http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

CORPUS ABG
ABG CORPUS

Aline de Lima Benevides
Universidade de São Paulo
benevides.aline12@gmail.com
Bruno Ferrari Guide
Universidade de São Paulo
bruno.fguide@gmail.com

RESUMO: Este artigo apresenta a metodologia empregada na compilação de um *corpus* linguístico do Português Brasileiro, o qual foi denominado de *Corpus* ABG, e no desenvolvimento de algumas ferramentas computacionais. O objetivo deste trabalho é reunir uma grande quantidade de textos, escritos e orais, que possa representar o falar brasileiro a fim de ser fonte de extração de dados fonológicos quantificados para duas pesquisas, a saber, Guide (2016) e Benevides (2017). O *corpus* contabiliza 3.616.625 ocorrências de palavras e 92.602 tipos de palavras, sendo que 1.938.805 ocorrências são provenientes dos *corpora* de fala e 1.676.820 ocorrências dos *corpora* escritos. Ancorado na metodologia da Linguística de *Corpus* e por meio de ferramentas computacionais desenvolvidas em Linguagem Python, o presente artigo divulga e disponibiliza à comunidade científica o *Corpus* ABG, as ferramentas computacionais (acentuador, categorizador de estruturas fonológicas, silabificador) e algumas informações fonológicas (acentuais e silábicas) já extraídas do *corpus*. Além disso, faz um convite a novas explorações dos dados a todos os pesquisadores que tiverem interesse.
PALAVRAS-CHAVE: *corpus* linguístico*;* linguística computacional; português brasileiro.

ABSTRACT: The present paper presents the task of compiling a linguistic corpus of Brazilian Portuguese, which was undertaken by the authors. It is called ABG Corpus, and this article is also about the computational tools developed for the task. Our main goal is to reunite a large amount of texts, both from spoken and written language to, in the best way possible, represent the Brazilian language in a way that we could use it as a database for our researches, Guide (2016) and Benevides (2017). The ABG corpus has 3.616.625 word tokens and 92.602 types of words, being that 1.938.805 of those tokens are from spoken language corpora and 1.676.820 tokens come from written corpora. Based on the corpus linguistics framework and through the use of computational tools developed using Python, this article shows and provides access to the ABG Corpus, the computational tools (stress marker, phonological structure identifier, syllabifier), as well as some phonological information (stress and syllable related), already present on the corpus. We end by inviting the community to further expand our findings and explore this new tool.
KEYWORDS: linguistic corpus; computational linguistics; Brazilian Portuguese.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

1 Introdução

O presente artigo pretende apresentar a metodologia empregada na compilação (coleta + tratamento dos dados) de um *corpus* linguístico, denominado *Corpus* ABG[1], e das ferramentas computacionais resultantes dessa empreitada, isto é, silabificador, transcritor fonológico, categorizador morfológico/lematizador e acentuador. Além disso, apresentará duas tarefas executadas computacionalmente: codificação de estruturas fonológicas e a extração de frequências das palavras no *corpus*.

O *corpus* foi compilado com a finalidade de ser fonte de dados de duas pesquisas linguísticas que investigaram a emergência de padrões fonológicos na atribuição do acento primário em Português Brasileiro (doravante PB) (cf. GUIDE, 2016; BENEVIDES, 2017). Diante da riqueza dos dados extraídos, apresentam-se, a seguir, as decisões metodológicas adotadas, bem como uma síntese dos dados quantitativos do *corpus*.

Este artigo estrutura-se da seguinte maneira: a seção 2 apresenta um breve panorama sobre a Linguística de *Corpus*; a seção 3 trata das justificativas para a compilação de um *corpus* para o PB; a seção 4 descreve a metodologia utilizada na compilação do *Corpus;* a seção 5 reserva-se à descrição das decisões metodológicas adotadas durante o processo de tratamento dos dados do *corpus*[2]; a seção 6 apresenta um panorama sobre os aspectos quantitativos do *Corpus* ABG e, por fim, na seção 7, sintetizam-se as principais contribuições que esta pesquisa traz para os estudos fonológicos e probabilísticos do idioma.

2 Linguística de *corpus*

A Linguística de *Corpus* (doravante LC) é uma abordagem dos estudos da linguagem que se preocupa com os métodos de compilação de *corpora* linguísticos. O objetivo da área é poder realizar estudos empíricos sobre a linguagem, assim como descrever os seus padrões de uso. Tais noções levam a compreender a linguagem como um sistema que pode ser descrito probabilisticamente (cf. SARDINHA, 2000a).

O principal objetivo da LC é estabelecer aos pesquisadores uma metodologia para a compilação de um *corpus* de pesquisa. Para Sanchez (1995 apud SARDINHA, 2000a, p. 338), a definição mais completa e abrangente para o termo *corpus* consiste em:

> Um conjunto de dados lingüísticos (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 lingüístico 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.

Essa exposição apresenta os principais conceitos que estão por trás das técnicas

1 O *corpus* foi desenvolvido pelos autores a fim de ser utilizado nas pesquisas de mestrado de ambos (GUIDE, 2016; BENEVIDES, 2017) na Universidade de São Paulo.

2 Todos os *scripts* desenvolvidos para essa tarefa estão disponibilizados publicamente no seguinte endereço: <https://github.com/SauronGuide/corpusABG>.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

de composição de *corpora* para a LC*,* no caso: autenticidade, especificidade, adequação, extensão e representatividade.

Para um *corpus* ser autêntico, ele deve conter somente porções de linguagem (escrita ou falada) produzidas por falantes nativos da língua, sem que essas tenham sido geradas para compor o *corpus* em formação. A seleção dos textos deve pautar-se na especificidade da linguagem a ser demonstrada por meio do cumprimento preciso dos critérios estabelecidos pelos pesquisadores, cabendo aos usuários se adequarem à busca de *corpora* específicos para o objeto em análise.

A compilação do *corpus* deve obedecer ainda a alguns parâmetros extensionais propostos na especificação da metodologia a ser adotada. A definição da extensão deve incluir uma quantidade diversificada de gêneros, a proporção de textos para cada um deles e o número de palavras que possuirá, sendo esse o critério correntemente apontado na literatura como primordial para determinar a representatividade de um *corpus* (cf. BIBER, 1993; SARDINHA, 2000a; 2000b; MCENERY & WILSON, 2001).

No entanto, segundo Biber (1993), não é possível falar em representatividade sem que seja estabelecida *a priori* a população a qual se deseja representar, já que "representatividade refere-se à extensão em que uma amostra inclui uma gama inteira de variabilidade da população"[3] (BIBER, 1993, p. 243). A representatividade passa, então, a ser concebida como a extensão e a diversificação da amostra de determinada população.

Contrapondo-se a essa concepção, Sardinha (2000b) afirma que nenhum *corpus* pode ser rotulado como uma amostra representativa, visto que só é possível falar em representatividade quando se conhece toda a população representada. Entretanto, segundo o autor, torna-se factível estimar o tamanho de um *corpus* por meio do número de palavras que ele possui, muito embora não se possa prever a extensão real de uma língua, como apresentado no Quadro 1.

Quadro 1: Tamanho relativo do corpus por número de palavras.

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

Fonte: Sardinha (2000b, p. 346).

O número de palavras e a diversidade de porções de linguagem tornam-se fundamentais para que vocábulos e padrões menos frequentes possam emergir. Baseado nessas afirmações, Sinclair (1996 apud SARDINHA, 2000a, p. 345) postula que um *corpus* deve "ser tão grande quanto a tecnologia permitir para a época". Atualmente, o tratamento dos dados tende a ser facilitado em decorrência da diversidade de ferramentas computacionais disponíveis; no entanto, muita mão de obra ainda é requerida para a seleção e compilação do *corpus*.

3 Trecho Original: "representativeness refers to the extent to which a sample includes the full range of variability in a population".

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

3 Justificativa para a compilação do *Corpus* ABG

Diversos *corpora* confiáveis podem ser encontrados disponíveis na internet gratuitamente para uso científico. Os mais famosos consistem em: Linguateca, ASPA, *Tycho Brahe*, C-ORAL-BRASIL e FrePOP.

A Linguateca (2015, s/p) consiste num "centro de recursos – distribuído – para o processamento computacional da língua portuguesa". Com uma equipe de pesquisadores de diversas localidades do mundo e com hospedagem (*site*) em Portugal, reúne diversos *links*/fontes de *corpora* e ferramentas (dicionários, buscadores, ferramentas computacionais etc) desenvolvidos para a Língua Portuguesa. Atualmente, é considerada a principal base de dados para estudos do português, sobretudo por conter uma requintada seleção de informação.

O Projeto ASPA (2014) consiste numa base de dados fonológicos *online,* o qual objetiva contribuir para o estudo do sistema fonológico do português, com uma metodologia pautada na Fonologia de Uso e na Teoria dos Exemplares. Para acessá-lo, é necessário um cadastro prévio de rápida obtenção. Nele, é possível realizar buscas ortográficas ou fonológicas por palavras, segmentos, tipos ou padrões fonológicos, obtendo suas frequências de tipo e de ocorrência.

O *Tycho Brahe*, também denominado *Corpus* Histórico do Português *Tycho Brahe* (GALVES & FARIA, 2010), como o nome especifica, reúne textos históricos escritos em português por autores nascidos entre 1380 e 1881. É composto de 68 textos com anotações morfológicas e sintáticas, os quais estão disponíveis na página do *corpus* (http://www.tycho.iel.unicamp.br/~tycho/corpus/).

O C-ORAL-BRASIL (RASO & MELLO, 2012) consiste num livro/DVD, o qual se volta para o estudo da fala espontânea do português brasileiro. O projeto é estruturado em duas fases: a primeira destinada à fala informal e a segunda à fala formal. Até o momento, apenas a primeira fase foi finalizada e encontra-se disponível por meio da compra do livro/DVD, possuindo, aproximadamente, 140 textos.

O FrePOP (*Frequency Patterns of Phonological Objects in Portuguese: Research and Applications*) (FROTA et al., 2010) é um banco de dados de textos escritos e falados, o qual objetiva fornecer a frequência de padrões fonológicos do português, tais como: palavra prosódica, caracteres ortográficos, clíticos, sílaba, segmentos, acento etc. As buscas podem ser realizadas por diversos filtros: variedade do português (europeu, brasileiro ou africano), por fatores sociais (idade, gênero, escolaridade e profissão), por período histórico (a partir do século XVI) e por tipo de *corpora* (falado, escrito, informal, conversação etc).

Diante da diversidade de *corpora* esboçada acima, a decisão de compilar um *corpus* deve-se à ausência de um *corpus* linguístico que fornecesse dados necessários para a computação de frequência por tipos silábicos e que permitisse acessar aos dados para a realização das mais variadas buscas.

Além de permitir pesquisas variadas, a compilação do *corpus*, que foi denominado de *Corpus* ABG, direcionou-se para o objeto de análise: o acento. Criou-se, portanto, um *corpus* que pudesse ser maximamente representativo da língua portuguesa para expressar os diferentes padrões acentuais que ocorrem nela. Para Sardinha (2000a, p. 348), "normalmente, corpora compilados em pequena escala por pesquisadores

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

individuais acabam sendo mais representativos do que os respectivos sub-corpora dos corpora gerais".

Outra vantagem diagnosticada refere-se à variabilidade, isso porque a especificação de uma única fonte de dados faz com que a variação seja mantida, a qual é sempre maior dentro do *corpus* do que entre *corpora* (cf. SARDINHA, 2000a; 2000b).

4 Metodologia de compilação do *Corpus* ABG

Conforme esboçado na seção 2, a primeira decisão a ser tomada antes do início da compilação de um *corpus* consiste na especificação da população que se deseja representar. No caso em questão, por se tratar de acento, pressupõe-se que os fenômenos que regem a atribuição acentual no português brasileiro são os mesmos independente da localidade do falante nativo do português, ou seja, um falante paulistano possui regras semelhantes aos cariocas, cearenses, gaúchos, amazonenses etc. Diante de tal fato, determinaram-se como população em investigação os falantes nativos do português brasileiro, não cabendo inquirir a cidade/estado em que os autores dos textos selecionados nasceram, apenas sua nacionalidade.

Num processo de amostragem balanceado, dada a abrangência de dialetos e as variedades diatópicas, o *corpus* deveria possuir a mesma quantidade de textos provenientes de todos os estados brasileiros. No entanto, dada a dificuldade em encontrar *corpora* linguísticos com transcrições de fala nas diferentes regiões – bem revisados e controlados – e a demanda temporal e braçal que tal processo poderia levar para a sua construção, resolveu-se utilizar como fonte de dados apenas os *corpora* disponibilizados na internet de fácil acesso, alcançando, com isso, o maior número de textos possível.

Além de textos falados (fala transcrita), buscaram-se textos que pudessem representar outra modalidade de utilização da língua, a escrita. Para tal, estabeleceram-se os seguintes tipos de textos: artigos, matérias jornalísticas e postagem de *blogs*. Essas três esferas visavam abarcar uma espécie de gradação de formalidade, possibilitando a emergência de léxicos variados, os quais, em geral, são específicos de certos registros.

Como o objetivo era expressar a língua em uso, determinou-se que a especificação da temporalidade dos textos consistia num critério primordial para manter tal decisão. Desse modo, os textos (falados e escritos) selecionados deveriam ser contemporâneos (cf. SARDINHA, 2000a), correspondendo às publicações posteriores ao ano de 2000. A quantidade de textos selecionados, as extensões e a metodologia empregada serão explicitadas a seguir, de acordo com o tipo de *corpus* compilado (falado ou escrito).

4.1 *Corpus* de fala

4.1.1 Projeto SP2010

O Projeto SP2010 (MENDES, 2010) é uma coletânea sociolinguística, coordenado pelo Prof. Dr. Ronald Beline Mendes da Universidade de São Paulo, que reúne amostras de fala paulistana. O *corpus* Projeto SP2010 compreende 60 transcrições de gravações com, aproximadamente, 1 hora de duração, coletados entre 2011 e 2013, de falantes que

Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

nasceram na cidade de São Paulo e que não se ausentaram dela por longo período de tempo. A amostra é estratificada por sexo, gênero, escolaridade e região da cidade.

Os arquivos de textos disponibilizados possuem o seguinte formato:

> SD1: ah... (vo)cê já (es)tá acostumado [risos]
> S1: ah (es)tou... (es)tou
> D1: (vo)cê falou que fez teatro também... né?
> S1: é... assim... não é... teatro... dá... aquela coisa... teatro amador né mais pra negócio de escola essas coisa... eu gostei...
> D1: sei sei
> S1: mas pa/ dei uma parada agora mas
> D1: ah... bacana
> (excerto extraído do documento 2010M29MPL-JairS-iva de Mendes (2010)).

As transcrições semiortográficas, como pode ser verificado no trecho acima, foram submetidas a um processo de limpeza, a qual foi realizada de modo automático por meio de um *script* desenvolvido em Linguagem Python. Os caracteres removidos consistiram em números, marcação de turno – Documentador (D) e Sujeito (S) –, pausas (...), truncamento (/), sinais de interrogação (?), exclamação (!), parênteses ( ) e colchetes ([]). As palavras transcritas com marcas de oralidade, tal como *(es)tou* e *(vo)cê*, tiveram seus sinais de pontuação (parênteses) excluídos para que a palavra ortográfica fosse mantida. Ainda, siglas que codificavam o tempo de gravação nas transcrições, como *seg* de segundos, e as marcas típicas de oralidade, como *ah, ahan, uhum*, também foram retiradas do *corpus*, já que não expressariam qualquer padrão acentual ou fonológico da língua.

4.1.2 Projeto SP2010 – parte 2

O Projeto SP2010, aqui denominado de parte 2, refere-se ao trabalho desenvolvido por Mendes & Oushiro (2012), projeto esse complementar ao Projeto SP2010. Ele contém 102 entrevistas gravadas desde 2009, seguindo os mesmos critérios metodológicos descrito na seção anterior.

Semelhante *script* de limpeza dos dados foi utilizado nessa versão do Projeto SP2010.

4.1.3 Iboruna

O *corpus* Iboruna (GONÇALVES, 2014) é um banco de dados pertencente ao Projeto ALIP (Amostra Linguística do Interior Paulista), que objetiva descrever o português falado na região de São José do Rio Preto. A amostra é composta por 151 entrevistas e 11 interações de fala em situações livres produzidas por falantes provenientes de Bady Bassit, Cedral, Guapiaçu, Ipiguá, Mirassol, Onda Verde e São José do Rio Preto, isto é, do Noroeste de São Paulo.

A metodologia empregada para coleta dos dados obedece aos critérios utilizados na Sociolinguística Variacionista, sendo assim estratificada por fatores sociais. A duração das gravações varia de 13 minutos a 1 hora e 30 minutos.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

Em decorrência da utilização de números, caracteres de pontuação (interrogação, parênteses, barra, etc), fez-se necessária a limpeza dos dados, a qual usou o mesmo *script* descrito anteriormente. A única especificidade deste *corpus* consiste no alto número de truncamentos, os quais, se não ambíguos, eram reparados pela palavra alvo, como *'crianç/'* para *criança* e, se ambíguos, eram excluídos do *corpus*, como *'livr/'* que poderia ser *livro, livre, livraria* – informações contextuais também foram utilizadas para determinar a palavra alvo.

4.1.4 C-ORAL-BRASIL

O C-ORAL-BRASIL (RASO & MELLO, 2012) contou com 139 textos de contextos informais. Em vista de a transcrição ser semiortográfica, a exclusão dos marcadores de fala e de transcrição novamente foi realizada por meios automáticos. Para esse *corpus*, adaptações foram necessárias para que as informações referentes aos falantes, como *ADR, *REN, também fossem eliminadas.

4.2 *Corpus* escrito

4.2.1 Artigos

A seleção de artigos científicos deveu-se à tentativa de representar uma porção de linguagem mais formal, a qual também compõe o uso linguístico.

Para evitar enviesamento temático, buscaram-se dez artigos para cada uma das cinco áreas temáticas escolhidas[4], isto é: economia, história, literatura, semiótica e sociologia, as quais foram extraídas de duas revistas científicas renomadas no assunto (1).

(1) Economia: *Economia e Sociedade e Estudos Econômicos*
História: *Anpuh- Brasil e Antíteses*
Literatura: *Cadernos do IL e Letrônica*
Semiótica: *Diadorim e Revista Alfa*
Sociologia: *Contemporânea e Estudos de Sociologia*

Dessas revistas, utilizaram-se as edições mais recentes que possuíssem textos produzidos por brasileiros – a inferência foi realizada através do sobrenome desses[5] – com extensão entre 15 e 30 páginas. Os artigos selecionados foram publicados a partir de novembro/dezembro de 2013.

4 Sabe-se que a simples determinação de áreas para compor um *corpus* já implica enviesamento temático; contudo, buscaram-se com essas escolhas diversificar os temas abordados a fim de que o léxico empregado pudesse ser o mais abrangente possível.

5 Somente após a compilação e transcrição total do *corpus* que se percebeu o equívoco no método de discriminação dos autores por meio dos seus sobrenomes. Métodos mais eficazes, como a busca pelo currículo Lattes, poderiam ter sido empregados.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

De maneira análoga aos textos falados, esses artigos precisaram passar por uma limpeza prévia, na qual foram excluídos os números e os sinais de pontuação.

4.2.2 Textos jornalísticos

Os textos jornalísticos enquadram-se nas porções de linguagem com formalidade mediana, já que, por serem publicados por grandes canais de comunicação, tendem a seguir a Gramática Normativa da Língua Portuguesa. Contudo, não precisam ser altamente formais, posto que devem ser de fácil leitura e entendimento para os diversos públicos consumidores.

Ao considerar a facilidade na leitura, o acesso aos textos e, principalmente, o alcance de leitores (em quantidade), supôs-se que os *sites* de jornais mais acessados seriam o da Folha de São Paulo e o do Estadão. Determinou-se, assim, que 15 edições completas seriam compiladas em dias alternados entre 29 de setembro de 2014 e 29 de outubro de 2014. Estima-se que mais de 1.500 textos foram compilados em conjunto.

Para esse *corpus*, informações referentes a datas, *websites*, *e-mails* foram excluídas.

4.2.3 Postagem de *blogs*

Diante da corrente utilização dos meios digitais e da inserção dos indivíduos cada vez maior nas mídias sociais, buscou-se uma fonte de extração de dados que pudesse representar minimamente essa nova característica social de uso da língua, ao mesmo tempo em que não empregasse formas ortográficas específicas dessas mídias, como *tbm, vc, smp, tecê* etc.

Selecionaram-se dez postagens recentes com mais de 2000 caracteres de 30 *blogs*, seguindo as seguintes dez temáticas: moda, *game*, maquiagem, culinária, música, natureza, viagem, animais de estimação, saúde e informática, três por área.

Para assegurar que os *blogs* eram visitados por muitos usuários, procuraram-se em buscadores os mais famosos e seguidos da *web* nos temas elencados acima (veja (2)). As postagens também passaram por uma limpeza automática, excluindo números, *e-mails*, datas etc.

(2) Moda: *Depois dos quinze, Coisas de Diva e Lala Rudge*
Game: *Mais de um blog de game, Game Over e Universo dos games*
Maquiagem: *Luciane Ferraes, Blog da Fabi e Boca rosa*
Culinária: *Cozinha travessa, Cuecas na cozinha e Sabor no prato*
Música: *Música brasileira, Música na veia e Somos música*
Natureza: *Natureza e Paz, Defensor da Natureza e Grupo ecológico*
Viagem: *Matraqueando, Preciso viajar e 360 meridianos*
Animais de Estimação: *Pet Care, Meu amigo Pet e Blog do gato*
Saúde: *Saúde Plena, Mulheres com saúde e Saúde Business 365*

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

Informática*: Blog da informática, Papo de informática e O segredo da informática*

5 A construção do *Corpus* ABG

A definição de uma metodologia consiste no primeiro passo durante o processo de compilação de um *corpus*, isso porque a simples seleção de textos nem sempre fornece os dados que os pesquisadores necessitam. Para tal, faz-se necessário preparar o *corpus* e buscar ferramentas que permitam a extração dos dados.

Reiterando, a compilação deste *corpus* objetivou extrair dados fonológicos a respeito do acento primário no português brasileiro. A necessidade de buscar padrões fez com que algumas ferramentas fossem desenvolvidas e utilizadas por meios computacionais, as quais consistem em: silabificador, transcritor fonológico, categorizador morfológico/lematizador, acentuador e codificador de estruturas fonológicas.

5.1 Decisões metodológicas

Os *corpora* apresentados acima foram subdivididos em dois *corpora*, um de fala e outro de escrita. Essa divisão é apenas metodológica, já que todos os dados serão utilizados para compor o *Corpus* ABG e para representar o uso concreto da língua. Ela tem a vantagem de permitir aos usuários a realização de buscas por tipo de texto pretendido.

Após essa separação, os *corpora* foram transformados em uma lista de palavras, a qual se tornou *input* para as próximas ferramentas. Diante da complexidade de definir o que é uma palavra, adota-se a seguinte concepção, a qual é correntemente utilizada nas metodologias computacionais[6].

> Temos evitado e continuaremos a evitar falar de "palavra", uma vez que a este termo não corresponde qualquer conceito linguístico preciso. Em uso aproximativo – e não linguístico –, o melhor que podemos fazer é identificar "palavra" com o segmento que na escrita portuguesa é precedido e seguido de um espaço em branco: em *os gatos já comeram* haveria assim quatro palavras (BARBOSA, 1994, p. 134).

Ressalta-se que a utilização de uma lista de palavra não permite que o significado de alguns vocábulos seja depreendido corretamente, posto que o contexto lexical foi excluído, impossibilitando assim a diferenciação de casos triviais como *manga* (fruta) e *manga* (de camisa), casos esses que seriam facilmente distinguidos pelo contexto. No entanto, como a consideração de tal informação contextual conduziria ao aumento significativo da complexidade do algoritmo e do custo computacional por requerer *tags* morfológico e sintático apropriados, decidiu-se que a extração de padrões por meio de uma lista de palavra já seria apropriada para os fins da referida pesquisa.

6 Assume-se aqui a concepção de palavra correntemente empregada pela Linguística Computacional, visto que considerar concepções de palavra, como a da Fonologia Prosódica, acarretariam enormes custos computacionais, já que o *corpus* não foi desenvolvido e não possui *taggers* apropriados para esse fim.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

5.2 A Linguagem Python

As ferramentas a serem expostas a seguir foram desenvolvidas em Linguagem Python[7], a qual consiste em uma linguagem de programação de alto nível que possibilita a execução de uma série de tarefas computacionais de modo amigável, rápido e preciso. A linguagem de programação é disponibilizada gratuitamente na internet e possui uma rede de usuários que se reúne em diversas comunidades para sanar dúvidas e se ajudar mutuamente. Aplicações desenvolvidas em Python são recorrentes no campo da Linguística Computacional (cf. BIRD, KLEIN & LOPER, 2009) e isso torna a escolha da linguagem uma tarefa simples.

5.2.1 Silabificador

O silabificador utilizado trata-se de um silabificador fonológico de autoria do Prof. Dr. Marcelo Barra Ferreira[8], da Universidade de São Paulo, o qual disponibilizou a ferramenta para a execução desta pesquisa.

Apesar de ter sido confeccionado com bases fonológicas, o silabificador foi usado para a separação de palavras ortográficas com a necessidade apenas de algumas adaptações na classificação dos sons por modo de articulação.

O programa foi realizado em Linguagem Python, tendo como premissa a escala de sonoridade[9], segundo a qual as vogais, por serem os elementos mais sonoros, são os primeiros componentes a preencherem a estrutura silábica, seguida da maximização dos ataques e, por fim, se ainda restar segmentos, eles comporão a posição de coda silábica (cf. COLLISCHONN, 2010; BISOL, 2013).

A cada sílaba formada, um elemento separador era inserido (-). Para que as bordas das palavras também fossem marcadas, adaptou-se o *script* para inserir um marcador de início (&) e um de fim de palavra (*), os quais se mostram necessários para a execução de regras fonológicas que se valem de tais informações.

5.2.2 Transcritor fonológico

Consoantes

A extração de padrões fonológicos requer a codificação do *corpus* em palavras fonológicas. Como a transcrição de milhões de palavras manualmente não seria uma tarefa trivial, resolveu-se implementar um programa que, a partir de informações grafêmicas, transcreveria os *inputs* ortográficos em *outputs* fonêmicos.

O primeiro passo consistiu em converter os dígrafos, como em *colher, manhã,*

7 Disponível gratuitamente através da *Python Foundation* no site <https://www.python.org/>.

8 Ao Prof. Dr. Marcelo Barra Ferreira, ficam os agradecimentos por disponibilizar o silabificador e, principalmente, por orientar a implementação das demais ferramentas computacionais.

9 A escala de sonoridade consiste numa classificação da sonoridade dos sons das línguas, a qual é correntemente utilizada nas teorias de sílabas como meio de “correlacionar a sonoridade relativa de um segmento com a posição que ele ocupa no interior da sílaba” (COLLISCHONN, 2010, p. 109).

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

chácara, Sheila e *carro*[10], em seus respectivos símbolos fonêmicos, *co**L**er, ma**N**ã, **S**ácara, **S**eila, ca**h**o,* e em apagar o grafema *h* que não tem correspondência fonológica em início de palavra, como *hora, honesto* e *habilidade* (veja o Quadro 2).

Quadro 2: Correspondência grafema-fonema na transcrição das consoantes e glides do *Corpus* ABG.

Fonema	Transcrição	Fonema	Transcrição
/p/	P	/tʃ/	T
/t/	T	/dʒ/	D
/k/	K	/m/	m
/b/	B	/n/	n
/d/	D	/ɲ/	N
/g/	G	/x/	h
/f/	F	/ɾ/	r
/v/	V	/l/	l
/s/	S	/ʎ/	L
/z/	Z	/j/	J
/ʒ/	J	/w/	W
/ʃ/	S		

Após a substituição dos dígrafos, o processo de epêntese[11] foi implementado. Nele, vocábulos que possuíam segmentos não permitidos em coda silábica tiveram suas sílabas reparadas com a inserção da vogal epentética [ɪ] ou *y*, *ad[ɪ]vogado, internet[ɪ], [ɪ]skate*, seguindo o símbolo de codificação do Quadro 2. Esses processos foram realizados primeiro, posto que uma ressilabificação seria necessária e, principalmente, porque alguns processos fonológicos, como a palatalização[12], eram dependentes de modificações grafemáticas. A partir de então, a computação é realizada linearmente a todas as letras, caractere por caractere.

Embora a correspondência grafema-fonema na Língua Portuguesa não seja transparente, isto é, um para um, ela não é tão caótica como no inglês. Essa relação pode, na maioria dos casos, ser prevista por regras que consideram os contextos fonológicos; contudo, em outros, essa relação é ambígua e apenas sub-regras podem ser inferidas. Esse fato impossibilitou a codificação precisa de alguns grafemas, conduzindo a adoção de algumas decisões metodológicas, as quais serão explicitadas adiante.

10 Ressalta-se que o dígrafo *ss*, de *pássaro*, não foi incluído nesta regra, visto que o processo fonológico de vozeamento que atinge o segmento *s* em posição intervocálica, ca[z]a, não se aplica ao [s] que tem correspondência ortográfica *ss*.

11 "No português, a epêntese se caracteriza pela inserção de uma vogal entre as consoantes em encontros consonantais que envolvam oclusivas, africadas, nasais ou fricativas. Por exemplo, *afta* [ˈafta] ~ [ˈafita] ou *dogma* [ˈdɔgma] ~ [ˈdɔgima]. Uma vogal epentética pode também ocorrer em final de palavra, como, por exemplo em *Varig* [ˈvaɾigi]. A vogal [i] é a vogal epentética recorrente no português" (CRISTÓFARO-SILVA, 2011, p. 99-100).

12 A palatalização é um "fenômeno pelo qual uma consoante adquire uma articulação palatal ou próxima à região palatal. No português brasileiro, ocorre a palatalização de oclusivas alveolares antes da vogal alta ou **glide palatal**" (CRISTÓFARO-SILVA, 2011, p. 168, grifo da autora).

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

As letras transparentes diagnosticadas na transcrição consistem em *k, v, f, j, p* e *b, Kléber, vaca, faca, jacaré, pedra* e *bola*, as quais foram transcritas com o mesmo caractere. Embora *t* e *d* tenham sido excluídos desse grupo, a única regra que os separa consiste na palatalização diante de [i], [ɪ] ou [j], no caso, dos símbolos *i, y* e *J, [dʒi]a, a[dʒɪ]vogado, [tʃi]a* e *interne[tʃɪ].*

Além deles, os grafemas consonantais *l, r, c, z, s* e *q* também apresentam comportamentos previsíveis por regras. O primeiro pode ser codificado como glide [w] ou *W* quando em posição de coda silábica, *go[w], po[w]vo, bo[w]sa*, nos demais contextos, corresponde à lateral alveolar [l] ou *l, [l]etra, go[l]eiro, b[l]usa*. O segundo, *r*, precisa de informações silábicas para ter sua transcrição realizada, isso porque *r* é produzido em início de palavra como fricativa velar desvozeada [x] ou *h, [x]ato, [x]ei, [x]elógio*, mas como tepe alveolar [ɾ] ou *r* em segunda posição de ataque complexo, *p[ɾ]ato, b[ɾ]asileiro, c[ɾ]avo,* e em ataque silábico no meio da palavra, *co[ɾ]ação, o[ɾ]igem, mo[ɾ]ango*. Para a posição de coda silábica, adotou-se a produção de tepe alveolar [ɾ] ou *r* como a padrão para esta transcrição, *po[ɾ]ta, ca[ɾ]ta, ho[ɾ]ta*; no entanto, sabe-se que há variação dialetal (cf. OUSHIRO, 2015).

O terceiro, *c*, exibe correspondência fonológica com a fricativa alveolar desvozeada [s] ou *s* diante de [i], [e] ou [ɛ], *[s]inema, [s]ereja*, *[s]ego,* e com a oclusiva velar desvozeada [k] diante de [a], [o], [ɔ] e [u], *[k]arruagem, [k]oragem, [k]ócegas, [k]urativo.* O quarto, *z*, apresenta-se como fricativa alveolar vozeada [z] ou *z* em ataque silábico, *[z]ebra, de[z]embro*, e quando seguida de consoantes vozeadas, *feli[z]mente, vora[z]mente,* caso contrário, é produzida como fricativa alveolar desvozeada [s] ou *s*, i. e., em final de palavra, *feli[s], vora[s].* Por fim, o grafema *s* é codificado como fricativa alveolar vozeada [z] ou *z* em posição intervocálica, *ca[z]a, repre[z]entar*, como fricativa alveolar desvozeada [s] ou *s* quando precedida por *s, pá[s]aro, proce[s]o*, e apagada quando precede *c, adolecente, conciência.* Por fim, o quinto, *q*, forma dígrafo com a vogal alta posterior [u], resultando em [q^{w}] se possuir a vogal baixa [a] como segmento subsequente, *[q^{w}]adro, [q^{w}]adrado*; todavia, nos demais contextos, i. e., o dígrafo *qu* seguido de *e* e *i* é codificado como oclusiva velar desvozeada [k], *[k]ente, [k]iabo.*

As consoantes nasais [m] e [n] foram tratadas separadamente, visto que o traço de nasalidade pode se espraiar para os segmentos ao seu redor. A primeira especificidade consiste nas ocorrências em posição de coda silábica, as quais, quando em meio de palavra, nasalizam a vogal precedente, como em *[ẽ]prego, [ẽ]bulatório*, quando em final de palavra, também nasalizam a vogal precedente e, ainda, fazem emergir um glide nasal, *com[ɐ̃w̃], ho[ẽj̃], híf[ẽj̃].* A segunda característica trata-se da manifestação das consoantes nasais [m] e [n] em posição de ataque silábico, *[m]açã, [n]avio, ca[m]elo, da[n]o[n]e.* A simples ocorrência em posição de ataque não traz ambiguidade para esta transcrição, porém, elas nasalizam as vogais em seu entorno apenas em alguns contextos, *ba[nɐ̃]na, [mɐ̃]no.* Essa aparente nasalização da vogal em posição tônica não é regular como em *canela, panela, flanela*, as quais não são produzidas com nasalidade

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

nas sílabas tônicas; contudo, são cabíveis de produções diferentes na sílaba pré-tônica: com ou sem nasalidade. Em decorrência da (aparente) maior proporção de não nasalização em tal contexto, determinou-se que essas vogais não seriam transcritas como nasais, já que a produção de nasalidade é obrigatória apenas em posição tônica e opcional em átona.

De modo semelhante, outro caso não totalmente previsível consiste no grafema *g*, o qual se manifesta como fricativa pós-alveolar vozeada [ʒ] ou *j* diante de [e], [ɛ] e [i], *[ʒ]elo, [ʒ]esto, [ʒ]inástica*, e como oclusiva velar vozeada [g] ou *g* diante de [a], [o], [ɔ] e [u], *[g]alinha, [g]ol, [g]ola, [g]ula.* No entanto, o último contexto é ambíguo se seguido por mais uma vogal, podendo formar uma consoante complexa como [g^{w}], *lin[g^{w}]iça, lin[g^{w}]ística,* ou ser expressa através da oclusiva velar vozeada [g], *[g]erra, [g]itarra.* Novamente, a ausência de uma regra fez com que uma das transcrições fosse assumida como *default*, a qual consiste na codificação de [g].

Sem dúvida, o grafema que exibe mais ambiguidade, i. e., menos transparência entre a forma grafemática e a fonêmica trata-se de *x*. Esse é produzido, na maioria dos casos, como [kɪs] quando em final de palavra, *láte[kɪs], tóra[kɪs];* como fricativa pós-alveolar desvozeada [ʃ] em início de palavra, *[ʃ]adrez, [ʃ]ícara*; como fricativa alveolar desvozeada [s] se precedido do fonema [s], formando os dois segmentos um só fonema [s], *e[s]eção, e[s]elente*; como fricativa alveolar desvozeada [s] diante de outros fonemas consonantais, *e[s]tinguir, e[s]clamação;* [kɪs] quando antecede *ão* ou *ia, cone[kɪs]ão, anore[kɪs]ia*[13]; como fricativa alveolar vozeada [z] entre vogais, sendo que a primeira deve ser a vogal anterior média alta [e], *e[z]ame, e[z]ercícios.* Já os casos restantes parecem não apresentar qualquer regularidade, *má[s]imo, ne[kɪs]o, o[kɪs]ítono, le[kɪs]ema, [ʃ]u[ʃ]a*, etc, sendo transcritos como *x*. Essa foi a forma encontrada para marcar diferentemente tais ocorrências que não são previstas por regras, assumindo-as como transcrições imprecisas, as quais devem ser tratadas com cautela.

O último grafema a ser codificado foi *ç*, visto que, apesar da sua regularidade, a transcrição precoce desse segmento como fricativa alveolar desvozeada [s] acarretaria na aplicação da regra de vozeamento intervocálica, a qual não se aplica em lexemas que tenham o grafema *ç, cora[s]ão, cacha[s]a.*

Vogais

As vogais, que não sofreram a codificação de nasalização apresentada anteriormente, foram transcritas com sua correspondente: se vogal oral, como oral; se nasal, como nasal, seguindo o Quadro 3.

13 Ressalta-se que essa sub-regra foi corrigida manualmente, posto que há vocábulos como *paixão, lixão, caixão* que não se enquadram nela.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

Quadro 3: Correspondência grafema-fonema na transcrição das vogais do *Corpus* ABG.

Vogais Orais		**Vogais Nasais**	
/a/	a	/ã/	A
/e/	e	/ẽ/	E
/ɛ/	3		
/i/	i	/ĩ/	I
/o/	o	/õ/	O
/ɔ/	0		
/u/	u	/ũ/	U
/ɪ/	y		
/ʊ/	w		
/ɐ/	@		

Para as vogais, fez-se necessária a elaboração de algoritmos que previam o comportamento dos ditongos. Determinou-se, então, que, após a localização das vogais *i* e *u* – únicos segmentos que podem ser glides no português –, os itens que os circundam deveriam ser observados. Se houvesse encontro vocálico com formação de ditongo, o *i* seria transcrito como [j] ou *J*, *p[ja]da, r[ej]*, e o *u* como [w] ou *W, árd[wo], l[ow]sa.* Todavia, se ambos coocorressem, o elemento em primeira posição comporia o núcleo da sílaba e o segundo se comportaria como glide, ramificando o núcleo, *grat[uj]to.* Além disso, nos ditongos com um segmento nasal, as vogais nasais deveriam espraiar a sua nasalidade para a vogal seguinte, resultando em *conclus[ẽw̃], m[ẽj̃].*

5.2.3 Categorizador morfológico e lematizador

O *tagger* utilizado para marcar as categorias morfológicas das palavras que compõem o *corpus* foi o *TreeTagger* (SCHMID, 2015), o qual consiste numa ferramenta de anotação de texto que fornece ao vocábulo a categoria morfológica e o lema, desenvolvido por Helmut Schmid da *University of Stuttgart*.

O algoritmo foi construído com base nas Cadeias de Markov de primeira ordem, para o qual “the next state (tag) depends only on the k preceding states (tag)” (SCHMID, 1995, p. 2), ou seja, as etiquetas são atribuídas baseadas numa janela de contexto que permite o algoritmo desambiguar as palavras que possam ter mais de uma categoria morfossintática. Para mais informações a respeito de como o categorizador foi construído, consulte Schmid (1994; 1995).

O *TreeTagger* possui etiquetador para diversas línguas, tais como inglês, espanhol, português, russo, alemão etc. A versão de etiquetagem para o português foi realizada pelo projeto FreeLing (2015), o qual emprega as seguintes categorias:

ADJ - adjetivo
ADV - advérbio
ADV + V - advérbio + verbo

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

CONJ - conjunção
DET - determinante
F - estrangeirismo
G - sigla
I - interjeição
NOM - substantivo
NUM - numeral
P - pronome
PR - pronome relativo
PRP – preposição
PRP + ADV – preposição + advérbio
PRP + DET - preposição + determinante
PRP + P - preposição + pronome
V - verbo
V + P – verbo + pronome

Além delas, precisou-se acrescentar uma categoria que contemplasse os compostos, posto que eles possuem um comportamento peculiar, principalmente no âmbito acentual. A etiqueta atribuída foi C, sendo a única a ser rotulada manualmente.

A segunda informação fornecida pelo *TreeTagger* foi o lema das palavras, o qual varia de palavra para palavra, correspondendo à forma canônica do vocábulo ou à sua entrada de dicionário.

5.2.4 Acentuador

O programa denominado de acentuador ABG tem como função receber uma palavra fonológica silabificada e retornar como *output* a palavra fonológica silabificada e acentuada.

O primeiro mecanismo adotado para marcar a localização acentual se valeu das regras de acentuação gráfica, para as quais o acento marcado é codificado por meio do acento gráfico, a saber, agudo ou circunflexo.

Basicamente, o *script* verifica se o *input* (palavra ortográfica) possui acento gráfico. Se houver, a vogal acentuada é substituída pelo símbolo que codifica o grafema no sistema de transcrição fonológico adotado (veja Quadro 4). Essa busca permitiu que todas as palavras proparoxítonas, paroxítonas com sílaba final pesada, as oxítonas com sílaba final leve, os hiatos *i* e *u*, dentre outros casos tivessem a marcação acentual facilmente detectada.

Quadro 4: Correspondência grafema-fonema na transcrição de vogais tônicas do *Corpus* ABG.

Vogal	**Transcrição**
/a/	1

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

/ã/	2
/e/	4
/ɛ/	5
/ɛ̃/	6
/i/	7
/ĩ/	8
/o/	9
/ɔ/	!
/õ/	#
/u/	$
/ũ/	%

Caso não houvesse acento gráfico no *input*, com exceção de *i* e *u* que em final de palavra são tônicos, o algoritmo deveria olhar para a classe morfológica da palavra e aplicar uma regra de acentuação baseada na proposta de peso silábico de Bisol (1994)[14]. A detecção da composição das sílabas, se pesadas ou leves, olhava para a estrutura da sílaba, como será esboçado na seção adiante. Isso se deve à necessidade de codificar diferentemente o segmento *s* em final de palavra, já que, segundo a autora, ele seria invisível às regras acentuais. A acentuação ocorre pela troca do símbolo que codifica o fonema de átono para tônico.

Após a marcação do acento de todos os vocábulos, as vogais átonas em final de palavra foram submetidas a um processo de redução vocálica, transformando-as em [ɐ] @, [ɪ] *y* e [ʊ] *w*, respectivamente*, a, e/i* e *o/u*. Apesar da aparente facilidade encontrada na transcrição das vogais, a diferenciação entre as vogais médias abertas e fechadas fora de posição tônica, como ocorre em *b[ɛ]líssima, b[ɔ]linha*, não é cabível de regras, por isso, elas foram transcritas como vogal média alta, [ɛ] e [ɔ].

5.2.5 Codificando estruturas fonológicas

Além das informações fonológicas e morfológicas descritas acima, resolveu-se acrescentar outras categorias fonológicas que pudessem ser relevantes para a busca de dados e para a extração de padrões. Para tal, fez-se um decodificador que recebesse um fonema e retornasse uma categoria fonológica segmental, a qual foi classificada como V (vogal), C (consoante), G (glide) e S (fricativa alveolar desvozeada [s] em final de palavra).

A separação de *s* nesse contexto tem como premissa as postulações realizadas por alguns teóricos, como Bisol (1994) e Hermans & Wetzels (2012), para os quais o português seria sensível ao peso silábico em posição final, contudo, esse segmento seria invisível a tal regra, não contribuindo com o peso silábico. Apesar de a presente pesquisa

14 Embora a hipótese métrica não preveja adequadamente a posição em que o acento recairá, adotou-se tal proposta como base para a codificação das sílabas tônicas, visto que as palavras com acento marcado já foram codificadas na primeira fase da acentuação, a qual olhou para o acento ortográfico, restando apenas os casos previstos pelas regras de peso silábico.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

não seguir qualquer proposta acentual, utilizou-se essa codificação para que elas fossem base para a realização do acentuador.

Outra informação que se mostrou relevante consiste na classificação do acento empregado nas palavras, i. e., a categorização em oxítona, paroxítona e proparoxítona, a qual também foi realizada de forma automática a partir da localização da posição tônica da palavra.

5.2.6 Contando a frequência

Como salientado anteriormente, a decisão de compilar um *corpus* deveu-se à necessidade de extrair padrões por meio do uso linguístico. Para que eles pudessem ser extraídos, a contabilização da frequência consistiu num dos métodos utilizados para tal.

As medidas de frequência das palavras foram feitas de dois modos. Primeiro, para determinados objetivos, é interessante saber apenas se a palavra ocorreu ou não no *corpus.* Nestes casos, o que foi utilizado é a medida de tipos, em que se uma palavra ocorre mais de uma vez num *corpus,* ela só é contada uma vez, desta maneira, uma palavra comum como *também* tem o mesmo peso que uma palavra rara como *pseudópode*.

No entanto, não é necessariamente verdade que as palavras raras e comuns do idioma tenham a mesma influência no comportamento da língua. Autores como Bybee (2001) apontam que o uso é fundamental para se compreender o funcionamento da fonologia de um idioma. Dessa forma, a segunda medida de frequência se dá na contagem de ocorrências, em que os termos encontrados são associados ao valor de uma contagem em todos os textos que compuseram o *corpus*. Neste caso, *também* obteria um peso muito maior do que *pseudópode*. Além disso, por existir uma divisão já pronta no conjunto de dados entre o *corpus* escrito e o oral, são também fornecidas no *Corpus* ABG as contagens de ocorrências das palavras em cada um dos subconjuntos.

A simples somatória da frequência de ocorrência não é suficiente para dizer o que é frequente ou infrequente dentro do *corpus*, uma vez que há uma quantidade pequena de palavras muito frequentes na língua, as quais incluem as palavras funcionais (preposição, pronomes, artigos), em contrapartida, uma quantidade enorme de palavras que ocorrem poucas vezes. Métodos que consideram essa distribuição são correntemente apontados na literatura como mais adequados para a análise de frequência.

5.3 Lei de Zipf

O Modelo de Zipf, ou curva de Zipf, “indica que a probabilidade de ocorrência de amostras é determinada por uma relação de poder da lei e pela posição de cada amostra” (ARAÚJO, 2013, p. 97)[15]. A predição é que um vocábulo mais frequente ocorrerá o dobro de vezes do segundo; o terceiro terá um terço das ocorrências do primeiro e assim sucessivamente. A curva que é estabelecida nesse modelo pode ser verificada na Figura 1.

15 Trecho Original: “states that the probability of occurrence of samples is determined by a power law relation and the rank of each sample”.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

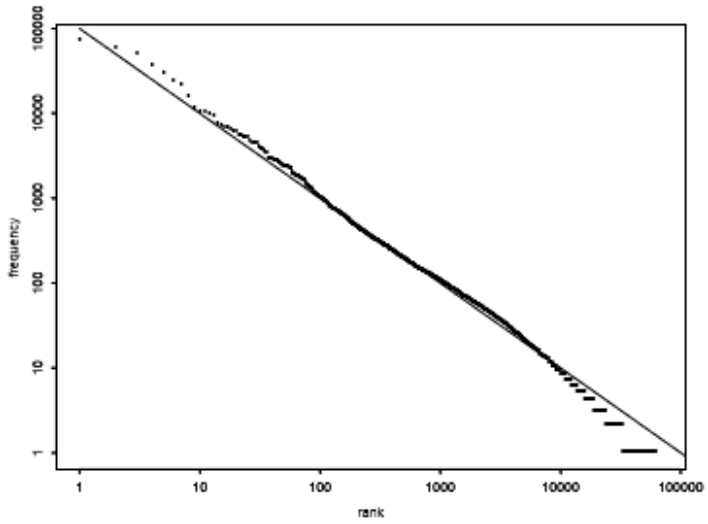

Figura 1: Lei de Zipf.
Fonte: Manning & Schütze (1999, p. 26).

A Figura 1 representaria basicamente a curva da frequência das palavras no léxico das línguas: poucas palavras ocorrem muitas vezes, um número médio de palavras tem média frequência e a maioria das palavras ocorre poucas vezes. Tal tendência pode ser verificada no *Corpus* ABG, o qual possui apenas 45 palavras – em geral, funcionais – que ocorreram mais de 10.000 vezes em um *corpus* com, aproximadamente, 93.000 tipos de palavras e 3.600.000 ocorrências de palavras. Para estabelecer o valor das palavras, é necessário calcular as suas constantes, como em (3):

(3) k ∞ 1 / r
Há uma constante *k* tal que $f \cdot r = k$
(extraído de MANNING & SCHÜTZE, 1999, p. 24)

Onde *k* é a constante, *f* é a frequência da palavra (frequência de ocorrência) e *r* é a posição na lista (posição no *ranking* das palavras pela frequência no *corpus*) (cf. MANNING & SCHÜTZE, 1999; ARAÚJO, 2013), ou seja, o cálculo consiste em $f \cdot r$. Esse cálculo é empregado para se descrever a distribuição de frequência das palavras.

Mais do que determinar a probabilidade de dada palavra no *corpus*, quer-se estabelecer níveis de frequência em que as palavras possam ser categorizadas e assumidas como tais. Ao tomar a relação exponencial que o Modelo de Zipf propõe, definiram-se cinco níveis de divisão do *Corpus* ABG, como pode ser observado em (4):

(4) Muito Frequente: + 10.000 ocorrências –> 46 palavras
Alta Frequência: 10.000 - 1.000 ocorrências –> 319 palavras
Média Frequência: 1.000 - 100 ocorrências –> 2.773 palavras
Baixa Frequência: 100 - 2 ocorrências –> 49.894 palavras
Pouco Frequente: 1 ocorrência –> 39.570 palavras

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

A divisão em níveis de frequência consiste numa decisão metodológica para facilitar a organização e o entendimento dos níveis de frequência. Semelhante processo será realizado na extração de outros padrões fonológicos, como na contagem de sílabas a serem apresentados em trabalhos futuros.

5.4 Organização do *output* do *Corpus* ABG

Todas as informações descritas acima compõem o banco de dados ABG, as quais foram separadas por vírgulas, constituindo assim um tipo de arquivo de tabela no formato '.csv' (*Comma Separated Value*) que permite a manipulação ágil e simples tanto em programas de edição de planilhas (como o Calc do LibreOffice) quanto diretamente nas interfaces com linguagens de programação (como o Python e o R).

O *output* do *Corpus* do ABG consiste em (5):

(5) [índice_palavra; palavra_ortográfica; categoria_morfológica; lema; transcrição_fonológica; transcrição_fonológica_acentuada; estrutura_silábica; categoria_acentual; frequência_geral; frequência_fala; frequência_escrita; nível_frequência]

A título de exemplo, veja a Figura 2:

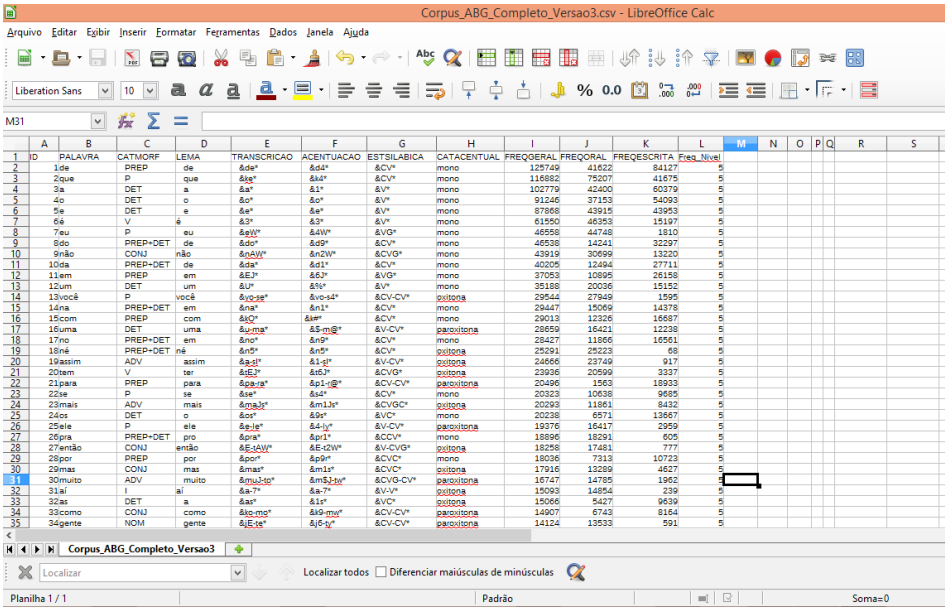

Figura 2: Exemplo de *output* do *Corpus* ABG.

Na planilha, é possível realizar buscas pelos filtros que o próprio Excel fornece.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

6 Dados quantitativos do *Corpus* ABG

O *Corpus* ABG é composto por 3.616.625 ocorrências de palavras e por 92.602 tipos de palavras, sendo que 1.938.805 ocorrências são provenientes dos *corpora* de fala e 1.676.820 ocorrências dos *corpora* escritos (veja a Tabela 1).

Tabela 1: Quantidade de ocorrência e de tipo por *corpus*.

Corpus	**Quantidade de Ocorrências**	**Quantidade de Tipos**
Corpus de Fala	1.938.805	36.484
Corpus Escrito	1.676.820	77.669
Corpus Geral	3.615.625	92.602

Embora o *corpus* de fala possua mais ocorrências do que o de escrita, como pode ser verificado na Tabela 1, o léxico é menos diversificado. Essa diferença exorbitante deve-se, sobretudo, a uma característica estilística da escrita, a qual tem de evitar a repetição, controle difícil de ser realizado em situações de comunicação oral. Além disso, os textos escritos possuem significativamente mais palavras estrangeiras do que as produções faladas, já que os textos científicos (artigos) contêm *abstracts* e o léxico utilizado em sites de computação e games tende a ter um alto número de vocábulos estrangeiros.

Outro dado revelador consiste na observação da distribuição desse léxico em classes gramaticais, conforme Tabela 2. Como inferido, os substantivos, adjetivos e verbos têm maior diversidade léxica do que as demais classes, muito embora os determinantes e as preposições sejam mais usados[16].

Tabela 2: Quantidade de tipos de palavras por classe gramatical.

Classe Gramatical	**Quantidade de Tipos**
Substantivo	37.048
Verbo	32.601
Adjetivo	11.847
Estrangeirismo	5.146
Verbo + Pronome	2.197
Compostos	1.486
Advérbio	1.101
Sigla	444
Pronome	218

16 O *tagger* utilizado para realizar a marcação morfológica do *Corpus*, o *TreeTagger*, possui um desempenho mediano para o português. Apesar disso, resolveu-se utilizá-lo a fim de minimizar o trabalho manual que uma morfologização requereria, realizando-se somente uma revisão as marcações feitas.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

Interjeição	183
Preposição + Pronome	91
Conjunção	69
Preposição	45
Numeral	44
Preposição + Determinante	40
Determinante	38
Preposição + Advérbio	5

Como o objeto teórico da compilação do *corpus* consistiu na extração de dados acentuais do Português Brasileiro, considerando para isso os padrões que emergem na língua, o primeiro dado a ser observado consiste na distribuição da tonicidade por sua categoria acentual, ou seja, monossílabo, oxítono, paroxítono, proparoxítono e acento na sílaba pré-antepenúltima (veja a Tabela 3). Embora não seja usual, ele refere-se a casos em que a epêntese da vogal [ɪ] cria uma nova sílaba no vocábulo, fato que desloca o acento para a quarta sílaba a partir da margem direita da palavra, como em *té[kɪ]nico, cá[pɪ]sula.*

Tabela 3: Quantidade de tipos de palavras por categoria acentual.

Categoria Acentual	**Quantidade de Tipos**
Quarta Sílaba	40
Proparoxítonos	3.797
Paroxítonos	65.511
Oxítonos	24.714
Monossílabos	1.525

A distribuição acentual do *Corpus* ABG assemelha-se em partes à proporção encontrada por Viaro & Guimarães-Filho (2007) (veja a Tabela 4). A principal diferença diagnosticada entre esses *corpora* consiste nos dados referentes às proparoxítonas. Contudo, essa diferença, provavelmente, é motivada pela quantidade de dados analisada e pelo tipo de *corpus*, visto que o de Viaro & Guimarães-Filho (2007) é composto de 150.875 palavras provenientes de textos escritos, os quais pertencem ao Dicionário Houaiss da Língua Portuguesa, enquanto o *Corpus* ABG tem mais de 90.000 tipos de palavras advindos de textos escritos e falados.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

Tabela 4: Distribuição percentual do acento por categoria em Viaro & Guimarães-Filho (2007, p. 31) e no *Corpus* ABG.

	Viaro & Guimarães-Filho (2007)	***Corpus* ABG**
Monossílabos	0.3% a 0.5%	1.6%
Oxítonos	25%	25.9%
Paroxítonos	62%	68.5%
Proparoxítonos	12%	4%
Pré-antepenúltimo	0.4%	0.04%

Os dados apresentados na Tabela 4 corroboram as postulações teóricas de que o acento paroxítono seria o padrão acentual do português em decorrência de sua predominância no léxico da língua e de sua produtividade. Por conseguinte, o segundo padrão seria o oxítono.

Além dos padrões acentuais, outros dados fonológicos podem ser extraídos do *Corpus* ABG, tais como estrutura de palavra por categoria acentual, estrutura silábica, tipos silábicos por tonicidade. Pode-se, ainda, contabilizar em termos de frequência as vogais mais frequentes em dada posição silábica, a quantidade de vocábulos com certa sequência sonora, dentre outras informações possíveis. A possibilidade de tecer buscas diversificadas consistiu no motivo central da compilação do *corpus,* o qual requer, em alguns casos, apenas o manuseio de planilhas e, em outros casos, a criação de *scripts* para buscas específicas.

O *Corpus* ABG encontra-se disponível para todos os pesquisadores que se interessarem e necessitarem de dados quantificados de padrões fonológicos do português brasileiro. Pretende-se, nos próximos trabalhos, apresentar mais dados e quantificações que já são possíveis de serem extraídos e utilizados do referido *corpus.*

7 Conclusão

Este artigo objetivou apresentar uma nova e útil ferramenta para os estudos dos padrões fonológicos do português brasileiro: o *Corpus* ABG. As informações necessárias para a realização dos estudos de Guide (2016) e Benevides (2017) apresentaram desafios metodológicos (e braçais) que instigaram o desenvolvimento de diversas ferramentas computacionais bastante interessantes a fim de fornecer um tratamento adequado aos dados de maneira rápida.

Apesar de o *Corpus* ABG ter tamanho médio (cf. SARDINHA, 2000b), ele possui informações quantificadas que permitiram investigações robustas sobre o funcionamento do acento no PB, bem como fornece outras informações fonológicas que podem servir de base para o desenvolvimento de outros estudos. A disponibilização do *corpus* e das ferramentas utilizadas no seu desenvolvimento por meio deste artigo é um convite a outros pesquisadores para utilizar esses dados em suas pesquisas.

Ressalta-se que diversas análises investigando diferentes relações entre as variáveis observadas no *corpus* podem ser realizadas. Trata-se de um *corpus* que ainda

Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

não foi explorado à exaustão, pois apresenta uma riqueza de dados que pode contribuir para diversas discussões no âmbito morfológico e fonológico do idioma.

Ademais, o conjunto de dados pode servir como substrato para diversos modelos probabilísticos de língua que visem investigar possíveis aspectos do funcionamento do sistema linguístico. Em Guide (2016), por exemplo, dois tipos de modelos probabilísticos foram implementados para investigar a questão do acento baseando suas probabilidades nos dados quantitativos deste *corpus*, outros modelos e outras questões teóricas podem ser investigadas se valendo deste mesmo conjunto de dados.

Referências

ARAÚJO, L. C. *Statistical Analyses in Language Usage.* 2013. 199 f. Tese (Doutorado em Engenharia Elétrica) - Escola de Engenharia, Universidade Federal de Minas Gerais, Belo Horizonte, 2013.

ASPA. *Projeto ASPA*: Avaliação Sonora do Português Atual. Disponível em: <http://www.projetoaspa.org/buscador2>. Acesso em: 19 dez. 2014.

BARBOSA, J. M. *Introdução ao Estudo da Fonologia e Morfologia do Português*. Coimbra: Livraria Almedina, 1994. 295 p.

BENEVIDES, A. L. de. *O acento primário em pseudopalavras*: uma abordagem experimental. 2017. 135 f. Dissertação (Mestrado) - Faculdade de Filosofia, Letras e Ciências Humanas, Departamento de Linguística. Universidade de São Paulo, São Paulo, 2017.

BIBER, D. Representativeness in Corpus Design. *Literaty and Linguistic Computing*, New York, v. 8, n. 4, p. 243-257, 1993.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python.* O'Reilly Media, 2009.

BISOL, L. O acento e o pé métrico. *Letras de Hoje,* Porto Alegre, v. 29, n. 4, p. 25-36, dez. 1994.

BISOL, L. A sílaba e seus constituintes. In: ABAURRE, M. B. M. (Org.). *A construção fonológica da palavra.* São Paulo: Contexto, 2013. v. III. p. 21-52.

BYBEE, J. *Phonology and Language Use.* Cambridge: Cambridge University Press, 2001. 238 p.

COLLISCHONN, G. A sílaba em português. In: BISOL, L. (Org.) *Introdução a estudos de fonoloia do português brasileiro. Porto* Alegre: EDIPUCRS, 2010. 5 ed. 286 p.

CRISTÓFARO-SILVA, T. *Dicionário de Fonética e Fonologia*. São Paulo: Contexto, 2011.

Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

239 p.

FREELING. *Etiquetas Eagles* (v. 2.0). Disponível em: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Portuguese-Tagset.html>. Acesso em: 15 abr. 2015.

FROTA et al. *FrePOP: Frequency Patterns of Phonological Objects in Portuguese:* Research and Applications. Laboratório de Fonética (CLUL), Faculdade de Letras da Universidade de Lisboa. 2010. Disponível em: <http://frepop.letras.ulisboa.pt/>. Acesso em: 15 out. 2015.

GALVES, C.; FARIA, P. *Tycho Brahe Parsed Corpus of Historical Portuguese.* 2010. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html>. Acesso em: 14 out. 2015.

GONÇALVES, S. C. L. *Banco de dados Iboruna*: amostras eletrônicas do português falado no interior paulista. Disponível em: <http://www.iboruna.ibilce.unesp.br/>. Acesso em: 08 out. 2014.

GUIDE, B. F. *Abordagem computacional para a questão do acento no português brasileiro.* 2016. 113 f. Dissertação (Mestrado) - Faculdade de Filosofia, Letras e Ciências Humanas, Departamento de Linguística, Universidade de São Paulo, São Paulo, 2016.

HERMANS, B.; WETZELS, W. L. Productive and unproductive stress patterns in Brazilian Portuguese. *Letras & Letras,* Uberlândia – MG, v. 28, n. 1, p. 77 – 11, jan./jun. 2012.

LINGUATECA. *Linguateca.* Disponível em: <http://www.linguateca.pt/>. Acesso em: 14 out. 2015.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing.* Londres: The MIT Press, 1999. 720 p.

MCENERY, T.; WILSON, A. *Corpus Linguistics*: an introduction. Edinburgh: Edinburgh University Press, 2001. 2 ed. 235 p.

MENDES, R. B. *Projeto SP2010*: Amostra da fala paulistana. 2010. Disponível em: <http://projetosp2010.fflch.usp.br>. Acesso em: 30 mar. 2015.

MENDES, R. B.; OUSHIRO, L. O paulistano no mapa sociolinguístico brasileiro. *Alfa*, Araraquara, v. 56, n. 3, p. 973-1001, 2012.

OUSHIRO, L. *Identidade na pluralidade*: avaliação, produção e percepção linguística na cidade de São Paulo. 2015. 372 f. Tese (Doutorado em Letras) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2015.

RASO, T.; MELLO, H. (Org.) *C-ORAL-BRASIL*: Corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012. 332 p.

Belo Horizonte, v. 10, n. 1, p. 139-163, jan.-jun. 2017 – ISSN 1983-3652
DOI: 10.17851/1983-3652.10.1.139-163

SARDINHA, T. B. Linguística de Corpus: Histórico e Problemática. *DELTA*, v. 16, n. 2, 2000a. p. 323-67.

SARDINHA, T. B. *O que é um corpus representativo?* Direct Papers 44. 2000b. Disponível em: <http://www.direct.f2s.com>. Acesso em: 02 set. 2014.

SCHMID, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. 1994. p. 1-9.

SCHMID, H. Improvements in part-of-specch tagging with a application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland, 1995, p. 1-9.

SCHMID, H. *TreeTagger*: a language independent part-of-speech tagger. 2015. Disponível em: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>. Acesso em: 15 abr. 2015.

VIARO, M. E.; GUIMARÃES-FILHO, Z. O. Análise quantitativa da freqüência dos fonemas e estruturas silábicas portuguesas. *Estudos Linguísticos*, São Paulo, XXXVI (1), p. 27-36, jan.-abr. 2007.

Recebido em 06 de março de 2017.
Aprovado em 01 de maio de 2017.