CAPÍTULO 12

VERIFICAÇÃO DE FREQUÊNCIA LEXICOLÓGICA PARA A CLASSIFICAÇÃO DE MATERIAL DIDÁTICO DE PORTUGUÊS LÍNGUA ADICIONAL

Maryelle Joelma Cordeiro
Carlos Antônio de Souza Perini

RESUMO: Neste trabalho apresentamos uma pesquisa em andamento que focaliza o estudo do léxico empregado nos livros didáticos de ensino de português como Língua Adicional (PLA). A pesquisa tem como objetivo contrastar a frequência do léxico de material didático de PLA com a frequência do léxico do português, apresentado na lista elaborada por Enzo Del Carratore e Jayme Laperuta Filho (Léxico de Frequência do Português falado na Cidade de São Paulo). Do ponto de vista lexicológico, como extensão dessa análise, pretende-se também produzir, a partir do léxico de frequência do Português e do léxico dos materiais didáticos selecionados, as listas do léxico para cada nível da Certificação de Proficiência em Língua Portuguesa (Celpe-Bras): Intermediário, Intermediário Superior, Avançado e Avançado Superior.
PALAVRAS-CHAVE: frequência lexicológica. ensino de PLA. Celpe-Bras. material didático.

ABSTRACT: In this paper we present an ongoing research that focuses on the study of the lexicon used in textbooks for teaching Portuguese as an Additional Language (PLA). The research aims to contrast the frequency of the lexicon from teaching textbooks PLA with the frequency of the spoken Portuguese lexicon, presented the list done by Enzo Del Carratore and Jayme Laperuta Filho (Léxico de Frequência do Português Falado na cidade de São Paulo). From the lexical point of view, as an extension of this analysis, we intend to produce, from the frequency of the Portuguese lexicon and the lexicon of selected teaching textbooks, the vocabulary list for each level of Proficiency Certification in Brazilian Portuguese (CELPE-Bras): Intermediate, Upper Intermediate, Advanced and Advanced Higher.
KEYWORDS: lexicological frequency. PLA teaching. Celpe-Bras. teaching textbooks.

1 I INTRODUÇÃO

Este artigo está dividido em quatro seções. Inicialmente são abordados os estudos da frequência do léxico do português brasileiro realizados por Biderman (1967). Posteriormente, apresentamos a lista de referência utilizada neste trabalho: Léxico de Frequência do Português falado na cidade de São Paulo, construído a partir de dados extraídos do Projeto NURC (Norma Urbana Culta), pelos pesquisadores Enzo Del Carratore e Jayme Laperuta Filho. Na terceira seção, apresentamos os níveis de competência do Celpe-Bras. Na quarta e última

seção, são apresentados os procedimentos metodológicos de preparação do material para a realização do contraste do léxico com as listas de frequência, utilizando-se softwares apropriados para este fim. Finalmente, nas considerações finais foi realizada a análise dos dados, com os gráficos que ilustram o contraste do léxico de cada livro didático com a lista de referência.

Acreditamos que o resultado das comparações pode auxiliar autores de livros didáticos no uso de textos autênticos para fins didáticos no ensino de PLA. Além disso, o material produzido poderia ser utilizado por professores de PLA, pois serve como suporte para a avaliação de textos de alunos, permitindo a classificação quanto à aprendizagem lexical e ao nível de competência linguística, de acordo com os níveis do Celpe-Bras.

2 I OS ESTUDOS DO LÉXICO DE FREQUÊNCIA DO PORTUGUÊS BRASILEIRO

Existem poucos estudos relacionados com o Léxico de Frequência do Português Brasileiro. Ao se realizar uma busca no portal de Periódicos da CAPES com as expressões "Léxico de Frequência" e "Português Brasileiro" não são retornados nenhum artigo ou quaisquer outros tipos de publicações ligados ao tema.

Podemos dizer que trata-se de um problema de caráter histórico, pois ainda na década de 60, Biderman (1969) já reconhecia essa deficiência nas pesquisas realizadas até então no Brasil:

> "Tanto quanto conheço os trabalhos lingüísticos realizados no Brasil (e para o português em geral), não sei de pesquisas sistemáticas realizadas nesse setor quer por linguistas nossos, quer aplicações a nossa língua. Conheço apenas alguns estudos esparsos dedicados a problemas específicos em português." (BIDERMAN (1967, p. 117)

A autora defende ainda que "[...] seria desejável que os estudantes inclinados aos estudos linguísticos tivessem uma formação estatística elementar". Uma fase posterior a essa consiste em institucionalizar um espaço comum para as pesquisas de dados linguísticos conciliados com elementos estatísticos.

Biderman (1967) reconhece a subjetividade dos trabalhos dos linguistas como um dificultador para os estatísticos como a repulsa dos linguistas para a forma com que os matemáticos utilizam a língua:

> [...] dificilmente coincidem os linguistas quanto à definição do vocabulário e mais ainda divergirão eles quando tiverem que decidir sobre as unidades léxicas em uma compilação vocabular. Se passarmos ao nível morfêmico e sintático, as divergências serão ainda maiores. Ora, a estatística precisa partir de critérios seguros e bem estabelecidos para proceder à compilação de suas amostras. [...] Se esses critérios não forem linguisticamente válidos [...], os resultados obtidos não terão significação lingüística. [...] levanta-se a grita dos lingüistas contra os seus confrades matemáticos. [...] queixam-se eles mui justamente de que alguns matemáticos utilizem a língua como instrumento de elucubrações abstratas, [...] esquecendo a língua como objetivo essencial de suas pesquisas. (BIDERMAN (1969, p. 118)

A partir da constatação da existência dessas divergências, BIDERMAN (1967, p. 119) afirma o que significa aplicar os métodos estatísticos no universo linguístico: "[...] a língua é um código cujos símbolos obedecem a certas frequências determinadas e previsíveis. [...] a língua é uma população e as realizações do discurso podem ser consideradas como amostras desse universo." E conclui justificando a utilidade dessas operações:

> "[..] tanto o lingüista preocupado essencialmente com a ciência da linguagem, como o historiador das línguas, o filólogo inclinado aos estudos literários e ao estabelecimento de textos, encontrará na prática da Estatística Lingüística um rico filão para explorar, revertendo-o em moeda sonante no comércio prático da sua ciência específica." (BIDERMAN, 1967, p. 119)

Dessa maneira, verificar a frequência do léxico como critério de avaliação de uma produção didática para ensino de língua em relação ao léxico de referência está no consenso de verificar a média de frequência do uso geral, sendo ou não aceito pela comunidade falante. Isso é uma tarefa complicada porque, além de isolar os homógrafos em uma quantificação computadorizada, Biderman (1967, p. 117) afirma que "toda realização do discurso comporta em maior ou menor grau uma escolha por parte do falante, ou do escritor, dos elementos léxicos, morfológicos e sintáticos disponíveis da língua no nível em que ele a atualiza." E, no caso do ensino de línguas ou de avaliação de proficiência linguística, existe uma escolha lexical para cada nível de ensino. Não é didático, por exemplo, ensinar o léxico de uma terminologia técnica seja de qualquer área do conhecimento (ou regionalismos) no nível iniciante de aprendizado de qualquer língua estrangeira.

Como veremos, se fazem necessários o estudo e a elaboração de listas do léxico de frequência para elaborar materiais de ensino de línguas, fazer planos de ensino ou exames de proficiência. Segundo Biderman (1967):

> "as pesquisas de Lexicoestatística visavam chegar a um diagnóstico da estrutura quantitativa do léxico das línguas com o objetivo de elaborar listas de freqüência de palavras para selecionar adequadamente o vocabulário a ser utilizado no ensino/aprendizagem do léxico". (BIDERMAN, 1967, p. 179)

Essas ações devem ser realizadas no ensino de português brasileiro, pois existe uma demanda crescente de estudos dessa língua, mas ainda carecemos de material didático para o ensino de PLA. Em parte, este artigo pretende trazer um ensaio experimental com a frequência do léxico de livros didáticos de ensino de PLA.

2.1 Lista de frequência do léxico do português falado na cidade de São Paulo

O Léxico de Frequência utilizado como referência neste artigo é uma contribuição do Projeto NURC[1]. Neste projeto, houve a intenção de elaborar o Léxico de Frequência da língua portuguesa contemporânea falada em todo o Brasil. Para isso, seria utilizado o inventário lexical coletado de São Paulo, Rio de Janeiro, Recife, Salvador e Porto

1. Sigla de "Norma Urbana Oral Culta". A obra de CASTILHO, A. T.; PRETI, D. (Org.): A linguagem falada culta na cidade de São Paulo, volume 1 traz a bibliografia completa sobre o Projeto NURC.

Alegre. No entanto, como havia outras prioridades, limitaram aos dados recolhidos na cidade de São Paulo (USP). Além disso, o levantamento lexical teve também como objetivo dominar as técnicas de quantificação e de processamento eletrônico do léxico de frequência. O trabalho do Projeto NURC resultou na lista do Léxico de Frequência utilizado como referência para a verificação lexicológica deste artigo.

2.2 Certificação de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras)

O Celpe-Bras é um exame que possibilita a Certificação de Proficiência em Língua Portuguesa para Estrangeiros. Desenvolvido e outorgado pelo Ministério da Educação (MEC), aplicado no Brasil e em outros países com o apoio do Ministério das Relações Exteriores (MRE) é o único certificado de proficiência em português como língua estrangeira reconhecido oficialmente pelo Governo Brasileiro. Internacionalmente, é aceito em empresas e instituições de ensino como comprovação de competência na língua portuguesa, e no Brasil é exigido pelas universidades para ingresso em cursos de graduação e em programas de pós-graduação, bem como para validação de diplomas de profissionais estrangeiros que pretendem trabalhar no país.[2]

2.2.1 Os níveis do Celpe-Bras:

O exame Celpe-Bras avalia a competência em língua portuguesa em quatro níveis que vão do Intermediário ao Avançado, não sendo avaliada, portanto, a competência para níveis básicos de proficiência. Os níveis atribuídos são:

Intermediário: Compreender textos orais e escritos sobre assuntos limitados, em contextos conhecidos e situações do cotidiano. Inadequações e interferências da língua materna são frequentes na escrita e na pronúncia, mas sem comprometimento da comunicação.

Intermediário Avançado: Compreender textos orais e escritos sobre assuntos limitados, em contextos conhecidos e situações do cotidiano. Inadequações e interferências da língua materna devem ser menos frequentes na escrita e na pronúncia do que no nível anterior.

Avançado: Domínio amplo da língua, capaz de produzir textos orais e escritos sobre assuntos variados, em contextos conhecidos e desconhecidos, sendo admitidas inadequações ocasionais na comunicação, sobretudo em contextos desconhecidos.

Avançado Superior: Domínio amplo da língua, capaz de produzir textos orais e escritos sobre assuntos variados, em contextos conhecidos e desconhecidos, com inadequações menos frequentes que no nível anterior.

2. Fonte: <http://celpebras.inep.gov.br/inscricao/> acesso em 30 de julho de 2016.

3 I PROCEDIMENTOS METODOLÓGICOS DESENVOLVIDOS NA PESQUISA

As listagens lexicais foram exportadas para programas de computador e foram empregadas variadas tecnologias digitais, aplicadas nessa ordem: (i) digitalização da lista de frequência lexicológica do português; (ii) isolamento e digitalização do léxico dos livros didáticos analisados (*Novo Avenida Brasil* e *Falar, Ler, Escrever … Português, um curso para estrangeiros*) e eliminação de redundâncias; (iii) uso de programas com *optical character recognition* (OCR) para conversão das imagens para o formato em que o computador seja capaz de fazer cálculos; (iv) correção manual dos erros feitos pelo computador; (v) conversão das planilhas para uma plataforma de gerência de banco de dados simples, para realizar as comparações utilizando a linguagem de consulta estruturada (SQL); (vi) geração de gráficos comparativos.

Quanto ao material didático selecionado, foi utilizado aquele adotado pelo curso de Português para Estrangeiros do Centro de Extensão da Faculdade de Letras da UFMG: os manuais *Novo Avenida Brasil* e *Falar, Ler, Escrever … Português, um curso para estrangeiros,* ilustrados na próxima página:

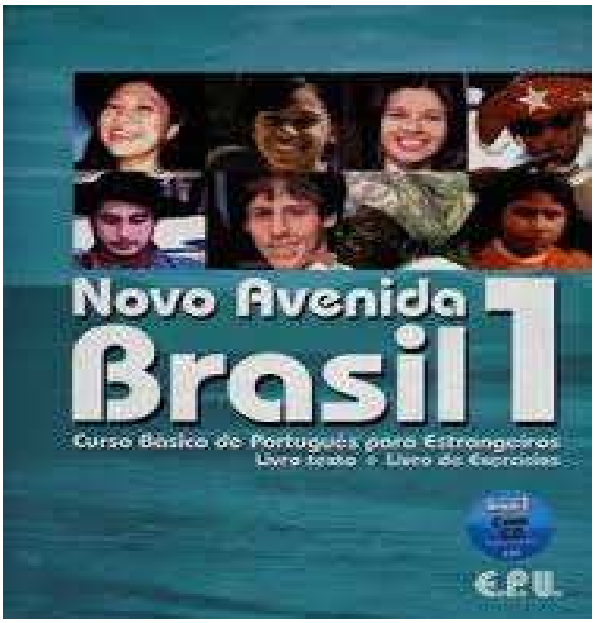


Figura 1 – Livro didático *Novo Avenida Brasil*

Fonte: <http://images.submarino.io/produtos/01/00/item/7150/6/7150643G1.jpg> acesso em 30 de julho de 2016.

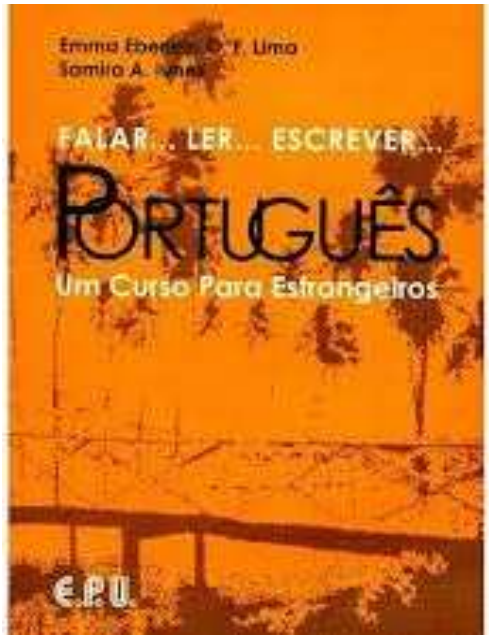
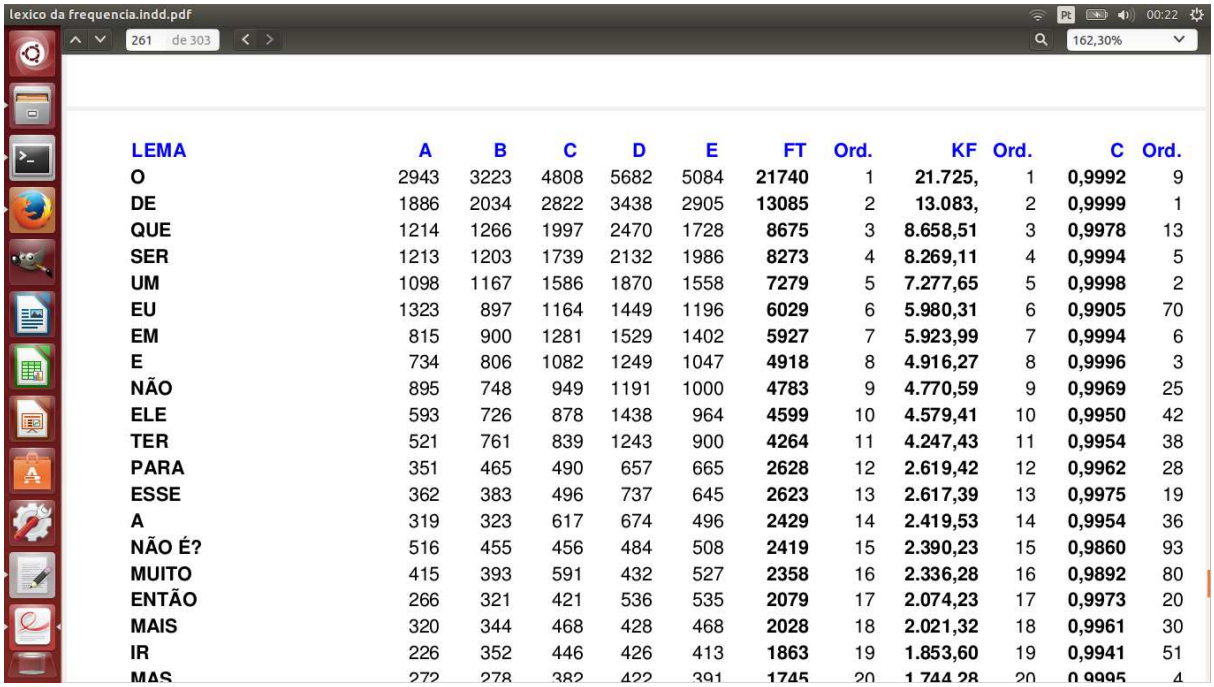


Figura 2 – Livro didático *Falar, Ler, Escrever … Português, um curso para estrangeiros*

Fonte: <http://images.submarino.io/produtos/01/00/item/176/6/176642GG.jpg> acesso em 30 de julho de 2016.

A seguir, são apresentadas imagens que ilustram os passos desenvolvidos na análise computacional.

Inicialmente, foi digitalizada a lista elaborada por Enzo Del Carratore e Jayme Laperuta Filho (Léxico de Frequência do Português falado na Cidade de São Paulo) e a lista digitalizada foi exportada para o software *PHPMyAdmin*.



LEMA	A	B	C	D	E	FT	Ord.	KF	Ord.	C	Ord.
O	2943	3223	4808	5682	5084	21740	1	21.725,	1	0,9992	9
DE	1886	2034	2822	3438	2905	13085	2	13.083,	2	0,9999	1
QUE	1214	1266	1997	2470	1728	8675	3	8.658,51	3	0,9978	13
SER	1213	1203	1739	2132	1986	8273	4	8.269,11	4	0,9994	5
UM	1098	1167	1586	1870	1558	7279	5	7.277,65	5	0,9998	2
EU	1323	897	1164	1449	1196	6029	6	5.980,31	6	0,9905	70
EM	815	900	1281	1529	1402	5927	7	5.923,99	7	0,9994	6
E	734	806	1082	1249	1047	4918	8	4.916,27	8	0,9996	3
NÃO	895	748	949	1191	1000	4783	9	4.770,59	9	0,9969	25
ELE	593	726	878	1438	964	4599	10	4.579,41	10	0,9950	42
TER	521	761	839	1243	900	4264	11	4.247,43	11	0,9954	38
PARA	351	465	490	657	665	2628	12	2.619,42	12	0,9962	28
ESSE	362	383	496	737	645	2623	13	2.617,39	13	0,9975	19
A	319	323	617	674	496	2429	14	2.419,53	14	0,9954	36
NÃO É?	516	455	456	484	508	2419	15	2.390,23	15	0,9860	93
MUITO	415	393	591	432	527	2358	16	2.336,28	16	0,9892	80
ENTÃO	266	321	421	536	535	2079	17	2.074,23	17	0,9973	20
MAIS	320	344	468	428	468	2028	18	2.021,32	18	0,9961	30
IR	226	352	446	426	413	1863	19	1.853,60	19	0,9941	51
MAS	272	278	382	422	391	1745	20	1.744,28	20	0,9995	4

Figura 3 - Lista do Léxico de Frequência do Português Falado em SP em formato digital

Fonte: <https://www.marilia.unesp.br/Home/Publicacoes/lexico da frequencia.indd.pdf> acesso em 30 de julho de 2016.

PHPMyAdmin

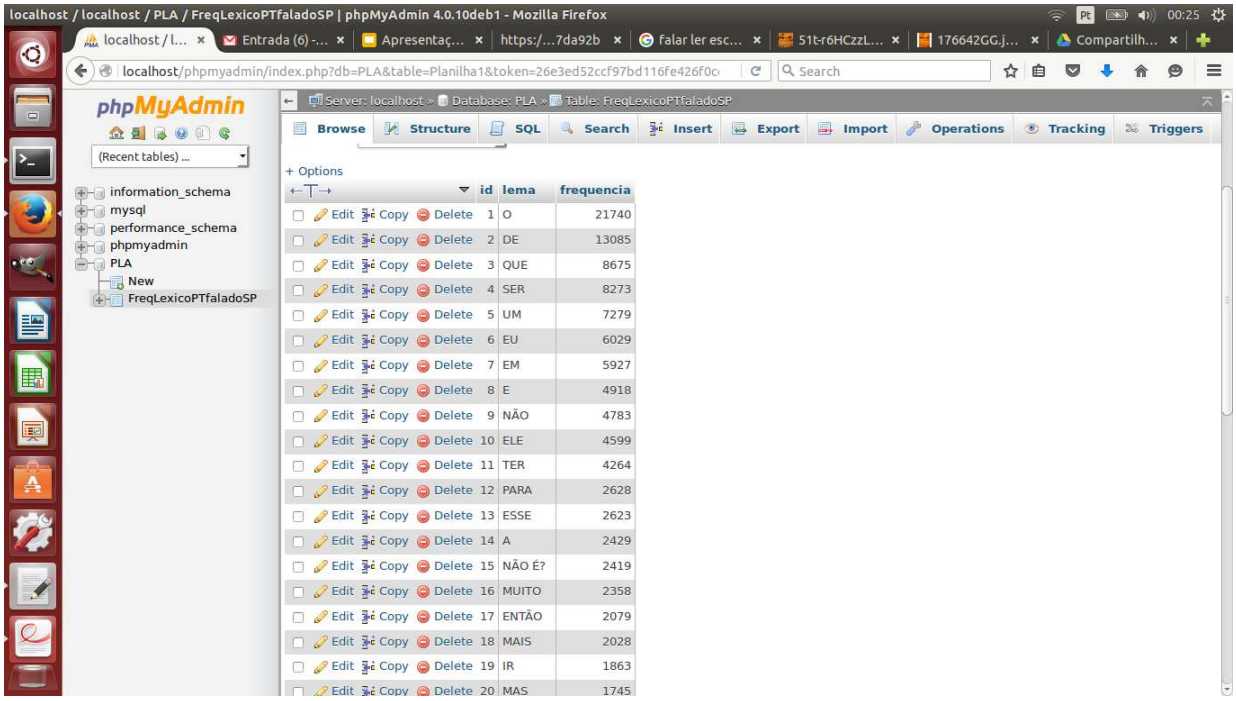


Figura 4 - Lista da fig. 2 exportada para o software

Fonte: Elaborado pelos autores com a utilização do software *PHPMyAdmin* (2018)

Posteriormente, os diálogos dos livros foram digitalizados e as imagens foram processadas com o programa gOcr[3], como ilustrado na figura abaixo. Esse programa possui o reconhecedor ótico de caracteres[4], que é capaz de converter a imagem do caracter em um formato que possa ser processado em editores de textos.

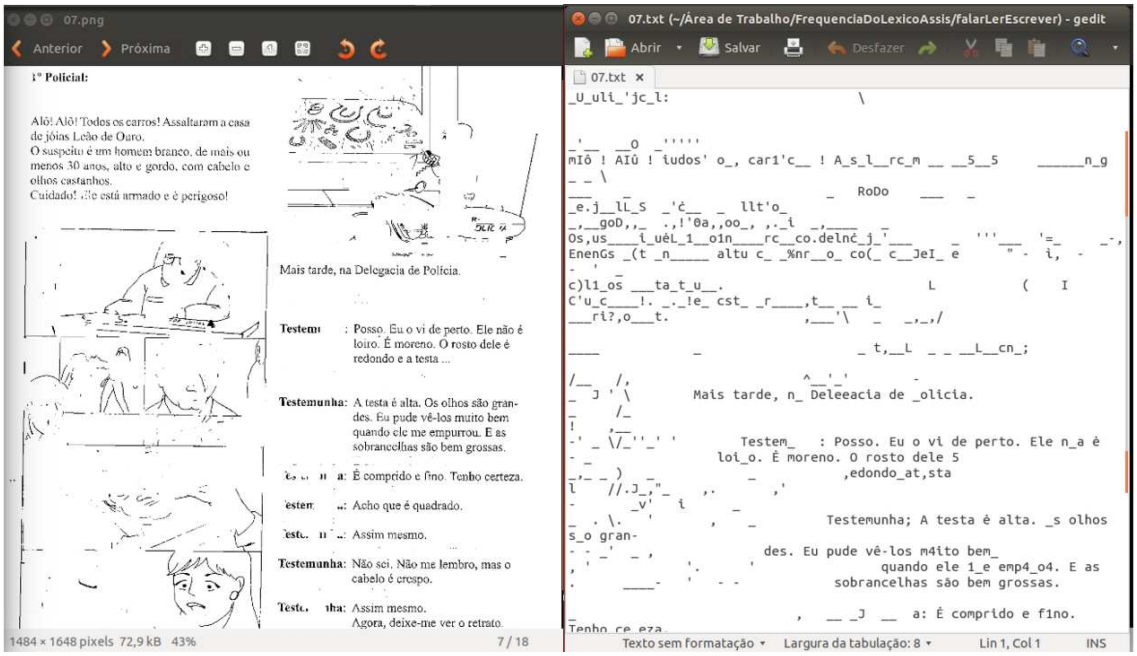


Figura 5 - Conversão de imagem com texto para texto editável com o gOcr

Fonte: Elaborado pelos autores com a utilização do software gOcr. (2018)

3. Programa gratuito de reconhecimento óptico de caracteres.
4. Do inglês, Optical Character Recognition.

Após a conversão ilustrada acima, o texto foi selecionado e copiado para um editor de texto, em que possíveis erros de conversão do Ocr foram corrigidos. Os erros são associados à semelhança de alguns caracteres. Como exemplo, em alguns casos a letra ‘e’ foi reconhecida como ‘c’ e vice-versa. Esses casos são muito raros e foram corrigidos um por um, a partir da releitura do texto convertido.

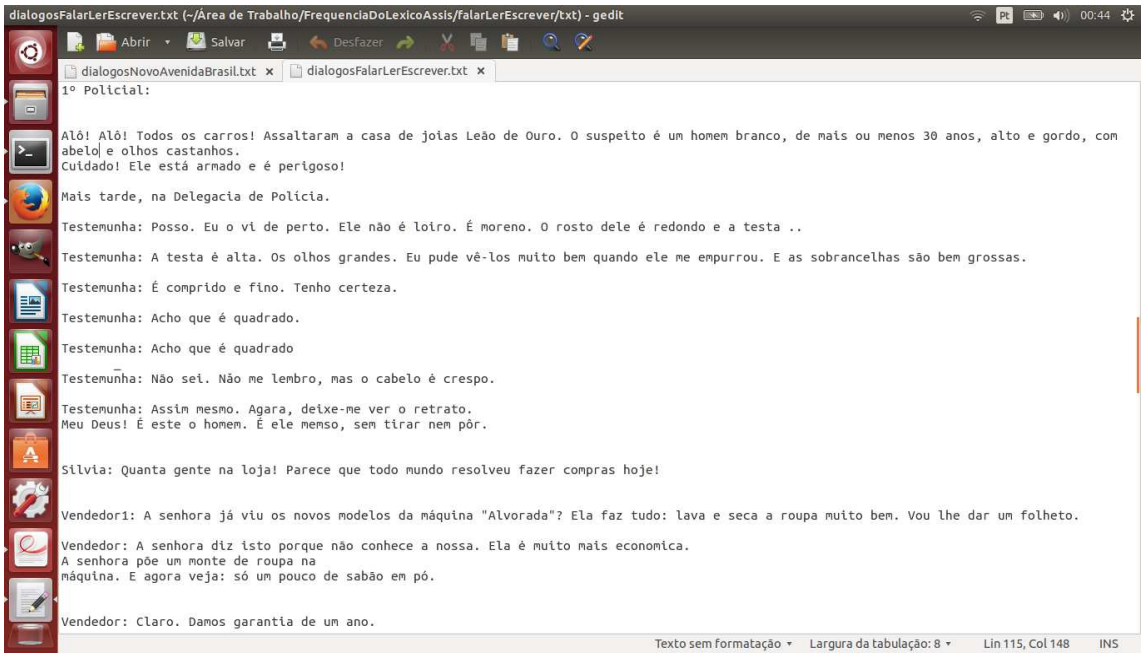
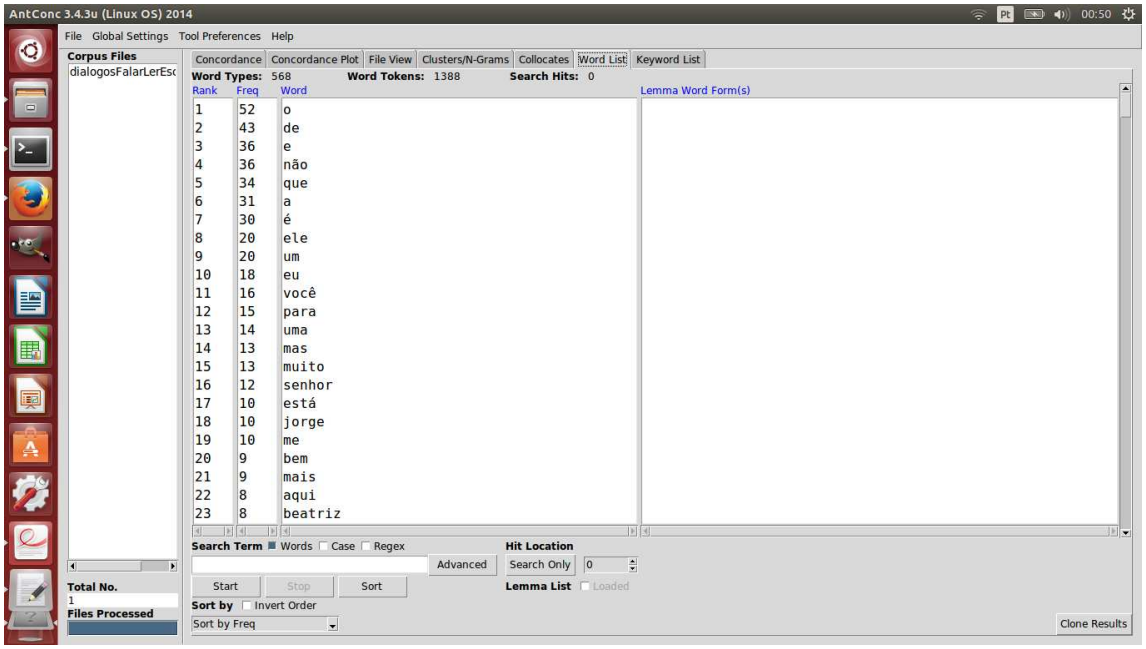


Figura 6 - Editor de texto com todo o conteúdo selecionado do livro

Fonte: Elaborado pelos autores com a utilização do software gEdit. (2018)

Depois que todos os textos foram inseridos no editor de texto, foi feita a contagem dos itens lexicais utilizando outro programa de computador: o *AntConc*[5].



5. Este programa permite a contagem dos itens lexicais dispondo o léxico por ordem alfabética ou por frequência.

Figura 7 - Geração da lista de frequência de “Falar, Ler, Escrever … Português, um curso para estrangeiros”, resultado do programa AntConc

Fonte: Elaborado pelos autores com a utilização do software *AntConc.* (2018)

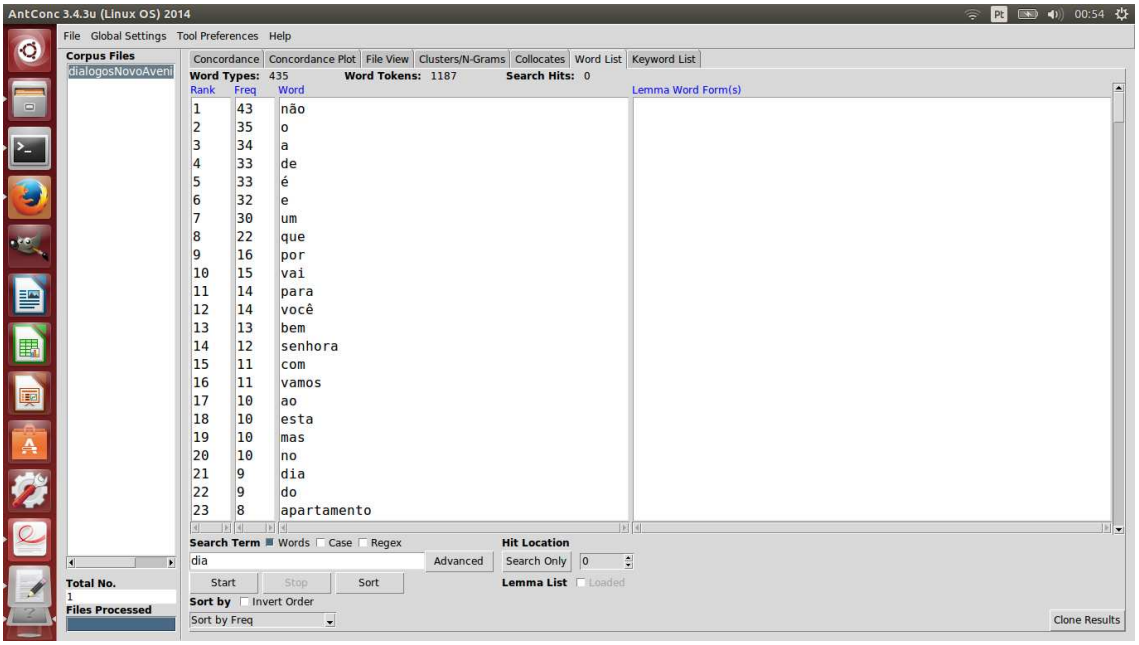
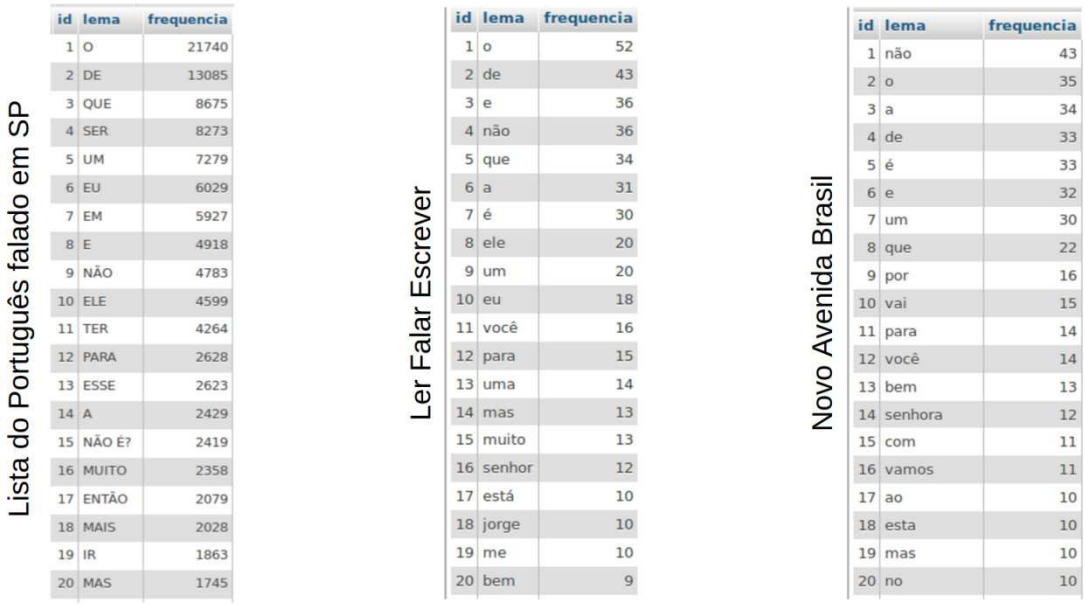


Figura 8 - Geração da lista de frequência de “Falar, Ler, Escrever … Português, um curso para estrangeiros”, resultado do programa AntConc

Fonte: Elaborado pelos autores com utilização do software *AntConc.* (2018)

Foi elaborada a conversão das listas para a plataforma de gerência de banco de dados (*PHPMyAdmin*) para a realização de comparações, utilizando a linguagem de consulta estruturada (SQL). Nessa mesma plataforma, foram colocadas as listas de frequência do léxico elaborada por Enzo Del Carratore e Jayme Laperuta Filho (Léxico de Frequência do Português falado na Cidade de São Paulo). Essa lista foi comparada com a lista do léxico dos materiais didáticos selecionados (*Novo Avenida Brasil* e *Falar, Ler, Escrever … Português, um curso para estrangeiros*), por meio de consultas com uso da linguagem SQL.



Lista do Português falado em SP

id	lema	frequencia
1	O	21740
2	DE	13085
3	QUE	8675
4	SER	8273
5	UM	7279
6	EU	6029
7	EM	5927
8	E	4918
9	NÃO	4783
10	ELE	4599
11	TER	4264
12	PARA	2628
13	ESSE	2623
14	A	2429
15	NÃO É?	2419
16	MUITO	2358
17	ENTÃO	2079
18	MAIS	2028
19	IR	1863
20	MAS	1745

Ler Falar Escrever

id	lema	frequencia
1	o	52
2	de	43
3	e	36
4	não	36
5	que	34
6	a	31
7	é	30
8	ele	20
9	um	20
10	eu	18
11	você	16
12	para	15
13	uma	14
14	mas	13
15	muito	13
16	senhor	12
17	está	10
18	jorge	10
19	me	10
20	bem	9

Novo Avenida Brasil

id	lema	frequencia
1	não	43
2	o	35
3	a	34
4	de	33
5	é	33
6	e	32
7	um	30
8	que	22
9	por	16
10	vai	15
11	para	14
12	você	14
13	bem	13
14	senhora	12
15	com	11
16	vamos	11
17	ao	10
18	esta	10
19	mas	10
20	no	10

Figura 9 - Listas geradas pelo PHPMyAdmin

Fonte: Elaborado pelos autores com utilização do software *PHPMyAdmin.* (2018)

As comparações realizadas foram feitas com o léxico selecionado com os diálogos dos livros didáticos. Do "*Novo Avenida Brasil"* foram recolhidos 13 diálogos com 1196 *tokens* e do *"Falar, Ler, Escrever … Português, um curso para estrangeiros"* foram recolhidos de 18 diálogos com 1393 *tokens*

CONSULTAS REALIZADAS

a) com a lista do livro "*Novo Avenida Brasil"*:

```
SELECT `lema ` FROM `NovoAvenidaBrasil`
WHERE `lema` IN (SELECT `lema` FROM `FreqLexicoPTfaladoSP`);
```

b) com a lista do livro *"Falar, Ler, Escrever … Português, um curso para estrangeiros"* :

```
SELECT `lema ` FROM `FalarLerEscrever`
WHERE `lema` IN (SELECT `lema` FROM `FreqLexicoPTfaladoSP`);
```

4 I CONSIDERAÇÕES FINAIS

Após a realização dos processamentos computacionais que permitem a comparação digitalizada do léxico apresentado nos manuais com a lista de frequência lexicológica do português falado em São Paulo, chegou-se aos dados descritos abaixo:

Livro didático	Itens	Lista de Referência
Novo Avenida Brasil	1196	207 (17,3%)
Falar, Ler, Escrever…Português, um curso para estrangeiros	1393	264 (18,9%)

Quadro 1 - Presença do léxico da lista de referência nos livros didáticos.

Fonte: Elaborado pelos autores. (2018)

Os textos selecionados do livro *Novo Avenida Brasil* possuem no total 1196 itens lexicais e o manual *Falar, Ler, Escrever … Português, um curso para estrangeiros* possui 1393. As colunas, a seguir, mostram quantos dos itens lexicais de suas respectivas listas foram encontrados nos materiais didáticos: 207 (17,3%) itens da lista de referência foram encontrados no léxico apresentado no livro *Novo Avenida Brasil* e 264 (18,9%) itens da mesma lista foram encontrados no léxico do livro *Falar, Ler, Escrever … Português, um curso para estrangeiros*. Outra maneira de visualizar os dados acima é por meio do gráfico abaixo.

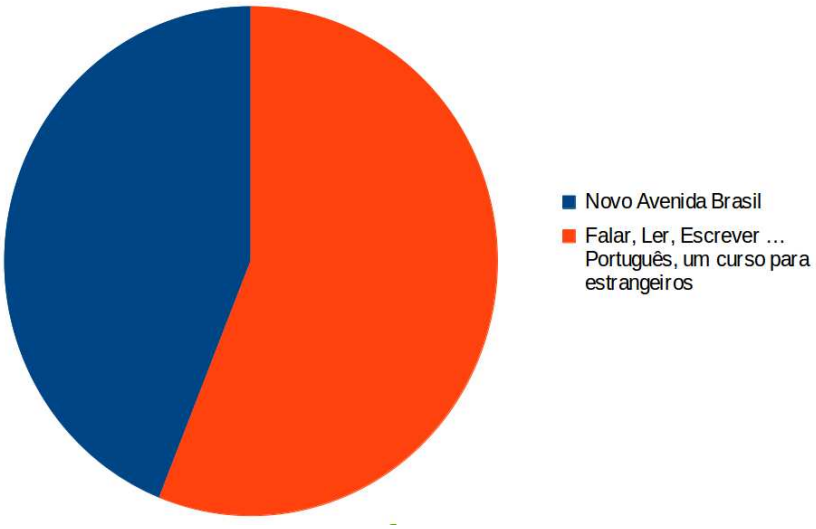


Gráfico 1 - Presença do léxico da lista de referência nos livros didáticos.

Fonte: Elaborado pelos autores. (2018)

A partir desse gráfico vemos que, dos materiais didáticos analisados, o livro "*Falar, Ler, Escrever … Português, um curso para estrangeiros"* apresentou melhor adequação que o "*Novo Avenida Brasil*".

A comparação entre as listas de frequência lexicológica com a lista de frequência usada como referência permitiu verificar a adequação do livro didático quanto a esse importante aspecto linguístico, que é o vocabulário da língua. Assim, essa comparação constitui um critério útil para avaliação de livros didáticos.

O resultado das comparações pode ainda auxiliar autores de livros didáticos na adaptação de textos autênticos para fins didáticos. Ainda não foi produzido o léxico de frequência para cada nível do Celpe-Bras. Essas listas poderiam ser utilizadas pelos

professores de PLA como suporte para a avaliação de textos de alunos, permitindo a classificação quanto à aprendizagem lexical e ao nível de competência linguística de acordo com o Celpe-Bras.

REFERÊNCIAS

BIDERMAN, M. T. C. **Estatística linguística**. Revista Alfa, São Paulo, v. 11, p. 117-128, 1967.

BIDERMAN, M. T. A. C. **The quantitative side of feature language: a Frequency Dictionary of Contemporary Brazilian Portuguese**. Alfa (São Paulo), v.42, n.esp., p. 161-181, 1998;

DEL CARRATORE, E. **Léxico de freqüência do português fala na cidade de São Paulo: projeto NURC**. Cultura Acadêmica (São Paulo), 2011.

LIMA, E. E. O. F. **Falar… Ler… Escrever… português. Um curso para estrangeiros**. EPU (São Paulo), 1999;

LIMA, E. E. O. F. **Novo Avenida Brasil.** EPU (São Paulo), 2008;