.600
MDNA/UN8_LIB_3	39	17	43.500
MDNA/UN8_LIB_4	174	37	21.200
MDNA/UN8_LIC	96	24	25.000
MDNA/UN8_LIABC	52	23	44.200
MDNA/UN8_LIABC_1	76	18	23.600
MDNA/UN8_LIABC_2	52	8	15.300
	U8	**Média**	**24.625**
MDNA/PR_2	46	8	17.300
MDNA/REV_2	24	4	16.600
MDNA/REV_2_1	17	7	41.100
MDNA/REV_2_10	4	2	50.000
MDNA/REV_2_2	3	1	33.300
MDNA/REV_2_3	33	15	45.400
MDNA/REV_2_4	17	0	0
MDNA/REV_2_5	14	5	35.700
MDNA/REV_2_6	10	2	20.000
MDNA/REV_2_7	17	3	17.600
MDNA/REV_2_8	32	6	18.700
MDNA/REV_2_9	13	3	23.000
	Rev e Pron 2	**Média**	**26.558**
MDNA/UN9_LIA	185	26	14.000
MDNA/UN9_LIA_1	63	21	33.300
MDNA/UN9_LIA_2	71	13	18.300
MDNA/UN9_LIB	38	16	42.100

MDNA/UN9_LIB_1	50	7	14.000
MDNA/UN9_LIB_2	48	4	8.300
MDNA/UN9_LIB_3	77	14	18.100
MDNA/UN9_LIB_4	69	15	21.700
MDNA/UN9_LIC	129	22	17.000
MDNA/UN9_LIC_1	90	9	10.000
MDNA/UN9_LIC_2	102	22	21.500
MDNA/UN9_LIABC	54	9	16.600
MDNA/UN9_LIABC_1	81	20	24.600
MDNA/UN9_LIABC_2	32	11	34.300
MDA/UN9_LIABC_L	209	41	19.600
	U9	**Média**	**20.893**
MDNA/UN10_LIA	30	1	3.300
MDNA/UN10_LIA_1	22	2	9.000
MDNA/UN10_LIA_10	13	7	53.800
MDNA/UN10_LIA_11	64	22	34.300
MDNA/UN10_LIA_12	113	42	37.100
MDNA/UN10_LIA_2	26	3	11.500
MDNA/UN10_LIA_3	28	5	17.800
MDNA/UN10_LIA_4	23	6	26.000
MDNA/UN10_LIA_5	41	13	31.700
MDNA/UN10_LIA_6	15	0	0
MDNA/UN10_LIA_7	15	8	53.300
MDNA/UN10_LIA_8	10	3	30.000
MDNA/UN10_LIA_9	15	6	40.000
MDNA/UN10_LIB	23	8	34.700
MDNA/UN10_LIB_1	63	7	11.100
MDNA/UN10_LIB_2	19	10	52.600
MDNA/UN10_LIB_3	66	10	15.100
MDNA/UN10_LIB_4	97	34	35.000
MDNA/UN10_LIB_5	90	28	31.100
MDNA/UN10_LIC	66	18	27.200
MDNA/UN10_LIC_1	109	36	33.000
MDNA/UN10_LIC_2	28	9	32.100
MDNA/UN10_LIC_3	53	10	18.800
MDNA/UN10_LIABC	95	26	27.300
MDNA/UN10_LIABC_1	77	26	33.700
MDNA/UN10_LIABC_2	60	18	30.000
MDA/UN10LIABC_L	168	32	19.000
	U10	**Média**	**27.722**
MDNA/UN11_LIA	57	9	15.700
MDNA/UN11_LIA_1	57	18	31.500
MDNA/UN11_LIA_2	78	29	37.100
MDNA/UN11_LIA_3	48	15	31.200
MDNA/UN11_LIB	20	6	30.000
MDNA/UN11_LIB_1	32	11	34.300
MDNA/UN11_LIB_2	38	10	26.300

MDNA/UN11_LIB_3	70	12	17.100
MDNA/UN11_LIB_4	152	20	13.100
MDNA/UN11_LIB_5	53	13	24.500
MDNA/UN11_LIC	87	10	11.400
MDNA/UN11_LIC_1	52	18	34.600
MDNA/UN11_LIC_2	168	44	26.100
MDNA/UN11_LIC_3	68	12	17.600
MDA/UN11LIC_V	150	28	18.600
MDNA/UN11_LIABC	77	18	23.300
MDNA/UN11_LIABC_1	87	22	25.200
MDNA/UN11_LIABC_2	104	13	12.500
MDA/UN11LIABC_L	151	34	22.500
	U11	**Média**	**23.821**
MDNA/UN12_LIA	37	3	8.100
MDNA/UN12_LIA_1	42	7	16.600
MDNA/UN12_LIA_2	39	6	15.300
MDNA/UN12_LIA_3	59	12	20.300
MDNA/UN12_LIA_4	73	14	19.100
MDNA/UN12_LIA_5	128	32	25.000
MDNA/UN12_LIA_6	48	8	16.600
MDNA/UN12_LIB	43	6	13.900
MDNA/UN12_LIB_1	18	5	27.700
MDNA/UN12_LIB_2	41	12	29.200
MDNA/UN12_LIB_3	58	13	22.400
MDNA/UN12_LIB_4	38	13	34.200
MDNA/UN12_LIB_5	101	21	20.700
MDNA/UN12_LIB_6	80	11	13.700
MDNA/UN12_LIB_7	144	28	19.400
MDNA/UN12_LIB_8	41	7	17.000
MDNA/UN12_LIC	38	9	23.600
MDNA/UN12_LIC_1	20	4	20.000
MDNA/UN12_LIC_2	69	22	31.800
MDNA/UN12_LIC_3	110	27	24.500
MDNA/UN12_LIC_4	176	53	30.100
MDNA/UN12_LIC_5	58	11	18.900
MDNA/UN12_LIABC	107	26	24.200
MDNA/UN12_LIABC_1	105	23	21.900
MDNA/UN12_LIABC_2	140	25	17.800
MDNA/UN12_LIABC_3	27	9	33.300
MDA/UN12LIABC_L	285	81	28.400
	U12	**Média**	**21.989**
MDNA/PR_3	49	13	26.500
MDNA/REV_3	55	10	18.100
MDNA/REV_3_1	14	6	42.800
MDNA/REV_3_10	15	6	40.000
MDNA/REV_3_11	22	9	40.900
MDNA/REV_3_12	28	13	46.400

MDNA/REV_3_13	10	2	20.000
MDNA/REV_3_14	19	2	10.500
MDNA/REV_3_15	118	33	27.900
MDNA/REV_3_2	23	12	52.100
MDNA/REV_3_3	38	7	18.400
MDNA/REV_3_4	21	9	42.800
MDNA/REV_3_5	35	8	22.800
MDNA/REV_3_6	12	1	8.300
MDNA/REV_3_7	31	8	25.800
MDNA/REV_3_8	14	6	42.800
MDNA/REV_3_9	22	5	22.700
	Rev e Pron 3	**Média**	**29.929**
MDNA/UN13_LIA	40	3	7.500
MDNA/UN13_LIA_1	16	0	0
MDNA/UN13_LIA_2	63	22	34.900
MDNA/UN13_LIA_3	55	8	14.500
MDNA/UN13_LIA_4	119	27	22.600
MDNA/UN13_LIA_5	109	22	20.100
MDNA/UN13_LIB	60	11	18.300
MDNA/UN13_LIB_1	77	18	23.300
MDNA/UN13_LIB_2	81	9	11.100
MDNA/UN13_LIB_3	217	61	28.100
MDNA/UN13_LIB_4	126	29	23.000
MDNA/UN13_LIC	108	25	23.100
MDNA/UN13_LIC_1	98	27	27.500
MDNA/UN13_LIC_2	80	15	18.700
MDNA/UN13_LIC_3	129	37	28.600
MDNA/UN13_LIC_4	69	8	11.500
MDNA/UN13_LIABC	104	22	21.100
MDNA/UN13_LIABC_1	99	14	14.100
MDNA/UN13_LIABC_2	77	19	24.600
MDNA/UN13_LIABC_3	12	1	8.300
MDA/UN13LIABC_L	227	34	14.900
	U13	**Média**	**18.848**
MDNA/UN14_LIA	52	7	13.400
MDNA/UN14_LIA_1	6	3	50.000
MDNA/UN14_LIA_2	78	15	19.200
MDNA/UN14_LIA_3	66	10	15.100
MDNA/UN14_LIA_4	7	0	0
MDNA/UN14_LIA_5	53	15	28.300
MDNA/UN14_LIA_6	52	9	17.300
MDNA/UN14_LIB	22	5	22.700
MDNA/UN14_LIB_1	33	9	27.200
MDNA/UN14_LIB_2	57	6	10.500
MDNA/UN14_LIB_3	34	10	29.400
MDNA/UN14_LIB_4	21	10	47.600
MDNA/UN14_LIB_5	76	16	21.000

MDNA/UN14_LIB_6	57	22	38.500
MDNA/UN14_LIB_7	10	1	10.000
MDNA/UN14_LIB_8	50	8	16.000
MDNA/UN14_LIC	18	3	16.600
MDNA/UN14_LIC_1	56	15	26.700
MDNA/UN14_LIC_2	5	1	20.000
MDNA/UN14_LIC_3	14	3	21.400
MDNA/UN14_LIC_4	32	15	46.800
MDNA/UN14_LIC_5	22	6	27.200
MDNA/UN14_LIC_6	87	20	22.900
MDNA/UN14_LIC_7	58	14	24.100
MDNA/UN14_LIABC	90	11	12.200
MDNA/UN14_LIABC_1	60	21	35.000
MDNA/UN14_LIABC_2	67	14	20.800
MDA/UN14LIABC_L	78	11	14.100
	U14	**Média**	**23.357**
MDNA/UN15_LIA	24	4	16.600
MDNA/UN15_LIA_1	36	11	30.500
MDNA/UN15_LIA_10	150	34	22.600
MDNA/UN15_LIA_2	19	1	5.200
MDNA/UN15_LIA_3	43	7	16.200
MDNA/UN15_LIA_4	4	1	25.000
MDNA/UN15_LIA_5	20	1	5.000
MDNA/UN15_LIA_6	14	2	14.200
MDNA/UN15_LIA_7	27	4	14.800
MDNA/UN15_LIA_8	18	4	22.200
MDNA/UN15_LIA_9	32	5	15.600
MDNA/UN15_LIB	52	8	15.300
MDNA/UN15_LIB_1	84	14	16.600
MDNA/UN15_LIB_2	86	12	13.900
MDNA/UN15_LIB_3	97	20	20.600
MDNA/UN15_LIB_4	36	6	16.600
MDNA/UN15_LIC	64	13	20.300
MDNA/UN15_LIC_1	47	17	36.100
MDNA/UN15_LIC_2	34	8	23.500
MDNA/UN15_LIC_3	22	6	27.200
MDNA/UN15_LIC_4	155	30	19.300
MDNA/UN15_LIC_5	53	13	24.500
MDNA/UN15_LIC_6	86	21	24.400
MDNA/UN15_LIC_7	105	22	20.900
MDNA/UN15_LIC_8	45	6	13.300
MDNA/UN15_LIABC	79	27	34.100
MDNA/UN15_LIABC_1	60	17	28.300
MDNA/UN15_LIABC_2	105	19	18.000
MDA/UN15LIABC_L	353	87	24.600
	U15	**Média**	**20.186**
MDNA/UN16_LIA	95	26	27.300

MDNA/UN16_LIA_1	79	17	21.500
MDNA/UN16_LIA_2	101	23	22.700
MDNA/UN16_LIB	18	5	27.700
MDNA/UN16_LIB_1	30	13	43.300
MDNA/UN16_LIB_2	46	2	4.300
MDNA/UN16_LIB_3	111	28	25.200
MDNA/UN16_LIB_4	101	23	22.700
MDA/UN16LIB_V	57	11	19.200
MDNA/UN16_LIC	35	9	25.700
MDNA/UN16_LIC_1	10	3	30.000
MDNA/UN16_LIC_2	48	14	29.100
MDNA/UN16_LIC_3	98	23	23.400
MDNA/UN16_LIC_4	36	7	19.400
MDNA/UN16_LIC_5	8	0	0
MDNA/UN16_LIC_6	57	10	17.500
MDNA/UN16_LIC_7	38	3	7.800
MDNA/UN16_LIABC	280	62	22.100
MDA/UN16LIABC_L	333	87	26.100
	U16	**Média**	**21.842**
MDNA/PR_4	92	19	20.600
MDNA/REV_4	25	6	24.000
MDNA/REV_4_1	118	33	27.900
MDNA/REV_4_10	36	8	22.200
MDNA/REV_4_11	10	4	40.000
MDNA/REV_4_12	41	15	36.500
MDNA/REV_4_13	8	1	12.500
MDNA/REV_4_2	64	9	14.000
MDNA/REV_4_3	64	9	14.000
MDNA/REV_4_4	45	11	24.400
MDNA/REV_4_5	16	1	6.200
MDNA/REV_4_6	23	2	8.600
MDNA/REV_4_7	24	2	8.300
MDNA/REV_4_8	52	10	19.200
MDNA/REV_4_9	35	7	20.000
	Rev e Pron 4	**Média**	**19.893**
MDNA/UN17_LIA	55	12	21.800
MDNA/UN17_LIA_1	25	6	24.000
MDNA/UN17_LIA_2	76	16	21.000
MDNA/UN17_LIA_3	112	41	36.600
MDNA/UN17_LIA_4	59	13	22.000
MDNA/UN17_LIA_5	14	1	7.100
MDNA/UN17_LIA_6	12	2	16.600
MDNA/UN17_LIB	17	1	5.800
MDNA/UN17_LIB_1	12	2	16.600
MDNA/UN17_LIB_2	25	6	24.000
MDNA/UN17_LIB_3	31	12	38.700
MDNA/UN17_LIB_4	92	30	32.600

MDNA/UN17_LIB_5	31	5	16.100
MDNA/UN17_LIB_6	27	4	14.800
MDNA/UN17_LIB_7	16	5	31.200
MDNA/UN17_LIB_8	92	24	26.000
MDNA/UN17_LIB_9	47	15	31.900
MDA/UN17LIB_V	186	47	25.200
MDNA/UN17_LIC	66	13	19.600
MDNA/UN17_LIC_1	43	5	11.600
MDNA/UN17_LIC_2	65	18	27.600
MDNA/UN17_LIC_3	119	28	23.500
MDNA/UN17_LIC_4	53	15	28.300
MDNA/UN17_LIABC	117	39	33.300
MDNA/UN17_LIABC_1	108	30	27.700
MDNA/UN17_LIABC_2	78	15	19.200
MDNA/UN17_LIABC_3	97	17	17.500
MDA/UN17LIABC_L	192	64	33.300
	U17	**Média**	**23.343**
MDNA/UN18_LIA	61	11	18.000
MDNA/UN18_LIA_1	67	15	22.300
MDNA/UN18_LIA_2	116	32	27.500
MDNA/UN18_LIA_3	58	9	15.500
MDNA/UN18_LIB	20	2	10.000
MDNA/UN18_LIB_1	20	4	20.000
MDNA/UN18_LIB_2	11	7	63.600
MDNA/UN18_LIB_3	27	7	25.900
MDNA/UN18_LIB_4	10	1	10.000
MDNA/UN18_LIB_5	148	42	28.300
MDNA/UN18_LIB_6	81	30	37.000
MDNA/UN18_LIC	19	3	15.700
MDNA/UN18_LIC_1	15	6	40.000
MDNA/UN18_LIC_10	71	13	18.300
MDNA/UN18_LIC_11	159	38	23.800
MDNA/UN18_LIC_12	76	19	25.000
MDNA/UN18_LIC_2	15	0	0
MDNA/UN18_LIC_3	19	5	26.300
MDNA/UN18_LIC_4	65	25	38.400
MDNA/UN18_LIC_5	26	7	26.900
MDNA/UN18_LIC_6	39	15	38.400
MDNA/UN18_LIC_7	22	4	18.100
MDNA/UN18_LIC_8	16	5	31.200
MDNA/UN18_LIC_9	10	3	30.000
MDNA/UN18_LIABC	173	47	27.100
MDNA/UN18_LIABC_1	126	27	21.400
MDNA/UN18_LIABC_2	101	23	22.700
MDA/UN18LIABC_L	277	70	25.200
	U18	**Média**	**25.236**
MDNA/UN19_LIA	33	3	9.000

MDNA/UN19_LIA_1	19	2	10.500
MDNA/UN19_LIA_2	25	6	24.000
MDNA/UN19_LIA_3	84	6	7.100
MDNA/UN19_LIA_4	100	28	28.000
MDNA/UN19_LIA_5	65	12	18.400
MDNA/UN19_LIB	88	21	23.800
MDNA/UN19_LIB_1	62	11	17.700
MDNA/UN19_LIB_2	69	17	24.600
MDNA/UN19_LIB_3	178	47	26.400
MDNA/UN19_LIB_4	73	19	26.000
MDNA/UN19_LIC	54	10	18.500
MDNA/UN19_LIC_1	55	9	16.300
MDNA/UN19_LIC_2	85	16	18.800
MDNA/UN19_LIC_3	74	19	25.600
MDNA/UN19_LIC_4	72	14	19.400
MDNA/UN19_LIC_5	71	21	29.500
MDNA/UN19_LIABC	196	31	15.800
MDNA/UN19_LIABC_1	174	45	25.800
MDNA/UN19_LIABC_2	193	46	23.800
MDNA/UN19_LIABC_3	99	29	29.200
MDA/UN19LIABC_L	248	71	28.600
	U19	**Média**	**21.218**
MDNA/UN20_LIA	5	0	0
MDNA/UN20_LIA_1	118	37	31.300
MDNA/UN20_LIA_2	85	10	11.700
MDNA/UN20_LIA_3	145	37	25.500
MDNA/UN20_LIA_4	53	9	16.900
MDNA/UN20_LIB	28	6	21.400
MDNA/UN20_LIB_1	138	49	35.500
MDNA/UN20_LIB_2	70	20	28.500
MDNA/UN20_LIB_3	66	21	31.800
MDNA/UN20_LIB_4	202	36	17.800
MDNA/UN20_LIB_5	27	5	18.500
MDNA/UN20_LIB_6	11	1	9.000
MDNA/UN20_LIB_7	20	1	5.000
MDNA/UN20_LIB_8	42	19	45.200
MDNA/UN20_LIB_9	8	1	12.500
MDNA/UN20_LIC_9	21	6	28.500
MDNA/UN20_LIABC	321	58	18.000
MDA/UN20LIABC_L	338	69	20.400
MDA/UN20LIB	141	29	20.500
	U20	**Média**	**20.947**
MDNA/PR_5	104	21	20.100
MDNA/REV_5	38	7	18.400
MDNA/REV_5_1	15	0	0
MDNA/REV_5_10	26	6	23.000
MDNA/REV_5_11	23	6	26.000

MDNA/REV_5_12	39	9	23.000
MDNA/REV_5_13	36	16	44.400
MDNA/REV_5_2	14	3	21.400
MDNA/REV_5_3	10	3	30.000
MDNA/REV_5_4	27	5	18.500
MDNA/REV_5_5	14	3	21.400
MDNA/REV_5_6	115	22	19.100
MDNA/REV_5_7	20	3	15.000
MDNA/REV_5_8	33	10	30.300
MDNA/REV_5_9	20	5	25.000
	Rev e Pron 5	**Média**	**22.373**

Apêndice 6

Listas de Convergência texto a texto (por grau de autenticidade)

MD vs. BP oral				
arquivo	**trigramas - texto**	**trigramas convergentes (BP oral)**	**% de convergência**	**grau de autenticidade**
MDNA/UN2_LIA	10	0	0	muito baixo
MDNA/UN2_LIA_2	4	0	0	
MDNA/UN2_LIB_2	10	0	0	
MDNA/UN3_LIA_1	34	0	0	
MDNA/UN7_LIA_3	9	0	0	
MDNA/UN7_LIA_5	7	0	0	
MDNA/UN7_LIA_6	8	0	0	
MDNA/REV_2_3	3	0	0	
MDNA/UN10_LIA_10	13	0	0	
MDNA/REV_3_8	14	0	0	
MDNA/UN18_LIB_2	11	0	0	
MDNA/UN4_LIB_1	26	1	3.800	
MDNA/UN10_LIA_4	23	1	4.300	
MDNA/UN16_LIB	18	1	5.500	
MDNA/UN1_LIC_1	17	1	5.800	
MDNA/UN2_LIC_1	17	1	5.800	
MDNA/UN5_LIB	17	1	5.800	
MDNA/UN5_LIA_1	33	2	6.000	
MDNA/UN14_LIB_1	33	2	6.000	
MDNA/UN8_LIA_1	15	1	6.600	
MDNA/UN14_LIC_3	14	1	7.100	
MDNA/UN10_LIA_2	26	2	7.600	
MDNA/UN17_LIB_1	12	1	8.300	
MDA/UN4LIABC_L	106	9	8.400	
MDNA/UN5_LIB_3	23	2	8.600	
MDNA/UN10_LIB	23	2	8.600	
MDNA/UN4_LIA_1	11	1	9.000	
MDNA/UN4_LIB_2	84	8	9.500	
MDNA/UN5_LIABC_1	51	5	9.800	
MDNA/UN17_LIA_3	112	11	9.800	
MDNA/UN2_LIA_4	10	1	10.000	
MDNA/UN2_LIB_1	10	1	10.000	
MDNA/UN18_LIC_9	10	1	10.000	
MDNA/UN4_LIC_1	39	4	10.200	
MDNA/UN5_LIA	38	4	10.500	
MDNA/UN10_LIB_2	19	2	10.500	
MDNA/UN10_LIA_3	28	3	10.700	

MDNA/UN3_LIA	17	2	11.700	baixo
MDNA/UN7_LIB_9	34	4	11.700	
MDNA/UN16_LIB_4	101	12	11.800	
MDNA/UN12_LIA_1	42	5	11.900	
MDNA/UN4_LIA_4	24	3	12.500	
MDNA/UN6_LIC	40	5	12.500	
MDNA/UN17_LIB_7	16	2	12.500	
MDNA/UN20_LIB_8	8	1	12.500	
MDNA/UN6_LIC_4	71	9	12.600	
MDNA/UN7_LIC	53	7	13.200	
MDNA/UN14_LIA_5	53	7	13.200	
MDNA/UN2_LIA_5	15	2	13.300	
MDNA/UN5_LIABC	52	7	13.400	
MDNA/UN2_LIC_2	37	5	13.500	
MDNA/UN1_LIB	36	5	13.800	
MDNA/UN16_LIC_6	57	8	14.000	
MDNA/UN13_LIA_2	63	9	14.200	
MDNA/UN17_LIA_5	14	2	14.200	
MDNA/UN11_LIB	20	3	15.000	
MDNA/UN5_LIB_4	33	5	15.100	
MDNA/UN7_LIA_1	13	2	15.300	
MDNA/UN8_LIB_3	39	6	15.300	
MDNA/UN9_LIABC_2	32	5	15.600	
MDNA/REV_3_14	19	3	15.700	
MDNA/UN9_LIA_1	63	10	15.800	
MDNA/UN17_LIB_2	25	4	16.000	
MDNA/UN1_LIC	31	5	16.100	
MDNA/UN4_LIABC	93	15	16.100	
MDNA/UN5_LIABC_2	31	5	16.100	
MDA/UN3LIABC_L	61	10	16.300	
MDNA/REV_2_1	24	4	16.600	
MDNA/UN9_LIABC	54	9	16.600	
MDNA/UN14_LIA_1	6	1	16.600	
MDNA/UN15_LIA_8	18	3	16.600	
MDNA/UN20_LIB_2	65	11	16.900	
MDNA/UN12_LIB_2	41	7	17.000	
MDNA/UN17_LIB_9	47	8	17.000	
MDNA/UN9_LIABC_1	81	14	17.200	
MDNA/UN11_LIC_2	168	29	17.200	
MDNA/REV_2	46	8	17.300	
MDNA/UN17_LIABC_1	108	19	17.500	
MDNA/UN5_LIC_7	17	3	17.600	
MDNA/UN8_LIABC	51	9	17.600	
MDNA/REV_2_8	17	3	17.600	
MDNA/UN14_LIB_3	34	6	17.600	
MDNA/UN2_LIABC_2	45	8	17.700	
MDNA/UN10_LIB_4	95	17	17.800	

MDNA/UN3_LIABC_2	22	4	18.100	
MDNA/UN10_LIB_3	66	12	18.100	
MDNA/UN10_LIABC_1	77	14	18.100	
MDNA/UN14_LIA_3	66	12	18.100	
MDNA/UN15_LIC_3	22	4	18.100	
MDNA/UN17_LIC	66	12	18.100	
MDNA/UN3_LIB_1	38	7	18.400	
MDNA/UN7_LIB	76	14	18.400	
MDNA/UN6_LIB_2	27	5	18.500	
MDNA/UN9_LIC_2	102	19	18.600	
MDNA/UN7_LIA_2	112	21	18.700	
MDNA/UN18_LIC_8	16	3	18.700	
MDNA/UN10_LIB_5	90	17	18.800	
MDNA/UN11_LIB_5	53	10	18.800	
MDNA/PR_5	104	20	19.200	
MDNA/UN11_LIC_1	51	10	19.600	
MDNA/UN17_LIABC	117	23	19.600	
MDNA/UN19_LIABC_1	172	34	19.700	
MDNA/UN1_LIABC_1	15	3	20.000	
MDNA/UN2_LIA_1	25	5	20.000	
MDNA/UN6_LIC_3	25	5	20.000	
MDNA/UN7_LIB_5	15	3	20.000	
MDNA/REV_2_7	10	2	20.000	
MDNA/UN10_LIA	30	6	20.000	
MDNA/UN10_LIA_6	15	3	20.000	
MDNA/UN13_LIB	60	12	20.000	
MDNA/UN16_LIC	35	7	20.000	
MDNA/UN16_LIC_1	10	2	20.000	
MDNA/REV_4_11	10	2	20.000	
MDNA/UN14_LIABC_1	59	12	20.300	
MDNA/UN7_LIABC	83	17	20.400	
MDNA/UN10_LIC_3	53	11	20.700	
MDNA/UN7_LIB_6	24	5	20.800	
MDNA/UN15_LIA	24	5	20.800	
MDNA/UN19_LIB_4	72	15	20.800	
MDNA/UN1_LIB_3	19	4	21.000	bom
MDNA/UN7_LIB_1	38	8	21.000	
MDA/UN16LIB_V	57	12	21.000	
MDNA/UN7_LIA	85	18	21.100	
MDNA/UN13_LIA_5	109	23	21.100	
MDNA/REV_5_2	14	3	21.400	
MDNA/UN17_LIB_8	92	20	21.700	
MDNA/UN8_LIC	96	21	21.800	
MDNA/REV_2_9	32	7	21.800	
MDNA/UN14_LIC_4	32	7	21.800	
MDNA/UN20_LIB_7	41	9	21.900	
MDNA/UN12_LIA_3	59	13	22.000	

MDNA/UN16_LIA_1	77	17	22.000	
MDNA/UN10_LIA_12	113	25	22.100	
MDNA/UN4_LIA_3	18	4	22.200	
MDNA/UN4_LIA_8	18	4	22.200	
MDNA/UN5_LIABC_3	27	6	22.200	
MDNA/UN12_LIABC_3	27	6	22.200	
MDNA/UN15_LIA_1	36	8	22.200	
MDNA/UN17_LIB_6	27	6	22.200	
MDNA/REV_5_3	9	2	22.200	
MDNA/UN19_LIC_2	85	19	22.300	
MDNA/UN12_LIABC	107	24	22.400	
MDNA/REV_3_7	31	7	22.500	
MDNA/UN8_LIA_4	53	12	22.600	
MDNA/UN4_LIA	79	18	22.700	
MDNA/UN5_LIC	22	5	22.700	
MDNA/UN5_LIC_3	22	5	22.700	
MDNA/UN7_LIA_7	79	18	22.700	
MDNA/REV_3_11	22	5	22.700	
MDNA/UN18_LIA_1	66	15	22.700	
MDNA/UN11_LIA_1	57	13	22.800	
MDNA/UN9_LIA	183	42	22.900	
MDNA/UN6_LIC_2	39	9	23.000	
MDNA/UN10_LIC_2	26	6	23.000	
MDNA/UN12_LIA_2	39	9	23.000	
MDNA/UN18_LIC_5	26	6	23.000	
MDA/UN17LIB_V	186	43	23.100	
MDNA/UN1_LIABC_3	43	10	23.200	
MDNA/UN12_LIA_4	73	17	23.200	
MDNA/UN10_LIABC_2	60	14	23.300	
MDNA/UN13_LIB_1	77	18	23.300	
MDNA/UN15_LIABC_1	60	14	23.300	
MDNA/UN3_LIB_2	47	11	23.400	
MDNA/UN4_LIA_2	17	4	23.500	
MDNA/UN4_LIB	119	28	23.500	
MDNA/REV_2_2	17	4	23.500	
MDNA/UN17_LIB	17	4	23.500	
MDNA/UN4_LIABC_1	72	17	23.600	
MDNA/UN12_LIB_4	38	9	23.600	
MDNA/UN4_LIA_7	105	25	23.800	
MDA/UN7LIABC_L	155	37	23.800	
MDNA/REV_3_4	21	5	23.800	
MDNA/UN19_LIB_3	176	42	23.800	
MDNA/UN5_LIC_6	54	13	24.000	
MDA/UN11LIC_V	150	36	24.000	
MDNA/REV_4	25	6	24.000	
MDNA/UN7_LIB_2	62	15	24.100	
MDNA/UN16_LIA_2	99	24	24.200	

MDNA/UN20_LIB_1	70	17	24.200	
MDNA/UN17_LIC_3	119	29	24.300	
MDNA/UN17_LIABC_2	78	19	24.300	
MDNA/UN12_LIA_5	127	31	24.400	
MDNA/UN13_LIC_1	98	24	24.400	
MDNA/UN15_LIC_6	86	21	24.400	
MDNA/UN17_LIC_4	53	13	24.500	
MDNA/UN17_LIC_2	65	16	24.600	
MDNA/UN5_LIB_1	97	24	24.700	
MDNA/UN10_LIC_1	109	27	24.700	
MDNA/REV_3_15	117	29	24.700	
MDNA/REV_4_1	117	29	24.700	
MDNA/UN5_LIB_2	12	3	25.000	
MDNA/UN7_LIA_4	8	2	25.000	
MDNA/PR_2	4	1	25.000	
MDNA/UN13_LIA	40	10	25.000	
MDNA/UN13_LIA_1	16	4	25.000	
MDNA/UN14_LIB_6	56	14	25.000	
MDNA/REV_4_7	24	6	25.000	
MDNA/UN19_LIC_4	72	18	25.000	
MDNA/UN11_LIABC_1	87	22	25.200	
MDA/UN11LIABC_L	150	38	25.300	
MDNA/UN3_LIC_1	51	13	25.400	
MDNA/REV_3	55	14	25.400	
MDNA/UN18_LIABC	173	44	25.400	
MDNA/UN19_LIABC_2	191	49	25.600	
MDNA/UN17_LIB_3	31	8	25.800	
MDNA/UN2_LIB	27	7	25.900	
MDNA/REV_3_12	27	7	25.900	
MDNA/UN15_LIA_7	27	7	25.900	
MDNA/UN9_LIB_1	50	13	26.000	
MDNA/UN19_LIA_4	100	26	26.000	
MDNA/UN19_LIABC_3	99	26	26.200	
MDNA/UN10_LIABC	95	25	26.300	
MDA/UN10LIABC_L	167	44	26.300	
MDNA/REV_3_3	38	10	26.300	
MDNA/UN19_LIA_1	19	5	26.300	
MDNA/REV_5	38	10	26.300	
MDNA/UN11_LIC_3	68	18	26.400	
MDNA/UN7_LIB_8	64	17	26.500	
MDNA/UN4_LIA_6	15	4	26.600	
MDNA/UN10_LIA_7	15	4	26.600	
MDNA/UN12_LIABC_1	105	28	26.600	
MDNA/REV_3_10	15	4	26.600	
MDNA/UN20_LIA_3	145	39	26.800	
MDNA/UN1_LIABC	26	7	26.900	
MDNA/UN10_LIA_11	63	17	26.900	

MDNA/UN14_LIA	52	14	26.900	
MDNA/UN7_LIABC_1	48	13	27.000	
MDNA/UN12_LIC	37	10	27.000	
MDNA/UN10_LIC	66	18	27.200	
MDNA/UN11_LIABC	77	21	27.200	
MDNA/UN18_LIC_7	22	6	27.200	
MDNA/UN3_LIC	51	14	27.400	
MDNA/UN3_LIABC	40	11	27.500	
MDNA/PR_1	29	8	27.500	
MDNA/UN9_LIB_4	69	19	27.500	
MDNA/UN8_LIA_6	83	23	27.700	
MDNA/UN13_LIB_4	126	35	27.700	
MDNA/UN14_LIC	18	5	27.700	
MDNA/UN15_LIB_4	36	10	27.700	
MDNA/UN18_LIB_5	148	41	27.700	
MDNA/UN19_LIC	54	15	27.700	
MDNA/REV_5_13	36	10	27.700	
MDNA/UN11_LIABC_2	104	29	27.800	
MDNA/UN12_LIC_4	176	49	27.800	
MDNA/UN15_LIABC	79	22	27.800	
MDNA/REV_5_6	115	32	27.800	
MDNA/UN4_LIC_3	86	24	27.900	
MDNA/UN4_LIC_4	43	12	27.900	
MDNA/UN13_LIC_3	129	36	27.900	
MDNA/UN6_LIA_2	64	18	28.100	
MDNA/UN11_LIB_1	32	9	28.100	
MDNA/UN13_LIABC	103	29	28.100	
MDNA/UN4_LIABC_2	81	23	28.300	
MDNA/UN16_LIB_3	109	31	28.400	
MDNA/UN2_LIABC_1	21	6	28.500	
MDNA/UN10_LIB_1	63	18	28.500	
MDNA/REV_3_1	14	4	28.500	
MDNA/UN13_LIA_4	119	34	28.500	
MDNA/UN15_LIA_6	14	4	28.500	
MDNA/UN16_LIC_3	98	28	28.500	
MDNA/UN19_LIC_5	70	20	28.500	
MDA/UN13LIABC_L	226	65	28.700	
MDNA/UN6_LIABC	111	32	28.800	
MDNA/UN8_LIA_3	52	15	28.800	
MDNA/REV_4_4	45	13	28.800	
MDNA/UN8_LIA	38	11	28.900	
MDNA/UN9_LIB	38	11	28.900	
MDNA/UN11_LIB_2	38	11	28.900	
MDNA/UN16_LIC_7	38	11	28.900	
MDNA/REV_5_12	38	11	28.900	
MDNA/UN16_LIABC	279	81	29.000	
MDNA/UN19_LIB_1	62	18	29.000	

MDNA/UN19_LIABC	196	57	29.000	
MDNA/UN15_LIB_3	96	28	29.100	
MDNA/UN1_LIABC_2	17	5	29.400	
MDNA/UN2_LIA_7	17	5	29.400	
MDNA/UN6_LIC_1	34	10	29.400	
MDNA/UN5_LIC_1	54	16	29.600	
MDNA/UN5_LIC_4	54	16	29.600	
MDNA/UN20_LIA_1	118	35	29.600	
MDNA/UN8_LIB	37	11	29.700	
MDNA/UN13_LIABC_2	77	23	29.800	
MDNA/UN14_LIA_2	77	23	29.800	
MDNA/UN14_LIABC_2	67	20	29.800	
MDNA/UN12_LIC_1	20	6	30.000	
MDNA/REV_3_13	10	3	30.000	
MDNA/UN4_LIC	73	22	30.100	
MDNA/UN5_LIC_8	43	13	30.200	
MDA/UN17LIABC_L	192	58	30.200	
MDNA/UN5_LIABC_4	92	28	30.400	
MDNA/REV_3_2	23	7	30.400	
MDNA/UN15_LIABC_2	105	32	30.400	
MDNA/UN6_LIA_1	88	27	30.600	
MDNA/UN8_LIB_4	173	53	30.600	
MDNA/UN20_LIB	137	42	30.600	
MDNA/UN8_LIA_2	13	4	30.700	
MDNA/UN8_LIB_2	52	16	30.700	
MDNA/REV_4_8	52	16	30.700	
MDNA/UN17_LIB_4	91	28	30.700	
MDNA/UN19_LIA_5	65	20	30.700	
MDNA/REV_5_10	26	8	30.700	
MDNA/UN4_LIC_2	81	25	30.800	
MDNA/UN5_LIA_2	55	17	30.900	
MDNA/UN15_LIB_1	84	26	30.900	
MDNA/UN9_LIC_1	90	28	31.100	
MDNA/UN15_LIC_8	45	14	31.100	
MDNA/UN1_LIB_2	48	15	31.200	
MDNA/UN2_LIB_4	32	10	31.200	
MDNA/REV_2_4	32	10	31.200	
MDNA/UN13_LIC_2	80	25	31.200	
MDNA/UN15_LIC	64	20	31.200	
MDNA/UN13_LIABC_1	99	31	31.300	
MDNA/UN9_LIC	127	40	31.400	
MDNA/UN6_LIC_5	57	18	31.500	
MDNA/UN15_LIA_2	19	6	31.500	
MDNA/UN18_LIC_3	19	6	31.500	
MDNA/UN6_LIA	79	25	31.600	
MDNA/UN10_LIA_5	41	13	31.700	
MDNA/UN12_LIB_8	41	13	31.700	

MDNA/UN3_LIB	50	16	32.000	alto
MDNA/UN3_LIC_2	50	16	32.000	
MDNA/UN5_LIB_5	50	16	32.000	
MDNA/UN13_LIB_3	215	69	32.000	
MDNA/UN15_LIC_5	53	17	32.000	
MDNA/UN17_LIA_1	25	8	32.000	
MDNA/UN18_LIABC_1	125	40	32.000	
MDNA/UN1_LIB_1	28	9	32.100	
MDNA/REV_1	176	57	32.300	
MDNA/UN12_LIB	43	14	32.500	
MDNA/PR_3	49	16	32.600	
MDNA/UN14_LIA_6	52	17	32.600	
MDNA/UN13_LIA_3	55	18	32.700	
MDNA/UN14_LIB_5	76	25	32.800	
MDNA/UN15_LIA_10	149	49	32.800	
MDA/UN9_LIABC_L	208	69	33.100	
MDNA/UN1_LIA_2	9	3	33.300	
MDNA/UN3_LIA_2	15	5	33.300	
MDNA/UN3_LIABC_1	39	13	33.300	
MDNA/UN8_LIA_5	30	10	33.300	
MDNA/UN9_LIB_2	48	16	33.300	
MDNA/UN10_LIA_9	15	5	33.300	
MDNA/REV_3_6	12	4	33.300	
MDNA/UN13_LIABC_3	12	4	33.300	
MDNA/UN16_LIB_1	30	10	33.300	
MDNA/UN16_LIC_4	36	12	33.300	
MDNA/UN17_LIA_2	75	25	33.300	
MDA/UN19LIABC_L	246	82	33.300	
MDA/UN20LIABC_L	140	47	33.500	
MDNA/UN18_LIA_2	116	39	33.600	
MDA/UN15LIABC_L	352	119	33.800	
MDNA/UN16_LIC_2	47	16	34.000	
MDNA/UN20_LIB_3	202	69	34.100	
MDNA/UN9_LIA_2	70	24	34.200	
MDA/UN16LIABC_L	332	114	34.300	
MDNA/UN12_LIC_5	58	20	34.400	
MDNA/UN14_LIABC	90	31	34.400	
MDNA/UN13_LIB_2	81	28	34.500	
MDNA/UN3_LIA_3	26	9	34.600	
MDNA/UN2_LIABC_3	23	8	34.700	
MDNA/REV_4_6	23	8	34.700	
MDNA/UN20_LIABC	338	118	34.900	
MDNA/UN9_LIB_3	77	27	35.000	
MDNA/UN14_LIC_7	57	20	35.000	
MDNA/UN18_LIB	20	7	35.000	
MDNA/REV_2_5	17	6	35.200	
MDNA/UN12_LIABC_2	139	49	35.200	

MDNA/UN20_LIC_9	321	113	35.200	
MDNA/UN5_LIC_2	31	11	35.400	
MDNA/UN17_LIB_5	31	11	35.400	
MDNA/UN14_LIC_6	87	31	35.600	
MDA/UN6LIABC_L	42	15	35.700	
MDNA/REV_2_6	14	5	35.700	
MDNA/UN11_LIB_3	70	25	35.700	
MDNA/UN16_LIA	95	34	35.700	
MDNA/UN19_LIA_3	84	30	35.700	
MDNA/UN6_LIA_3	97	35	36.000	
MDNA/UN1_LIA_1	11	4	36.300	
MDNA/UN14_LIB	22	8	36.300	
MDNA/UN20_LIB_5	11	4	36.300	
MDNA/REV_4_12	41	15	36.500	
MDA/UN12LIABC_L	283	104	36.700	
MDNA/UN14_LIB_2	57	21	36.800	
MDNA/UN18_LIB_6	81	30	37.000	
MDNA/UN20_LIB_4	27	10	37.000	
MDNA/UN15_LIA_3	43	16	37.200	
MDNA/PR_4	91	34	37.300	
MDNA/UN1_LIA	16	6	37.500	
MDNA/UN12_LIA_6	48	18	37.500	
MDNA/UN12_LIB_6	80	30	37.500	
MDNA/UN15_LIA_9	32	12	37.500	
MDNA/UN16_LIC_5	8	3	37.500	
MDNA/UN18_LIABC_2	101	38	37.600	
MDNA/UN12_LIA	37	14	37.800	
MDA/UN18LIABC_L	275	104	37.800	
MDNA/UN2_LIB_5	21	8	38.000	
MDNA/UN18_LIC_12	76	29	38.100	
MDNA/UN15_LIB_2	86	33	38.300	
MDNA/UN7_LIB_4	39	15	38.400	
MDNA/UN8_LIB_1	65	25	38.400	
MDNA/REV_2_10	13	5	38.400	
MDNA/UN15_LIB	52	20	38.400	
MDNA/UN19_LIB	88	34	38.600	
MDNA/UN12_LIB_1	18	7	38.800	
MDNA/UN12_LIB_7	144	56	38.800	
MDNA/REV_4_10	36	14	38.800	
MDNA/UN15_LIC_4	154	60	38.900	
MDNA/UN6_LIB_3	69	27	39.100	
MDNA/UN11_LIA_2	78	31	39.700	
MDNA/UN2_LIC	10	4	40.000	
MDNA/UN10_LIA_8	10	4	40.000	
MDNA/UN14_LIB_7	10	4	40.000	
MDNA/UN14_LIC_2	5	2	40.000	
MDNA/UN18_LIC_4	65	26	40.000	

MDNA/REV_5_1	15	6	40.000	
MDNA/REV_5_7	20	8	40.000	
MDNA/REV_5_9	20	8	40.000	
MDNA/UN18_LIC_11	157	63	40.100	
MDNA/UN11_LIA	57	23	40.300	
MDNA/UN15_LIC_1	47	19	40.400	
MDNA/UN12_LIC_2	69	28	40.500	
MDNA/UN17_LIA_4	59	24	40.600	
MDNA/UN10_LIA_1	22	9	40.900	
MDNA/REV_3_9	22	9	40.900	
MDNA/UN14_LIC_5	22	9	40.900	
MDNA/UN15_LIC_7	105	43	40.900	
MDA/UN14LIABC_L	78	32	41.000	
MDNA/UN18_LIC_6	39	16	41.000	
MDNA/UN6_LIB	72	30	41.600	
MDNA/UN17_LIA_6	12	5	41.600	
MDNA/UN19_LIC_3	74	31	41.800	
MDNA/UN14_LIB_4	19	8	42.100	
MDNA/REV_3_5	35	15	42.800	
MDNA/UN14_LIC_1	56	24	42.800	
MDNA/REV_4_9	35	15	42.800	
MDNA/UN20_LIB_9	21	9	42.800	
MDNA/REV_5_5	14	6	42.800	
MDNA/UN13_LIC	107	46	42.900	
MDNA/UN18_LIA	60	26	43.300	
MDNA/UN19_LIB_2	69	30	43.400	
MDNA/UN4_LIA_5	39	17	43.500	
MDNA/UN12_LIB_5	101	44	43.500	
MDNA/UN18_LIC_10	71	31	43.600	
MDNA/UN19_LIC_1	55	24	43.600	
MDNA/UN11_LIA_3	48	21	43.700	
MDNA/UN5_LIC_5	25	11	44.000	
MDNA/UN19_LIA_2	25	11	44.000	
MDNA/UN7_LIB_3	68	30	44.100	
MDA/UN3LIC_P	18	8	44.400	
MDNA/UN15_LIA_5	20	9	45.000	
MDNA/UN20_LIB_6	20	9	45.000	
MDNA/UN11_LIB_4	150	68	45.300	
MDNA/UN6_LIB_1	44	20	45.400	
MDNA/UN12_LIC_3	110	50	45.400	
MDNA/UN1_LIA_3	13	6	46.100	
MDNA/UN17_LIABC_3	97	45	46.300	
MDA/UN20LIB	28	13	46.400	
MDNA/UN18_LIA_3	58	27	46.500	
MDNA/UN18_LIC_1	15	7	46.600	
MDNA/REV_4_2	64	30	46.800	
MDNA/REV_4_3	64	30	46.800	

MDNA/UN15_LIC_2	34	16	47.000	
MDNA/UN17_LIA	55	26	47.200	
MDNA/UN18_LIC	19	9	47.300	
MDNA/REV_5_11	23	11	47.800	
MDNA/UN8_LIABC_2	52	25	48.000	
MDNA/UN14_LIB_8	50	24	48.000	
MDNA/UN11_LIC	87	42	48.200	
MDNA/UN19_LIA	33	16	48.400	
MDNA/UN20_LIA_4	53	26	49.000	
MDNA/UN20_LIA_2	85	42	49.400	
MDNA/UN8_LIABC_1	76	38	50.000	
MDNA/UN18_LIB_4	10	5	50.000	
MDNA/UN12_LIB_3	58	30	51.700	
MDNA/REV_5_4	27	14	51.800	
MDNA/UN13_LIC_4	69	36	52.100	
MDNA/UN16_LIB_2.txt	46	24	52.100	
MDNA/UN18_LIC_2	15	8	53.300	
MDNA/UN17_LIC_1	43	23	53.400	
MDNA/UN18_LIB_1	20	11	55.000	
MDNA/UN2_LIABC	16	9	56.200	
MDNA/UN20_LIA	5	3	60.000	
MDNA/UN2_LIB_3	13	8	61.500	
MDNA/REV_4_13	8	5	62.500	
MDNA/UN18_LIB_3	27	17	62.900	
MDNA/UN7_LIB_7	25	16	64.000	
MDNA/REV_5_8	33	22	66.600	
MDNA/UN15_LIA_4	4	3	75.000	
MDNA/REV_4_5	16	12	75.000	
MDNA/UN14_LIA_4	7	6	85.700	

MD vs. BP escrito				
arquivo	**trigramas - texto**	**trigramas convergentes (BP escrito)**	**% de convergência**	**grau de autenticidade**
MDNA/UN2_LIB_4	32	0	0	muito baixo
MDNA/REV_2_4	17	0	0	
MDNA/UN10_LIA_6	15	0	0	
MDNA/UN13_LIA_1	16	0	0	
MDNA/UN14_LIA_4	7	0	0	
MDNA/UN16_LIC_5	8	0	0	
MDNA/UN18_LIC_2	15	0	0	
MDNA/UN20_LIA	5	0	0	
MDNA/REV_5_1	15	0	0	
MDNA/UN10_LIA	30	1	3.300	
MDNA/UN16_LIB_2	46	2	4.300	
MDNA/UN15_LIA_5	20	1	5.000	

MDNA/UN20_LIB_7	20	1	5.000	
MDNA/UN15_LIA_2	19	1	5.200	
MDNA/UN8_LIB	37	2	5.400	
MDNA/UN17_LIB	17	1	5.800	
MDNA/REV_4_5	16	1	6.200	
MDNA/UN7_LIB_5	15	1	6.600	
MDNA/UN17_LIA_5	14	1	7.100	
MDNA/UN19_LIA_3	84	6	7.100	
MDNA/UN13_LIA	40	3	7.500	
MDNA/UN8_LIB_2	52	4	7.600	
MDNA/UN16_LIC_7	38	3	7.800	
MDNA/UN3_LIB	50	4	8.000	
MDNA/UN3_LIC_2	50	4	8.000	
MDNA/UN12_LIA	37	3	8.100	
MDNA/UN9_LIB_2	48	4	8.300	
MDNA/REV_3_6	12	1	8.300	
MDNA/UN13_LIABC_3	12	1	8.300	
MDNA/REV_4_7	24	2	8.300	
MDNA/UN4_LIC_2	81	7	8.600	
MDNA/REV_4_6	23	2	8.600	
MDNA/UN10_LIA_1	22	2	9.000	
MDNA/UN19_LIA	33	3	9.000	
MDNA/UN20_LIB_6	11	1	9.000	
MDNA/UN2_LIABC_1	21	2	9.500	
MDNA/UN2_LIB_2	10	1	10.000	
MDNA/UN9_LIC_1	90	9	10.000	
MDNA/UN14_LIB_7	10	1	10.000	
MDNA/UN18_LIB	20	2	10.000	
MDNA/UN18_LIB_4	10	1	10.000	
MDNA/UN1_LIB_3	19	2	10.500	
MDNA/REV_3_14	19	2	10.500	
MDNA/UN14_LIB_2	57	6	10.500	
MDNA/UN19_LIA_1	19	2	10.500	
MDNA/UN2_LIB	27	3	11.100	baixo
MDNA/UN10_LIB_1	63	7	11.100	
MDNA/UN13_LIB_2	81	9	11.100	
MDNA/UN11_LIC	87	10	11.400	
MDNA/UN10_LIA_2	26	3	11.500	
MDNA/UN13_LIC_4	69	8	11.500	
MDNA/UN17_LIC_1	43	5	11.600	
MDNA/UN20_LIA_2	85	10	11.700	
MDNA/UN5_LIC_5	25	3	12.000	
MDNA/UN7_LIB_7	25	3	12.000	
MDNA/UN14_LIABC	90	11	12.200	
MDNA/UN2_LIABC	16	2	12.500	
MDNA/UN11_LIABC_2	104	13	12.500	
MDNA/REV_4_13	8	1	12.500	

MDNA/UN20_LIB_9	8	1	12.500	
MDNA/UN3_LIB_2	47	6	12.700	
MDNA/UN5_LIC_2	31	4	12.900	
MDNA/UN11_LIB_4	152	20	13.100	
MDNA/UN15_LIC_8	45	6	13.300	
MDNA/UN5_LIB_1	97	13	13.400	
MDNA/UN14_LIA	52	7	13.400	
MDNA/UN12_LIB_6	80	11	13.700	
MDNA/UN1_LIB	36	5	13.800	
MDNA/UN12_LIB	43	6	13.900	
MDNA/UN15_LIB_2	86	12	13.900	
MDNA/UN9_LIA	185	26	14.000	
MDNA/UN9_LIB_1	50	7	14.000	
MDNA/REV_4_2	64	9	14.000	
MDNA/REV_4_3	64	9	14.000	
MDNA/UN13_LIABC_1	99	14	14.100	
MDA/UN14LIABC_L	78	11	14.100	
MDNA/UN15_LIA_6	14	2	14.200	
MDNA/UN13_LIA_3	55	8	14.500	
MDNA/UN6_LIA_1	88	13	14.700	
MDNA/UN15_LIA_7	27	4	14.800	
MDNA/UN17_LIB_6	27	4	14.800	
MDA/UN13LIABC_L	227	34	14.900	
MDNA/REV_5_7	20	3	15.000	
MDNA/UN10_LIB_3	66	10	15.100	
MDNA/UN14_LIA_3	66	10	15.100	
MDNA/UN8_LIA_3	52	8	15.300	
MDNA/UN8_LIB_1	65	10	15.300	
MDNA/UN8_LIABC_2	52	8	15.300	
MDNA/UN12_LIA_2	39	6	15.300	
MDNA/UN15_LIB	52	8	15.300	
MDNA/UN18_LIA_3	58	9	15.500	
MDNA/UN15_LIA_9	32	5	15.600	
MDNA/UN11_LIA	57	9	15.700	
MDNA/UN18_LIC	19	3	15.700	
MDNA/UN19_LIABC	196	31	15.800	
MDNA/UN6_LIB_1	44	7	15.900	
MDNA/UN4_LIABC_2	81	13	16.000	
MDNA/UN14_LIB_8	50	8	16.000	
MDNA/UN17_LIB_5	31	5	16.100	
MDNA/UN15_LIA_3	43	7	16.200	
MDNA/UN19_LIC_1	55	9	16.300	
MDNA/UN4_LIA	79	13	16.400	
MDNA/UN4_LIA_3	18	3	16.600	
MDNA/UN4_LIA_4	24	4	16.600	
MDNA/UN4_LIA_8	18	3	16.600	
MDNA/REV_2	24	4	16.600	

MDNA/UN9_LIABC	54	9	16.600	
MDNA/UN12_LIA_1	42	7	16.600	
MDNA/UN12_LIA_6	48	8	16.600	
MDNA/UN14_LIC	18	3	16.600	
MDNA/UN15_LIA	24	4	16.600	
MDNA/UN15_LIB_1	84	14	16.600	
MDNA/UN15_LIB_4	36	6	16.600	
MDNA/UN17_LIA_6	12	2	16.600	
MDNA/UN17_LIB_1	12	2	16.600	
MDNA/UN20_LIA_4	53	9	16.900	
MDNA/UN9_LIC	129	22	17.000	
MDNA/UN12_LIB_8	41	7	17.000	
MDNA/UN6_LIB_3	70	12	17.100	
MDNA/UN11_LIB_3	70	12	17.100	
MDNA/PR_1	29	5	17.200	
MDNA/PR_2	46	8	17.300	
MDNA/UN14_LIA_6	52	9	17.300	
MDNA/UN3_LIABC	40	7	17.500	
MDNA/UN16_LIC_6	57	10	17.500	
MDNA/UN17_LIABC_3	97	17	17.500	
MDNA/UN1_LIC_1	17	3	17.600	
MDNA/UN4_LIA_2	17	3	17.600	
MDNA/REV_2_7	17	3	17.600	
MDNA/UN11_LIC_3	68	12	17.600	
MDNA/UN6_LIA	79	14	17.700	
MDNA/UN19_LIB_1	62	11	17.700	
MDNA/UN4_LIB_2	84	15	17.800	
MDNA/UN10_LIA_3	28	5	17.800	
MDNA/UN12_LIABC_2	140	25	17.800	
MDNA/UN20_LIB_4	202	36	17.800	
MDNA/UN4_LIC_1	39	7	17.900	
MDNA/UN15_LIABC_2	105	19	18.000	
MDNA/UN18_LIA	61	11	18.000	
MDNA/UN20_LIABC	321	58	18.000	
MDNA/UN9_LIB_3	77	14	18.100	
MDNA/REV_3	55	10	18.100	
MDNA/UN18_LIC_7	22	4	18.100	
MDNA/UN9_LIA_2	71	13	18.300	
MDNA/UN13_LIB	60	11	18.300	
MDNA/UN18_LIC_10	71	13	18.300	
MDNA/REV_3_3	38	7	18.400	
MDNA/UN19_LIA_5	65	12	18.400	
MDNA/REV_5	38	7	18.400	
MDNA/UN19_LIC	54	10	18.500	
MDNA/UN20_LIB_5	27	5	18.500	
MDNA/REV_5_4	27	5	18.500	
MDA/UN11LIC_V	150	28	18.600	

MDNA/UN1_LIA	16	3	18.700	
MDNA/REV_2_8	32	6	18.700	
MDNA/UN13_LIC_2	80	15	18.700	
MDNA/UN10_LIC_3	53	10	18.800	
MDNA/UN19_LIC_2	85	16	18.800	
MDNA/UN12_LIC_5	58	11	18.900	
MDA/UN6LIABC_L	42	8	19.000	
MDA/UN10LIABC_L	168	32	19.000	
MDNA/UN12_LIA_4	73	14	19.100	
MDNA/REV_5_6	115	22	19.100	
MDNA/UN5_LIABC	52	10	19.200	
MDNA/UN7_LIABC	83	16	19.200	
MDNA/UN14_LIA_2	78	15	19.200	
MDA/UN16LIB_V	57	11	19.200	
MDNA/REV_4_8	52	10	19.200	
MDNA/UN17_LIABC_2	78	15	19.200	
MDNA/UN15_LIC_4	155	30	19.300	
MDNA/UN12_LIB_7	144	28	19.400	
MDNA/UN16_LIC_4	36	7	19.400	
MDNA/UN19_LIC_4	72	14	19.400	
MDNA/UN3_LIC	51	10	19.600	
MDA/UN9_LIABC_L	209	41	19.600	
MDNA/UN17_LIC	66	13	19.600	
MDA/UN4LIABC_L	106	21	19.800	
MDNA/UN2_LIC	10	2	20.000	
MDNA/REV_2_6	10	2	20.000	
MDNA/UN12_LIC_1	20	4	20.000	
MDNA/REV_3_13	10	2	20.000	
MDNA/UN14_LIC_2	5	1	20.000	
MDNA/REV_4_9	35	7	20.000	
MDNA/UN18_LIB_1	20	4	20.000	
MDNA/UN13_LIA_5	109	22	20.100	
MDNA/PR_5	104	21	20.100	
MDNA/UN12_LIA_3	59	12	20.300	
MDNA/UN15_LIC	64	13	20.300	
MDNA/UN8_LIA_6	83	17	20.400	
MDA/UN20LIABC_L	338	69	20.400	
MDA/UN20LIB	141	29	20.500	
MDNA/UN15_LIB_3	97	20	20.600	
MDNA/PR_4	92	19	20.600	
MDNA/UN8_LIA_4	53	11	20.700	
MDNA/UN12_LIB_5	101	21	20.700	
MDNA/UN14_LIABC_2	67	14	20.800	
MDNA/UN4_LIC_4	43	9	20.900	
MDNA/UN15_LIC_7	105	22	20.900	
MDNA/UN7_LIB_1	38	8	21.000	bom
MDNA/UN14_LIB_5	76	16	21.000	

MDNA/UN17_LIA_2	76	16	21.000	
MDNA/UN7_LIA	85	18	21.100	
MDNA/UN13_LIABC	104	22	21.100	
MDNA/UN8_LIB_4	174	37	21.200	
MDNA/UN14_LIC_3	14	3	21.400	
MDNA/UN18_LIABC_1	126	27	21.400	
MDNA/UN20_LIB	28	6	21.400	
MDNA/REV_5_2	14	3	21.400	
MDNA/REV_5_5	14	3	21.400	
MDNA/UN9_LIC_2	102	22	21.500	
MDNA/UN16_LIA_1	79	17	21.500	
MDNA/UN9_LIB_4	69	15	21.700	
MDNA/UN5_LIA_2	55	12	21.800	
MDNA/UN17_LIA	55	12	21.800	
MDNA/UN4_LIC	73	16	21.900	
MDNA/UN12_LIABC_1	105	23	21.900	
MDNA/UN17_LIA_4	59	13	22.000	
MDNA/UN16_LIABC	280	62	22.100	
MDNA/UN15_LIA_8	18	4	22.200	
MDNA/REV_4_10	36	8	22.200	
MDNA/UN18_LIA_1	67	15	22.300	
MDNA/UN12_LIB_3	58	13	22.400	
MDA/UN11LIABC_L	151	34	22.500	
MDNA/UN13_LIA_4	119	27	22.600	
MDNA/UN15_LIA_10	150	34	22.600	
MDNA/REV_3_9	22	5	22.700	
MDNA/UN14_LIB	22	5	22.700	
MDNA/UN16_LIA_2	101	23	22.700	
MDNA/UN16_LIB_4	101	23	22.700	
MDNA/UN18_LIABC_2	101	23	22.700	
MDNA/REV_3_5	35	8	22.800	
MDNA/UN1_LIB_2	48	11	22.900	
MDNA/UN14_LIC_6	87	20	22.900	
MDNA/UN2_LIB_3	13	3	23.000	
MDNA/REV_1	178	41	23.000	
MDNA/UN7_LIB_4	39	9	23.000	
MDNA/REV_2_9	13	3	23.000	
MDNA/UN13_LIB_4	126	29	23.000	
MDNA/REV_5_10	26	6	23.000	
MDNA/REV_5_12	39	9	23.000	
MDNA/UN13_LIC	108	25	23.100	
MDNA/UN5_LIC_8	43	10	23.200	
MDNA/UN11_LIABC	77	18	23.300	
MDNA/UN13_LIB_1	77	18	23.300	
MDNA/UN7_LIB_8	64	15	23.400	
MDNA/UN16_LIC_3	98	23	23.400	
MDNA/UN15_LIC_2	34	8	23.500	

MDNA/UN17_LIC_3	119	28	23.500	
MDNA/UN4_LIABC_1	72	17	23.600	
MDNA/UN8_LIABC_1	76	18	23.600	
MDNA/UN12_LIC	38	9	23.600	
MDNA/UN18_LIC_11	159	38	23.800	
MDNA/UN19_LIB	88	21	23.800	
MDNA/UN19_LIABC_2	193	46	23.800	
MDNA/UN6_LIC_3	25	6	24.000	
MDNA/REV_4	25	6	24.000	
MDNA/UN17_LIA_1	25	6	24.000	
MDNA/UN17_LIB_2	25	6	24.000	
MDNA/UN19_LIA_2	25	6	24.000	
MDNA/UN14_LIC_7	58	14	24.100	
MDNA/UN12_LIABC	107	26	24.200	
MDNA/UN15_LIC_6	86	21	24.400	
MDNA/REV_4_4	45	11	24.400	
MDNA/UN11_LIB_5	53	13	24.500	
MDNA/UN12_LIC_3	110	27	24.500	
MDNA/UN15_LIC_5	53	13	24.500	
MDNA/UN9_LIABC_1	81	20	24.600	
MDNA/UN13_LIABC_2	77	19	24.600	
MDA/UN15LIABC_L	353	87	24.600	
MDNA/UN19_LIB_2	69	17	24.600	
MDNA/UN5_LIABC_4	92	23	25.000	
MDNA/UN7_LIA_4	8	2	25.000	
MDNA/UN7_LIABC_1	48	12	25.000	
MDNA/UN8_LIC	96	24	25.000	
MDNA/UN12_LIA_5	128	32	25.000	
MDNA/UN15_LIA_4	4	1	25.000	
MDNA/UN18_LIC_12	76	19	25.000	
MDNA/REV_5_9	20	5	25.000	
MDNA/UN11_LIABC_1	87	22	25.200	
MDNA/UN16_LIB_3	111	28	25.200	
MDA/UN17LIB_V	186	47	25.200	
MDA/UN18LIABC_L	277	70	25.200	
MDNA/UN20_LIA_3	145	37	25.500	
MDNA/UN3_LIABC_1	39	10	25.600	
MDNA/UN19_LIC_3	74	19	25.600	
MDNA/UN16_LIC	35	9	25.700	
MDNA/UN7_LIB_2	62	16	25.800	
MDNA/REV_3_7	31	8	25.800	
MDNA/UN19_LIABC_1	174	45	25.800	
MDNA/UN18_LIB_3	27	7	25.900	
MDNA/UN10_LIA_4	23	6	26.000	
MDNA/UN17_LIB_8	92	24	26.000	
MDNA/UN19_LIB_4	73	19	26.000	
MDNA/REV_5_11	23	6	26.000	

MDNA/UN11_LIC_2	168	44	26.100	
MDA/UN16LIABC_L	333	87	26.100	
MDNA/UN11_LIB_2	38	10	26.300	
MDNA/UN18_LIC_3	19	5	26.300	
MDNA/UN6_LIC_1	34	9	26.400	
MDNA/UN19_LIB_3	178	47	26.400	
MDNA/PR_3	49	13	26.500	
MDNA/UN14_LIC_1	56	15	26.700	
MDNA/UN18_LIC_5	26	7	26.900	
MDNA/UN18_LIABC	173	47	27.100	
MDNA/UN5_LIC	22	6	27.200	
MDNA/UN5_LIC_3	22	6	27.200	
MDNA/UN10_LIC	66	18	27.200	
MDNA/UN14_LIB_1	33	9	27.200	
MDNA/UN14_LIC_5	22	6	27.200	
MDNA/UN15_LIC_3	22	6	27.200	
MDNA/UN10_LIABC	95	26	27.300	
MDNA/UN16_LIA	95	26	27.300	
MDNA/UN4_LIB	120	33	27.500	
MDNA/UN13_LIC_1	98	27	27.500	
MDNA/UN18_LIA_2	116	32	27.500	
MDNA/UN17_LIC_2	65	18	27.600	
MDA/UN3LIC_P	18	5	27.700	
MDNA/UN5_LIC_1	54	15	27.700	
MDNA/UN5_LIC_4	54	15	27.700	
MDNA/UN12_LIB_1	18	5	27.700	
MDNA/UN16_LIB	18	5	27.700	
MDNA/UN17_LIABC_1	108	30	27.700	
MDNA/REV_3_15	118	33	27.900	
MDNA/REV_4_1	118	33	27.900	
MDNA/UN2_LIA_1	25	7	28.000	
MDNA/UN19_LIA_4	100	28	28.000	
MDNA/UN6_LIA_2	64	18	28.100	
MDNA/UN13_LIB_3	217	61	28.100	
MDNA/UN4_LIA_5	39	11	28.200	
MDNA/UN14_LIA_5	53	15	28.300	
MDNA/UN15_LIABC_1	60	17	28.300	
MDNA/UN17_LIC_4	53	15	28.300	
MDNA/UN18_LIB_5	148	42	28.300	
MDA/UN12LIABC_L	285	81	28.400	
MDNA/UN7_LIA_5	7	2	28.500	
MDNA/UN20_LIB_2	70	20	28.500	
MDNA/UN20_LIC_9	21	6	28.500	
MDNA/UN13_LIC_3	129	37	28.600	
MDA/UN19LIABC_L	248	71	28.600	
MDNA/UN4_LIABC	93	27	29.000	
MDNA/UN5_LIABC_2	31	9	29.000	

MDNA/UN16_LIC_2	48	14	29.100	
MDNA/UN6_LIA_3	99	29	29.200	
MDNA/UN12_LIB_2	41	12	29.200	
MDNA/UN19_LIABC_3	99	29	29.200	
MDNA/UN5_LIB	17	5	29.400	
MDNA/UN5_LIB_4	34	10	29.400	
MDNA/UN14_LIB_3	34	10	29.400	
MDNA/UN19_LIC_5	71	21	29.500	
MDNA/UN6_LIB	74	22	29.700	
MDNA/UN2_LIB_1	10	3	30.000	
MDNA/UN5_LIB_5	50	15	30.000	
MDNA/UN10_LIA_8	10	3	30.000	
MDNA/UN10_LIABC_2	60	18	30.000	
MDNA/UN11_LIB	20	6	30.000	
MDNA/UN16_LIC_1	10	3	30.000	
MDNA/UN18_LIC_9	10	3	30.000	
MDNA/REV_5_3	10	3	30.000	
MDNA/UN12_LIC_4	176	53	30.100	
MDNA/UN1_LIABC_3	43	13	30.200	
MDNA/UN4_LIC_3	86	26	30.200	
MDNA/UN7_LIB	76	23	30.200	
MDNA/REV_5_8	33	10	30.300	
MDNA/UN7_LIB_3	69	21	30.400	
MDNA/UN15_LIA_1	36	11	30.500	
MDNA/UN3_LIA_3	26	8	30.700	
MDNA/UN2_LIABC_2	45	14	31.100	alto
MDNA/UN10_LIB_5	90	28	31.100	
MDNA/UN11_LIA_3	48	15	31.200	
MDNA/UN17_LIB_7	16	5	31.200	
MDNA/UN18_LIC_8	16	5	31.200	
MDNA/UN20_LIA_1	118	37	31.300	
MDNA/UN3_LIB_1	38	12	31.500	
MDNA/UN8_LIA	38	12	31.500	
MDNA/UN11_LIA_1	57	18	31.500	
MDNA/UN10_LIA_5	41	13	31.700	
MDNA/UN12_LIC_2	69	22	31.800	
MDNA/UN20_LIB_3	66	21	31.800	
MDNA/UN17_LIB_9	47	15	31.900	
MDNA/UN10_LIC_2	28	9	32.100	
MDNA/UN7_LIB_9	34	11	32.300	
MDNA/UN17_LIB_4	92	30	32.600	
MDNA/UN7_LIA_2	113	37	32.700	
MDA/UN7LIABC_L	155	51	32.900	
MDNA/UN10_LIC_1	109	36	33.000	
MDNA/UN1_LIA_2	9	3	33.300	
MDNA/UN2_LIB_5	21	7	33.300	
MDNA/UN3_LIA_2	15	5	33.300	

MDNA/UN3_LIC_1	51	17	33.300	
MDNA/UN4_LIA_6	15	5	33.300	
MDNA/UN5_LIC_6	54	18	33.300	
MDNA/UN5_LIABC_3	27	9	33.300	
MDNA/UN7_LIA_3	9	3	33.300	
MDNA/UN8_LIA_1	15	5	33.300	
MDNA/UN8_LIA_5	30	10	33.300	
MDNA/REV_2_2	3	1	33.300	
MDNA/UN9_LIA_1	63	21	33.300	
MDNA/UN12_LIABC_3	27	9	33.300	
MDNA/UN17_LIABC	117	39	33.300	
MDA/UN17LIABC_L	192	64	33.300	
MDNA/UN10_LIABC_1	77	26	33.700	
MDNA/UN7_LIC	53	18	33.900	
MDNA/UN15_LIABC	79	27	34.100	
MDNA/UN12_LIB_4	38	13	34.200	
MDNA/UN9_LIABC_2	32	11	34.300	
MDNA/UN10_LIA_11	64	22	34.300	
MDNA/UN11_LIB_1	32	11	34.300	
MDNA/UN1_LIABC	26	9	34.600	
MDNA/UN11_LIC_1	52	18	34.600	
MDNA/UN10_LIB	23	8	34.700	
MDNA/UN13_LIA_2	63	22	34.900	
MDNA/UN6_LIC	40	14	35.000	
MDNA/UN6_LIC_5	57	20	35.000	
MDNA/UN10_LIB_4	97	34	35.000	
MDNA/UN14_LIABC_1	60	21	35.000	
MDNA/UN1_LIABC_2	17	6	35.200	
MDNA/UN1_LIC	31	11	35.400	
MDNA/UN4_LIA_7	107	38	35.500	
MDNA/UN20_LIB_1	138	49	35.500	
MDNA/UN1_LIB_1	28	10	35.700	
MDNA/REV_2_5	14	5	35.700	
MDNA/UN15_LIC_1	47	17	36.100	
MDNA/UN1_LIA_1	11	4	36.300	
MDNA/REV_4_12	41	15	36.500	
MDNA/UN17_LIA_3	112	41	36.600	
MDNA/UN18_LIB_6	81	30	37.000	
MDNA/UN10_LIA_12	113	42	37.100	
MDNA/UN11_LIA_2	78	29	37.100	
MDNA/UN5_LIB_3	24	9	37.500	
MDNA/UN1_LIA_3	13	5	38.400	
MDNA/UN8_LIA_2	13	5	38.400	
MDNA/UN18_LIC_4	65	25	38.400	
MDNA/UN18_LIC_6	39	15	38.400	
MDNA/UN14_LIB_6	57	22	38.500	
MDNA/UN17_LIB_3	31	12	38.700	

MDNA/UN5_LIABC_1	51	20	39.200	
MDNA/UN1_LIABC_1	15	6	40.000	
MDNA/UN10_LIA_9	15	6	40.000	
MDNA/REV_3_10	15	6	40.000	
MDNA/REV_4_11	10	4	40.000	
MDNA/UN18_LIC_1	15	6	40.000	
MDNA/UN3_LIABC_2	22	9	40.900	
MDNA/REV_3_11	22	9	40.900	
MDNA/UN3_LIA	17	7	41.100	
MDNA/REV_2_1	17	7	41.100	
MDNA/UN6_LIABC	113	47	41.500	
MDNA/UN5_LIB_2	12	5	41.600	
MDNA/UN7_LIB_6	24	10	41.600	
MDNA/UN7_LIA_7	79	33	41.700	
MDNA/UN9_LIB	38	16	42.100	
MDNA/UN6_LIB_2	28	12	42.800	
MDNA/UN7_LIA_1	14	6	42.800	
MDNA/REV_3_1	14	6	42.800	
MDNA/REV_3_4	21	9	42.800	
MDNA/REV_3_8	14	6	42.800	
MDNA/UN16_LIB_1	30	13	43.300	
MDNA/UN8_LIB_3	39	17	43.500	
MDA/UN3LIABC_L	61	27	44.200	
MDNA/UN8_LIABC	52	23	44.200	
MDNA/REV_5_13	36	16	44.400	
MDNA/UN5_LIA	38	17	44.700	
MDNA/UN20_LIB_8	42	19	45.200	
MDNA/UN4_LIA_1	11	5	45.400	
MDNA/REV_2_3	33	15	45.400	
MDNA/UN2_LIC_2	37	17	45.900	
MDNA/UN4_LIB_1	26	12	46.100	
MDNA/REV_3_12	28	13	46.400	
MDNA/UN2_LIA_5	15	7	46.600	
MDNA/UN14_LIC_4	32	15	46.800	
MDNA/UN2_LIA_7	17	8	47.000	
MDNA/UN2_LIC_1	17	8	47.000	
MDNA/UN14_LIB_4	21	10	47.600	
MDNA/UN2_LIABC_3	23	11	47.800	
MDNA/UN6_LIC_4	71	34	47.800	
MDNA/UN6_LIC_2	39	19	48.700	
MDNA/UN2_LIA_2	4	2	50.000	
MDNA/UN2_LIA_4	10	5	50.000	
MDNA/REV_2_10	4	2	50.000	
MDNA/UN14_LIA_1	6	3	50.000	
MDNA/UN3_LIA_1	35	18	51.400	
MDNA/REV_3_2	23	12	52.100	
MDNA/UN10_LIB_2	19	10	52.600	

MDNA/UN5_LIC_7	17	9	52.900	
MDNA/UN10_LIA_7	15	8	53.300	
MDNA/UN10_LIA_10	13	7	53.800	
MDNA/UN5_LIA_1	33	18	54.500	
MDNA/UN7_LIA_6	8	5	62.500	
MDNA/UN18_LIB_2	11	7	63.600	
MDNA/UN2_LIA	10	7	70.000	

B. ANEXOS

Anexo 1

Capa do livro *Muito prazer* – fale o português do Brasil (2008)

Anexo 2

Sumário do livro *Muito prazer* – fale o português do Brasil (2008)

Unidade 1

MUITO PRAZER

LIÇÃO A

GRAMÁTICA Pronomes pessoais e verbo ser

AMPLIAÇÃO DO VOCABULÁRIO O alfabeto

PANORAMA Cumprimentos

LIÇÃO B

GRAMÁTICA Artigo Definido e pronome adjetivo possessivo; pronome pessoal

AMPLIAÇÃO DO VOCABULÁRIO Nacionalidade e profissões

PANORAMA Despedidas

LIÇÃO C

AMPLIAÇÃO DO VOCABULÁRIO Nacionalidade e profissões

LIÇÕES A, B e C

LEITURA E REDAÇÃO Mensagens Instantâneas

CONSOLIDAÇÃO LEXICAL Expressões

Unidade 2

Este é o meu amigo Paulo

LIÇÃO A

PANORAMA Apresentações

GRAMÁTICA Pronomes demonstrativos

AMPLIAÇÃO DO VOCABULÁRIO Expressões e Inversão

LIÇÃO B

PANORAMA Números I

GRAMÁTICA Verbos: precisar e ligar (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Verbos

LIÇÃO C

PANORAMA Respostas a respeito de pessoas

GRAMÁTICA Pronomes demonstrativos e verbo morar + preposição EM (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Relacionamentos

LIÇÕES A, B e C

LEITURA E REDAÇÃO Recados

CONSOLIDAÇÃO LEXICAL Verbos

Unidade 3

Quantos anos ele tem?

LIÇÃO A

PANORAMA Números II

GRAMÁTICA Pronomes possessivos e verbo ter (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Relacionamentos

LIÇÃO B

PANORAMA Números III e documentos

GRAMÁTICA Verbo poder (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Documentos oficiais e extra-oficiais

LIÇÃO C

PANORAMA Fuso horário

GRAMÁTICA Verbo: querer + preposição de (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Horas I

LIÇÕES A, B e C

LEITURA E REDAÇÃO Hora de Verão

CONSOLIDAÇÃO LEXICAL Horário de Atividades

Unidade 4

Táxi!

LIÇÃO A

PANORAMA Dinheiro

GRAMÁTICA Futuro simples e ir + verbo

AMPLIAÇÃO DO VOCABULÁRIO Dinheiro

LIÇÃO B

PANORAMA Tipos de restaurante

GRAMÁTICA ‘Mas’ e ‘nem ... nem’

AMPLIAÇÃO DO VOCABULÁRIO Menu I

LIÇÃO C

PANORAMA Na praça de alimentação

GRAMÁTICA Estar com + substantivo/ Estar com vontade de + verbo/ Enquanto... (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Menu II

LIÇÕES A, B e C

LEITURA E REDAÇÃO Restaurantes no RJ – Naturais

CONSOLIDAÇÃO LEXICAL Comidas e Bebidas

Pronúncia do português – parte 1

Revisão das unidades 1 a 4

Unidade 5

Que semana, hein?

LIÇÃO A

PANORAMA Horários

GRAMÁTICA Verbos: querer; estar e achar (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Adjetivos

LIÇÃO B

PANORAMA Dias da semana e meses do ano

GRAMÁTICA Verbos: estudar e trabalhar (presente do indicativo); locuções adverbiais de tempo

AMPLIAÇÃO DO VOCABULÁRIO Horas II

LIÇÃO C

PANORAMA Procurando algo

GRAMÁTICA Pronomes e advérbios interrogativos

AMPLIAÇÃO DO VOCABULÁRIO Revisão e ampliação de verbos

LIÇÕES A, B e C

LEITURA E REDAÇÃO Minha rotina

CONSOLIDAÇÃO LEXICAL Verbos

Unidade 6

Vamos pro cinema, Ana?

LIÇÃO A

PANORAMA Entretenimentos no fim de semana

GRAMÁTICA Verbos: ir, ter que e poder (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Outros entretenimentos

LIÇÃO B

PANORAMA Família

GRAMÁTICA Verbos: ir, achar e gostar (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Família

LIÇÃO C

PANORAMA Rotina: a família de Ronaldo Gomes

GRAMÁTICA Verbo estar + verbo – “r” + NDO; verbo saber

AMPLIAÇÃO DO VOCABULÁRIO Meios de transporte e verbos de locomoção

LIÇÕES A, B e C

LEITURA E REDAÇÃO Terminal Rodoviário Tietê

CONSOLIDAÇÃO LEXICAL Árvore Genealógica

Unidade 7

Atrasada de novo, Valquíria?

LIÇÃO A

PANORAMA A locomoção na cidade de São Paulo

GRAMÁTICA Verbo: ser, estar e vir (presente do indicativo); Sugestão: Por que... não...?

AMPLIAÇÃO DO VOCABULÁRIO Verbos

LIÇÃO B

PANORAMA Aonde ir no seu bairro

GRAMÁTICA Verbo ter (= existir)

AMPLIAÇÃO DO VOCABULÁRIO O que fazer no seu bairro

LIÇÃO C

PANORAMA O que fazer no fim de semana

GRAMÁTICA Verbos: saber, conhecer, preferir (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Na academia

LIÇÕES A, B e C

LEITURA E REDAÇÃO Parques de São Paulo

CONSOLIDAÇÃO LEXICAL Características

Unidade 8

Eu gostaria de ver um apartamento para comprar

LIÇÃO A

PANORAMA Planta de imóveis

GRAMÁTICA Verbos: gostar (futuro do pretérito), ver (Imperativo), ficar (=localizações); expressão: dar para

AMPLIAÇÃO DO VOCABULÁRIO Preposições e locuções prepositivas

LIÇÃO B

PANORAMA O seu bairro

GRAMÁTICA Pronomes indefinidos: algum, muito, nenhum

AMPLIAÇÃO DO VOCABULÁRIO Tipos de imóveis

LIÇÃO C

PANORAMA Mobília

GRAMÁTICA Adjetivos: comparativo

AMPLIAÇÃO DO VOCABULÁRIO Partes da casa

LIÇÕES A, B e C

LEITURA E REDAÇÃO Classificados

CONSOLIDAÇÃO LEXICAL Móveis e Imóveis

Pronúncia do português – parte 2

Revisão das unidades 5 a 8

Unidade 9

A gente faz ginástica na mesma academia

LIÇÃO A

PANORAMA Descrição física I

GRAMÁTICA Verbos: olhar, ser, ter, gostar e fazer (pretérito imperfeito do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Outros termos para descrição física

LIÇÃO B

PANORAMA Descrição física II

GRAMÁTICA Verbos: correr, ir e fazer (tempo)

AMPLIAÇÃO DO VOCABULÁRIO Cores I; Vestuário I

LIÇÃO C

PANORAMA Descrição de personalidade

GRAMÁTICA Verbos: lembrar-se e parecer (presente do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Cores II; Vestuário II

LIÇÕES A, B e C

LEITURA E REDAÇÃO O significado das cores

CONSOLIDAÇÃO LEXICAL Cores

Unidade 10

Estou com gripe

LIÇÃO A

PANORAMA A saúde e os remédios

GRAMÁTICA Verbos: sarar e tomar (pretérito perfeito do indicativo); pronome indefinido

AMPLIAÇÃO DO VOCABULÁRIO Corpo Humano I

LIÇÃO B

PANORAMA Tipos de tratamento médico

GRAMÁTICA Verbo ser; Estrutura com verbo ser + sujeito + que; advérbio de freqüência

AMPLIAÇÃO DO VOCABULÁRIO Alguns sintomas

LIÇÃO C

PANORAMA Descrição de condição física ou emocional

GRAMÁTICA Verbo ficar, estar e ter (pretérito perfeito do indicativo); advérbio de intensidade

AMPLIAÇÃO DO VOCABULÁRIO Estados emocionais e sentimentos: adjetivos e substantivos

LIÇÕES A, B e C

LEITURA E REDAÇÃO Cortando o mal pelas raízes

CONSOLIDAÇÃO LEXICAL O corpo humano

Unidade 11

Você é bom em História do Brasil?

LIÇÃO A

PANORAMA Um pouco de História

GRAMÁTICA Verbos: começar, permanecer, ser e ver (pretérito perfeito do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Algumas regências

LIÇÃO B

PANORAMA Festas Juninas

GRAMÁTICA Verbos: preparar, fazer, divertir-se e ser (Pretérito imperfeito do indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Ser ou estar + particípio passado (= adjetivo)

LIÇÃO C

PANORAMA Lendas

GRAMÁTICA Verbo fazer (pretérito perfeito do indicativo); pronome indefinido; Pretérito Perfeito x Pretérito Imperfeito

AMPLIAÇÃO DO VOCABULÁRIO Tipos de histórias

LIÇÕES A, B e C

LEITURA E REDAÇÃO Festas Juninas – Tradição e Comidas Típicas

CONSOLIDAÇÃO LEXICAL Colocações – regências verbais

Unidade 12

Estou a fim de uma moqueca

LIÇÃO A

PANORAMA Tipos de comida

GRAMÁTICA Verbo: andar + complemento; Mais-que-perfeito composto; diminutivo e aumentativo

AMPLIAÇÃO DO VOCABULÁRIO Divisão da forma de servir os pratos; tipos de comidas e de restaurantes

LIÇÃO B

PANORAMA Costumes

GRAMÁTICA Perfeito x Imperfeito do indicativo; estar + '-ndo' (Imperfeito do Indicativo); superlativo

AMPLIAÇÃO DO VOCABULÁRIO Costumes brasileiros e expressões

LIÇÃO C

PANORAMA Convites

GRAMÁTICA Verbos: estar (Pretérito perfeito do indicativo) e dizer (Presente do Indicativo)

AMPLIAÇÃO DO VOCABULÁRIO Lugares e atividades

LIÇÕES A, B e C

LEITURA E REDAÇÃO Made in Brazil para o Japão

CONSOLIDAÇÃO LEXICAL Formas no cardápio de servir comidas e bebidas

pronúncia do português – parte 3

revisão das unidades 9 a12

Unidade 13

Estou fazendo planos para viajar

LIÇÃO A

PANORAMA Planos

GRAMÁTICA Conjunção coordenada conclusiva e derivação

AMPLIAÇÃO DO VOCABULÁRIO Prefixos e sufixos

LIÇÃO B

PANORAMA Passado, presente, futuro: transportes e comunicações

GRAMÁTICA Pretérito perfeito composto e superlativo

AMPLIAÇÃO DO VOCABULÁRIO Objetos

LIÇÃO C

PANORAMA Os tempos atuais

GRAMÁTICA Futuro do Pretérito e pretérito mais-que-perfeito composto

AMPLIAÇÃO DO VOCABULÁRIO Atividades/Ensino no Brasil

LIÇÕES A, B e C

LEITURA E REDAÇÃO Destaques

CONSOLIDAÇÃO LEXICAL Tipos de filmes

Unidade 14

Alô? Quem fala?

LIÇÃO A

PANORAMA Comunicação

GRAMÁTICA Pronomes Pessoais I

AMPLIAÇÃO DO VOCABULÁRIO Nível de formalidade – recados

LIÇÃO B

PANORAMA Telefonemas para empresas

GRAMÁTICA Pronomes Pessoais II

AMPLIAÇÃO DO VOCABULÁRIO Expressões usadas ao telefone I

LIÇÃO C

PANORAMA Ao telefone

GRAMÁTICA Pronomes Pessoais III

AMPLIAÇÃO DO VOCABULÁRIO Expressões usadas ao telefone II

LIÇÕES A, B e C

LEITURA E REDAÇÃO Mulheres passam cinco anos ao telefone, diz estudo

CONSOLIDAÇÃO LEXICAL Ao telefone

Unidade 15

Quer ir ao cinema comigo na quinta?

LIÇÃO A

PANORAMA Bate-papo

GRAMÁTICA Discurso direto e indireto: perguntas e declarações

AMPLIAÇÃO DO VOCABULÁRIO Comunicação via computador

LIÇÃO B

PANORAMA Recados

GRAMÁTICA Discurso direto e indireto: ordens e declarações

AMPLIAÇÃO DO VOCABULÁRIO Comunicação escrita ou oral

LIÇÃO C

PANORAMA Eventos

GRAMÁTICA Posição dos pronomes que atuam como objetos

AMPLIAÇÃO DO VOCABULÁRIO Expressões com partes do corpo

LIÇÕES A, B e C

LEITURA E REDAÇÃO Aulas de inglês já migram para a Web

CONSOLIDAÇÃO LEXICAL Comunicação escrita e oral

Unidade 16

Imagine fazer uma viagem de bicicleta!

LIÇÃO A

PANORAMA Viagem

GRAMÁTICA Regência Verbal e Nominal

AMPLIAÇÃO DO VOCABULÁRIO Tipos de viagem e lugares para hospedagem

LIÇÃO B

PANORAMA Reservas

GRAMÁTICA Futuro do Subjuntivo

AMPLIAÇÃO DO VOCABULÁRIO Hotel

LIÇÃO C

PANORAMA Lembranças ou souvenires

GRAMÁTICA Verbos e Expressões

AMPLIAÇÃO DO VOCABULÁRIO Tipos de lembranças

LIÇÕES A, B e C

LEITURA E REDAÇÃO Abrolhos – BA

CONSOLIDAÇÃO LEXICAL Regência Verbal e Nominal

pronúncia do português – parte 4

revisão das unidades 13 a 16

Unidade 17

Os patins foram inventados por um belga em 1760

LIÇÃO A

PANORAMA Invenções

GRAMÁTICA Voz passiva I – tempos simples

AMPLIAÇÃO DO VOCABULÁRIO Mais invenções

LIÇÃO B

PANORAMA Máquinas

GRAMÁTICA Voz passiva II – tempos compostos

AMPLIAÇÃO DO VOCABULÁRIO Eletroeletrônicos

LIÇÃO C

PANORAMA Consertos

GRAMÁTICA Futuro do subjuntivo – Verbos irregulares

AMPLIAÇÃO DO VOCABULÁRIO Carros

LIÇÕES A, B e C

LEITURA E REDAÇÃO Novas regras para renovação de CNH

CONSOLIDAÇÃO LEXICAL Carros

Unidade 18

Vou para outro setor na nova empresa

LIÇÃO A

PANORAMA Emprego

GRAMÁTICA Presente do Subjuntivo I – verbos regulares

AMPLIAÇÃO DO VOCABULÁRIO O mercado de trabalho e o futuro

LIÇÃO B

PANORAMA Fenômenos da natureza

GRAMÁTICA Presente do Subjuntivo II – verbos irregulares A

AMPLIAÇÃO DO VOCABULÁRIO Natureza

LIÇÃO C

PANORAMA Impostos

GRAMÁTICA Presente do subjuntivo II – verbos irregulares B

AMPLIAÇÃO DO VOCABULÁRIO Impostos

LIÇÕES A, B e C

LEITURA E REDAÇÃO Profissão do futuro

CONSOLIDAÇÃO LEXICAL Impostos e taxas

Unidade 19

Se eu fosse você compraria um jornal para procurar emprego

LIÇÃO A

PANORAMA Jogos de azar

GRAMÁTICA Imperfeito do Subjuntivo com futuro do Pretérito

AMPLIAÇÃO DO VOCABULÁRIO Jogos de Azar

LIÇÃO B

PANORAMA Casamento

GRAMÁTICA Imperfeito do Subjuntivo com expressões

AMPLIAÇÃO DO VOCABULÁRIO Casamento

LIÇÃO C

PANORAMA Vocação profissional

GRAMÁTICA Pretérito Perfeito do Subjuntivo

AMPLIAÇÃO DO VOCABULÁRIO Escola

LIÇÕES A, B e C

LEITURA E REDAÇÃO Objetividade no currículo é a senha para entrevista

CONSOLIDAÇÃO LEXICAL Jogos, casamento, escola

Unidade 20

O que você teria feito diferente na sua vida?

LIÇÃO A

PANORAMA Balanço do ano

GRAMÁTICA Futuro do Pretérito Composto

AMPLIAÇÃO DO VOCABULÁRIO Fim do ano

LIÇÃO B

PANORAMA Arrependimento

GRAMÁTICA Pretérito mais-que-perfeito do Subjuntivo

AMPLIAÇÃO DO VOCABULÁRIO Pensamentos sobre erros; Arrependimentos

LIÇÃO C

PANORAMA Conselhos

GRAMÁTICA Imperfeito do Subjuntivo + Futuro do pretérito composto

AMPLIAÇÃO DO VOCABULÁRIO Conselhos

LIÇÕES A, B e C

LEITURA E REDAÇÃO Sucesso profissional: suas metas para o ano que vem (e os anos seguintes...)

CONSOLIDAÇÃO LEXICAL Planos

pronúncia do português – parte 5

revisão das unidades 17 a 20

Anexo 3

MATERIAIS DIDÁTICOS BRASILEIROS DE ENSINO DE PORTUGUÊS COMO LÍNGUA ESTRANGEIRA/SEGUNDA LÍNGUA[1]

Ano de publicação	Título	Autor(es)13	Editora	ISBN	Compo-nentes[2]	Público-alvo/ nível[3]
1954 (1ª edição)	*Português para estrangeiros* (v. 1 e 2)	MARCHANT, Mercedes	Porto Alegre: Age	8585627212	LA	"Estrangeiros de qualquer nacionalidade."
1969 (1ª edição)	*Português*: conversação e gramática	MAGRO, Haydée S; PAULA, Paulo de	São Paulo: Pioneira/Brazilian American Cultural Institute	8522101094	LA, K7	Básico e intermediário
1976 (2ª edição)	*Português básico para estrangeiros*	MONTEIRO, Sylvio	São Paulo: Ibrasa	8534801169	LA, K7 (1)	Nível básico
1981 (1ª edição)	*Falando, lendo, escrevendo português*: um curso para estrangeiros[4]	LIMA, Emma Eberlein Oliveira Fernandes; IUNES, Samira Abirad	São Paulo: EPU	85-12-54010-9	LA, LE, LP, LR, LT, G (al, fr, ing), CD/K7 para LA (3), CD/K7 para LE (4)	"Adultos e adolescentes a partir dos 13 anos, de qualquer nacionalidade. Leva o aluno totalmente principiante até o nível intermediário."

1. Tabela adaptada do artigo de Diniz (2007*):* Mudanças discursivas em livros didáticos brasileiros de ensino de Português como Língua Estrangeira. *Portuguese Language Journal, v. 2.*
2. Foram adotadas as seguintes siglas: LA (livro do aluno), LE (livro de exercícios), LP (livro do professor), LR (livro de respostas), LT (livro de testes), G (glossário), al (alemão), esp (espanhol), fr (francês), ing (inglês). Os números entre parênteses indicam a quantidade de CDs ou fitas K7 que fazem parte da coleção.
3. As informações que constam nesta coluna foram retiradas dos prefácios e/ou quarta-capas dos livros do aluno.
4. Posteriormente editado com o título de *Falar... ler... escrever... português*: um curso para estrangeiros (ISBN: 85-12-54310-8).

1984 (1ª edição)	*Tudo bem?* Português para a nova geração (v. 1 e 2)	PONCE, Maria Harumi Otuki; BURIM, Silvia; FLORISSI, Susanna	São Paulo: SBS	8587343270 (v.1) 858734384X (v.2)	LA, CD (2)	"Voltado às necessidades do público jovem."
1986	*Avenida Brasil*: curso básico de português para estrangeiros (v. 1 e 2)	LIMA, Emma Eberlein Oliveira Fernandes; ROHRMANN, Lutz; ISHIHARA, Tokiko; BERGWEILER, Cristián González; IUNES, Samira Abirad	São Paulo: EPU	85-12-54700-6 (v. 1) 85-12-54750-2 (v. 2)	LA, LE, LP, CD/K7 (2), G (al, esp, ing, fr)	"Destina-se a estrangeiros de qualquer nacionalidade, adolescentes e adultos que queiram aprender Português para poder comunicar-se com brasileiros e participar de sua vida cotidiana."
1989 (1ª edição)	*Fala Brasil*	COUDRY, Pierre; FONTÃO DO PATROCÍNIO, Elizabeth	Campinas: Pontes	85-7113-082-5	LA, LE, CD/K7 (2)	"Falantes de qualquer idioma."
1990 (1ª edição)	*Português como segunda língua*	ALMEIDA, Marilú Miranda Montenegro e; GUIMARÃES, Lucia Angelina Cid Loureiro	Rio de Janeiro: Ao Livro Técnico	85-215-0534-5	LA	"O livro tem como objetivo suprir as necessidades encontradas no estudo do Português". "Destina-se a alunos que já tenham noções da língua."
1990 (1ª edição)	*Português via Brasil*. um curso avançado para estrangeiros	LIMA, Emma Eberlein Oliveira Fernandes; IUNES, Samira Abirad	São Paulo: EPU	85-12-54380-2	LP, LA, K7	"Pessoas que tenham terminado o curso básico de Português como língua estrangeira e desejam prosseguir seus

						estudos em nível intermediário e avançado.”
1991 (1ª edição)	*Aprendendo português do Brasil*: um curso para estrangeiros	LAROCA, Maria Nazaré de Carvalho; BARA, Nadime; PEREIRA, Sonia Maria da Cunha	Campinas: Pontes	85-7113-065-5	LA, LE, LP, CD/K7 (1)	“O livro tem como objetivo dar condições ao aluno estrangeiro de dominar, em pouco tempo, as estruturas fundamentais da Língua Portuguesa, nas modalidades oral e escrita.”
1994	*Português para estrangeiros infanto-juvenil*. Português para estrangeiros nível avançado	MARCHANT, Mercedes	Porto Alegre: Age	8574970301 (infanto-juvenil); 858562728X (avançado)	LA, K7 para nível básico (1), K7 para nível avançado (1)	Níveis básico e avançado “Crianças e adolescentes cuja língua materna é o espanhol”.
1997	*Um Português bem brasileiro* (níveis 1 a 4)	Fundação Centro de Estudos Brasileiros (FUNCEB)	Buenos Aires: Loyola	987-96351-0-8 (nível 1) 987-96351-2-4 (nível 2) 987-96351-0-8 (nível 3) 987-96351-6-7 (nível 4)	LA	Hispano-falantes
1999 (1ª edição)	*Bem-vindo*: a língua portuguesa no mundo da comunicação	PONCE, Maria Harumi Otuki de; BURIM, Silvia R. B. Andrade; FLORISSI, Susanna	São Paulo: SBS	85-7583-063-5	LA, LE, LP, LR, CD/K7 (4)	“Público de jovens e adultos de qualquer nacionalidade que queira aprender

						português, com sotaque brasileiro, como língua estrangeira." Nível iniciante. até o pós-intermediário.
2000 (1ª edição)	*Conhecendo o Brasil* – curso de português para falantes de espanhol	Fundação Centro de Estudos Brasileiros (FUNCEB)	Buenos Aires: Akian	987-96351-5-9	Livro, K7 (2), vídeo (3)	"Preparado especialmente para falantes de espanhol." Nível básico.
2000 (1ª edição)	*Sempre amigos*: fala Brasil para jovens	FONTÃO DO PATROCÍNIO, Elizabeth	Campinas: Pontes	85-7113-140-8	LA, LP	"Público jovem." O último dos seis módulos do livro é dedicado, especificamente, a falantes de espanhol.
2001 (1ª edição)	*Interagindo em português*: textos e visões do Brasil (v. 1 e 2)	HENRIQUES, Eunice Ribeiro; GRANNIER, Daniele Marcelle	Brasília: Thesaurus	85-7062-254-6 (v. 1) 85-7062-253-8 (V. 2)	LA, K7	Iniciante (v. I) Intermediário (v. II) Avançado (v. III, no prelo).
2002	*Passagens* – português do Brasil para estrangeiros	CELLI, Rosine	Campinas: Pontes	8571131643	LA, LR, CD, CD-ROM	"Adolescentes e adultos."
2003	*Diálogo Brasil*: curso intensivo de português para estrangeiros	LIMA, Emma Eberlein Oliveira Fernandes; IUNES, Samira Abirad; LEITE, Marina Ribeiro	São Paulo: EPU	85-12-54220-9	LA, LP, CD/K7 (2), G (al, fr, ing, esp)	"Destinado a um público adulto, a profissionais de todas as áreas que necessitem de um aprendizado

						seguro e relativamente rápido, aplicando-se também a um público jovem." "Abrange o ensino da língua desde suas primeiras noções, chegando ao final do nível intermediário."
2005 (1ª edição)	*Estação Brasil*: português para estrangeiros	BIZON, Ana Cecília; FONTÃO DO PATROCÍNIO, Elizabeth	Campinas: Átomo	85-7670-015-8	LA, CD (1)	"Alunos aprendizes que já alcançaram uma proficiência média em PLE; alunos que desejam se preparar para o exame de proficiência Celpe-Bras."
2006 (1ª edição)	*Panorama Brasil*: ensino do português no mundo dos negócios	PONCE, Harumi de; BURIM, Silvia; FLORISSI, Susanna	São Paulo: Galpão	8599311042	LA, CD (2)	"Livro voltado para o mundo dos negócios. Ideal para alunos de nível intermediário e avançado, é uma importante ferramenta para educadores que trabalham com diretores, executivos e demais funcionários de empresas que vêm trabalhar no Brasil."

2008	*Muito prazer* – fale o Português do Brasil	Fernandes, Glaucia; Ferreira, Telma de Lurdes São Bento; Ramos, Vera Lúcia	São Paulo: Disal	978-85-7844-005-3	LA, CD	"O livro é um curso de português para estrangeiros que tem como objetivo capacitar o aluno, de qualquer nacionalidade, a aprender o Português falado no Brasil e a comunicar-se com precisão e fluência. Com abordagem nova, combina as melhores características das abordagens mais modernas de ensino de língua estrangeira."
2009	*Novo Avenida Brasil* 1 e 2. Curso Básico de Português para estrangeiros	LIMA, Emma Eberlein Oliveira Fernandes; ROHRMANN, Lutz; ISHIHARA, Tokiko; IUNES, Samira Abirad; BERGWEILER,Cristián González	São Paulo: EPU	978-85-12-54520-2 (v. 1) 978-85-12-54570-7 (v. 2)	LA, CD	"Destina-se a estrangeiros de qualquer nacionalidade, adolescentes e adultos, que queiram aprender Português para poderem comunicar-se com brasileiros e participar de sua vida cotidiana."