Revista de Estudos da Linguagem, v. 26, n. 2, p. 769-792, 2018

Análise de textos enciclopédicos da *Simple English Wikipedia* e da *Wikipedia*: algumas discussões para o ensino de língua inglesa

Analysis of encyclopedic texts from Simple English Wikipedia and Wikipedia: some discussions for English language teaching

Eduardo Batista da Silva
Universidade Estadual de Goiás, Morrinhos, Goiás / Brasil
eduardo.silva@ueg.br

Resumo: Tomamos como objeto de pesquisa o conteúdo lexical do texto enciclopédico, mais precisamente o perfil lexical de textos presentes em duas enciclopédias colaborativas: uma destinada a aprendizes de língua inglesa (*Simple English Wikipedia*) e outra destinada a um público falante nativo de língua inglesa (*Wikipedia*). Nosso objetivo geral é apresentar o texto enciclopédico como um recurso didático para o enriquecimento e prática de vocabulário em língua inglesa. Os objetivos específicos são os seguintes: 1) proceder uma análise do perfil lexical de artigos da *Simple English Wikipedia* e da *Wikipedia*; 2) comparar os artigos nas duas enciclopédias e 3) checar se os artigos adaptados da enciclopédia destinada aos aprendizes realmente empregam vocabulário mais elementar. O embasamento teórico recorre aos estudos de Lexicologia (NATION, 2001, 2003, 2015) e da Linguística de Córpus (BERBER SARDINHA, 2004, 2012). Com relação à metodologia, os 35 melhores artigos da *Simple English Wikipedia*, na opinião do editor do site, foram convertidos no formato texto simples e posteriormente analisados pelo *software VocabProfile*, versão 4, um programa on-line que divide um texto em faixas de frequência lexical. Após o processamento dos arquivos, o *VocabProfile* verificou o perfil lexical dos textos enciclopédicos. Os resultados indicam que, do ponto de vista léxico-quantitativo, não há

eISSN: 2237-2083
DOI: 10.17851/2237-2083.26.2.769-792

diferença significativa entre o perfil lexical da *Simple English Wikipedia* e da *Wikipedia*. As duas enciclopédias se diferenciam primordialmente pela extensão dos artigos.

Palavras-chave: vocabulário; língua inglesa; *Wikipedia*; *Simple English Wikipedia*.

Abstract: We tackle the lexical content of encyclopedic texts as our research object, more precisely, the lexical profile of texts in two collaborative encyclopedias: one is designed for English language learners (Simple English Wikipedia) and the other is designed for an English-speaking audience (Wikipedia). We aim at introducing the encyclopedic text as a pedagogical resource for the enhancement and practice of vocabulary in English. Our specific goals are the following: 1) proceed an analysis of the lexical profile of texts from Simple English Wikipedia and Wikipedia; 2) compare the texts of both encyclopedias and 3) check whether the adapted texts from the encyclopedia designed for English language learners do employ more simple vocabulary. The theoretical background resorts to studies concerning Lexicology (NATION, 2001, 2003, 2015) and Corpus Linguistics (BERBER SARDINHA, 2004, 2012). Regarding methodology, the 35 best articles from Simple English Wikipedia, in the editor´s opinion, were converted to simple text format and later analyzed by the software VocabProfile, version 4, an online software that divides a text into frequency bands. After processing the files, VocabProfile verified the lexical profile of the encyclopedic texts. The findings show that, from a lexicoquantitative perspective, there is no significant difference between the lexical profile in Simple English Wikipedia and Wikipedia. Both encyclopedias primarily differ in terms of article length.

Keywords: vocabulary; english language; Wikipedia; Simple English Wikipedia.

Recebido em 8 de setembro de 2017

Aceito em 17 de outubro de 2017

1 Introdução

Este trabalho faz parte de um projeto de pesquisa em andamento na Universidade Estadual de Goiás intitulado "Estudos em Lexicologia e Linguística de Córpus para o Professor de Língua Inglesa". O presente estudo toma como objeto de pesquisa o conteúdo lexical do texto enciclopédico, mais precisamente o perfil lexical de textos presentes em duas enciclopédias colaborativas: uma destinada a aprendizes de língua inglesa (*Simple English Wikipedia*) e outra destinada a um público falante nativo de língua inglesa (*Wikipedia*). *Wikipedias* são lugares nos quais as pessoas trabalham em equipe para escrever enciclopédias em diferentes línguas.

Para contextualizar o instrumento no qual se insere nosso objeto de pesquisa, apresentamos, nos próximos parágrafos, as características das duas enciclopédias e, na sequência, detemo-nos na *Simple English Wikipedia*.

Desde a sua criação, no ano de 2001, a *Wikipedia* vem sendo utilizada como uma obra de consulta gratuita, de fácil acesso, destacando-se pelo quesito confiabilidade, na maioria das vezes, e atraindo milhões de visitantes diariamente.

No que concerne à utilização dos artigos escritos de forma colaborativa na prática de leitura em língua inglesa, vale ressaltar que, além da *Wikipedia*, existe outra enciclopédia colaborativa chamada *Simple English Wikipedia* (doravante, SEW), criada no ano de 2004, que se propõe a ser um recurso informacional destinado a um público que inclui estudantes, crianças, adultos com dificuldades de aprendizagem ou de leitura e aprendizes de inglês. Outrossim, outras pessoas usam-na graças à linguagem simples, o que possibilita o conhecimento de conceitos com os quais não têm familiaridade.

No final do mês de agosto de 2017, a SEW contava com 127.167 artigos em seu banco de dados (SIMPLE ENGLISH WIKIPEDIA, 2017a). Em comparação, a *Wikipedia* em língua inglesa, na mesma época, possuía 5.468.078 artigos (WIKIPEDIA, 2017a).

No tocante ao léxico empregado, na SEW, opta-se pela utilização de palavras simples da língua inglesa, acompanhada de estruturas gramaticais também mais simples. Ao preocupar-se com a qualidade do conteúdo léxico-gramatical de seus artigos, uma equipe de editores avalia todos os novos artigos ou suas atualizações. Na redação dos textos, existe uma preocupação em usar um repertório lexical mais básico e

frases mais curtas, com o intuito de tornar a leitura dos aprendizes mais fácil. Os colaboradores são estimulados a expandir os artigos, adicionando detalhes e adotando o vocabulário básico, sem a premissa de que os textos criados devam necessariamente ser curtos. Com base na orientação de que apenas 2.000 palavras são suficientes para se escrever um bom artigo (SIMPLE ENGLISH WIKIPEDIA, 2016), não fica claro se esse quantitativo relaciona-se à quantidade de palavras consideradas isoladamente, sem repetições (*types*) ou todas as ocorrências de palavras no artigo (*tokens*).

Para os autores dos textos, sugere-se que tomem como parâmetro e procurem as palavras em quatro grandes listas de palavras (SIMPLE ENGLISH WIKIPEDIA, 2016), a saber: *Basic English 850* (*"Basic"*é um acrônimo para *British American Scientific International Commercial*). Essa lista de palavras foi criada por Charles Kay Ogden em 1935. Trata-se de uma tentativa de explicar conceitos considerados complexos com 850 palavras básicas do inglês: são 100 palavras denominadas "*operations*", 400 palavras na categoria "*things*", 100 "*general*", 200 "*picturable words*" e 50 "*opposites*"); *Basic English 1500* (uma lista mais avançada que a *Basic English* 850, que contém, na verdade, mais de 2.600 palavras, constituída das 850 palavras da *Basic English*; 179 palavras internacionais; 50 substantivos internacionais; 12 nomes de áreas científicas; 50 palavras sobre o tempo e números, entre outros); *Voice of America Special English Word Book* (lista que contém 1.580 palavras com 6 categorias gramaticais, 8 termos que denominam os mais conhecidos órgãos do corpo humano, 32 termos científicos, 5 prefixos, entre outros) e uma lista de Inglês Simplificado da *European Association of Aerospace Manufacturers* (lista criada para auxiliar engenheiros a escrever manuais de maneira a tornar a redação mais simples. No entanto, a lista da associação não se encontra disponível no site).

Uma vez que a *Simple English Wikipedia* importa-se com a seleção e utilização do vocabulário presente no corpo de seus verbetes, partimos da hipótese de que seu conteúdo lexical diferencia-se da *Wikipedia* , que potencialmente contém palavras simplificadas.

Frente ao exposto, nossa pesquisa tem o objetivo geral de apresentar o texto enciclopédico como um recurso didático para o enriquecimento e prática de vocabulário em língua inglesa. Os objetivos específicos são os seguintes: 1) proceder uma análise do perfil lexical de artigos da *Simple English Wikipedia* e da *Wikipedia*; 2) comparar os

artigos nas duas enciclopédias e 3) checar se os artigos adaptados da enciclopédia destinada aos aprendizes realmente empregam vocabulário mais elementar em seus textos.

A fundamentação teórica recorrerá basicamente à Lexicologia e Linguística de Córpus. Trata-se de campos independentes de investigação que, explorados conjuntamente, enriquecerão nossas reflexões.

2 Fundamentação teórica

A utilização de obras de consulta como apoio na formação linguística de modo geral não constitui uma novidade propriamente dita. Na área de língua inglesa, existem trabalhos relacionados especialmente à Lexicografia Pedagógica ou à Terminologia Aplicada, comprometidos com a associação entre obras de consulta e ensino. Empreendemos aqui uma discussão que tangencia a mesma linha, porém, recorrendo à enciclopédia, que , ao nosso ver, constitui ainda um recorte incipiente no contexto brasileiro de ensino de língua inglesa. A fim de situar o estudo de uma obra de consulta como a enciclopédia, adaptamos para um mapa conceitual o esquema desenvolvido por Welker (2005, p. 44):

1.	**OBRAS DE CONSULTA**
1.1	**Dicionário de língua**
1.1.1	***impresso/convencional***
1.1.1.1	*monolíngue*
1.1.1.1.1	*geral*
1.1.1.1.2	*especial*
1.1.1.2	*bi/multilíngue*
1.1.1.2.1	*geral*
1.1.1.2.2	*especial*
1.1.2	***eletrônico***
1.1.2.1	*monolíngue*
1.1.2.1.1	*geral*
1.1.2.1.2	*especial*
1.1.2.2	*bi/multilíngue*
1.1.2.2.1	*geral*
1.1.2.2.2	*especial*

1.2	**Outras obras de consulta**
1.2.1	***impresso/convencional***
1.2.1.1	*enciclopédias*
1.2.1.2	*atlas*
1.2.1.3	*almanaques*
1.2.1.4	*etc*
1.2.2	***eletrônico***
1.2.2.1	*enciclopédias*
1.2.2.2	*atlas*
1.2.2.3	*almanaques*
1.2.2.4	*etc*

No que se refere às enciclopédias eletrônicas, o mapa conceitual mostra que tanto as enciclopédias em formato papel (impresso/convencional) quanto as enciclopédias eletrônicas gozam de mesma importância quanto à categorização hierárquica. Entretanto, quanto ao nível 1.2 (outras obras de consulta), o referido autor falha em registrar a diferenciação entre obras monolíngues e bi/multilíngues. Salientamos que o escopo de nossa pesquisa reside no nível 1.2.2.1, monolíngue (em língua inglesa).

Cremos ser oportuna a explicação de Rey-Debove (1971) quanto à especificidade da enciclopédia, ou melhor, da definição enciclopédica: o dicionário de língua diz o que significa o signo leão, ao passo que a enciclopédia diz e mostra o que é um leão. Trazendo essa noção para o contexto da pesquisa, partimos do princípio de que, no artigo enciclopédico, encontram-se disponíveis inúmeras palavras para descrever – em alguns casos, exaustivamente – o objeto ou fenômeno. Dessa forma, encontramos um vasto repositório linguístico que pode ser explorado para fins didáticos.

Ao voltar nossas atenções para o vocabulário utilizado no texto enciclopédico, ressaltamos que temos interesse direto no vocabulário mais frequente da língua inglesa, ou seja, no vocabulário fundamental. Iniciamos na sequência algumas discussões inseridas na seara da Lexicologia, que pode ser definida como uma divisão da Linguística, cuja preocupação é o estudo científico do repertório de palavras presentes em um idioma, ou seja, o léxico propriamente dito. Sua manifestação mais concreta encontra-se no vocabulário, as palavras em uso pelos falantes.

Sob os auspícios da Lexicologia, podemos explorar, analisar, refletir, comparar e identificar as unidades léxicas. Os benefícios da leitura para o aprendiz de língua estrangeira são destacados por Nation (2015), especialmente o contato com textos simplificados por parte dos iniciantes.

Acatamos as ideias de Nation (2001, 2003) quanto ao fato de que, já que algumas palavras ocorrem com uma frequência muito maior que outras, as mais frequentes revelam-se potencialmente mais úteis aos alunos – conhecimento que é um pré-requisito seminal para o planejamento de um programa de vocabulário e para a tomada de decisão no dia a dia sobre como lidar com determinadas palavras.

Com relação à frequência e extensão das palavras, Nation (2001, 2003), propõe a divisão do vocabulário em língua inglesa em quatro grupos. O primeiro grupo é constituído por palavras de alta frequência composto de aproximadamente 2.000 famílias de palavras. Correspondem de 80% a 95% das palavras que ocorrem em um texto qualquer. O segundo grupo abarca as palavras acadêmicas que costumam ocorrer em textos acadêmicos e não fazem parte das 2.000 palavras mais frequentes. Essas palavras compõem entre 8,5% e 10% de um texto qualquer. Para Nation (2003), a melhor lista para essa classe é a Lista de Palavras Acadêmicas. No entanto, Silva (2015) detecta deficiências nessa lista e propõe outra lista acadêmica com forte apelo estatístico. O terceiro grupo abrange as palavras técnicas, ou seja, palavras comuns e de significado restrito à determinada área de especialidade. Por fim, podemos visualizar outro grupo dentro do qual seriam inseridas as palavras de baixa frequência, ou seja, que não fazem parte dos grupos citados anteriormente.

Leffa (2000) faz uma síntese a respeito do vocabulário e destaca seu imprescindível papel para o aprendizado de uma língua, o que faz do léxico instrumento fundamental.

> Partindo do princípio de que a simples instrução específica do vocabulário não garante a compreensão de leitura, o aluno deve aprender as palavras novas dentro de um contexto significativo, que pode ser dado por relações intratextuais, onde o significado da palavra desconhecida pode ser inferenciado dentro do próprio texto, e por relações intertextuais, considerando aí as disciplinas do currículo escolar. (LEFFA, 2000, p. 37).

Sob essa perspectiva, um trabalho direcionado para o enriquecimento lexical por meio de verbetes enciclopédicos pode

contemplar as ideias expostas pelo referido autor. O mesmo autor ressalta a importância de uma tríade fundamental no trabalho linguístico: a seleção do vocabulário que deve ser ensinado (o que), os textos que serão usados (com o que) e as estratégias empregadas (como). Ora, tendo em mente o que será ensinado, com quais instrumentos e de que maneira serão aplicados, percebemos pontos-chave para o aprendizado do aluno.

Nesse sentido, o uso de dispositivos que são acessíveis e de uso comum para todos tornam-se excelentes instrumentos de ensino. Essa visão abarca nossa proposta de trabalho com textos enciclopédicos como *input* para o desenvolvimento do repertório lexical em língua inglesa. Frente à responsabilidade que o professor de língua inglesa carrega quanto à instrução lexical, concordamos com a seguinte afirmação:

> Cabe ao professor incluir o vocabulário nas suas preocupações ao preparar suas aulas, propondo atividades em que determinadas palavras consideradas chave sejam explicitamente ensinadas. Dessa forma, o professor chama a atenção do aluno para aquelas palavras, possibilitando uma maior discussão e reflexão sobre elas, o que é imprescindível para facilitar sua retenção. (RODRIGUES, 2006, p. 17).

A tarefa do professor de problematizar o conhecimento do léxico leva a práticas que envolvam o vocabulário no contexto da sala de aula de língua estrangeira. Como mostram Oliveira e Silva (2016), pode-se aprimorar a língua inglesa via leitura, a fim de manter o aprendiz em contato com conteúdo linguístico do programa pedagógico. Soma-se ao material oficial de um curso de idiomas, por exemplo, um valioso repositório de estudo disponível on-line contendo uma grande variedade de textos, destinado a um público com faixa etária e nível de proficiência diversos. Destacamos, assim, a inserção do texto enciclopédico como mais um recurso para se estudar a linguagem (tanto para o professor de língua inglesa como para o aprendiz).

Até o momento, discutimos aspectos mais quantitativos estudados pela Lexicologia, no contexto do ensino de língua inglesa. Subsidiariamente, acreditamos que a Linguística de Córpus (doravante, LC) desempenhe um papel fundamental com foco na descrição, compreensão, ensino, entre outros. Uma maneira de se estudar a linguagem ou de como chegar até ela, pode ser por meio das contribuições dessa área. A utilização de córpus sempre foi um recurso

empregado em pesquisas linguísticas (ALUISIO; ALMEIDA, 2006). Para tanto, usaremos a abordagem da LC como corpo de linguagem natural (autêntica) que pode ser usado como base para pesquisa linguística. Na presente pesquisa, essa abordagem possibilitará a coleta dos dados necessários para a análise, fornecendo evidências advindas do processamento de uma grande quantidade de textos e das palavras neles presentes. Dessa maneira, realizamos uma exploração do conteúdo lexical dos textos enciclopédicos para aplicações no ensino.

Recorremos à conceituação de Berber Sardinha (2004) para definir essa abordagem linguística:

> A Linguística de Córpus ocupa-se da coleta e exploração de corpora, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador. (BERBER SARDINHA, 2004, p. 325).

A investigação no âmbito da LC leva em consideração uma série de fatores no desenvolvimento de *corpora* eletrônicos, como destaca Berber Sardinha (2004, 2012): origem: os dados devem ser autênticos; propósito: o córpus deve ter a finalidade de ser objeto de estudo; composição: os dados do córpus devem ser criteriosamente escolhidos; formatação: os dados devem ser legíveis por computadores; representatividade: deve representar uma linguagem ou variedade; extensão: deve ser vasto para se tornar representativo.

O trabalho com córpus também exige a observância de certas características para que possamos enxergar uma tipologia. Berber Sardinha (2004, 2012) destaca sete elementos: modo: pode ser oral ou escrito; tempo: sincrônico, diacrônico, histórico e contemporâneo; seleção: amostragem, monitor, dinâmico, estático e equilibrado; conteúdo: especializado, regional e multilíngue; autoria: aprendiz (não nativo) e língua nativa; disposição interna: paralelo e alinhado; finalidade: estudo, referência e treinamento.

Berber Sardinha (2012) destaca que o avanço dos recursos computacionais contribuiu para conferir rapidez e capacidade de processamento de dados linguísticos. Ao adotar uma abordagem empirista da linguagem, compreendida como um sistema probabilístico, sugere que

existe uma primazia dos dados provenientes da observação da linguagem, em geral, reunidos sob a forma de um córpus (BERBER SARDINHA, 2004).

Embora o escopo dessa área de estudos da linguagem possa ser definido em termos do que as pessoas fazem com corpora, seria um engano supor que a LC seja somente um meio rápido de descrever como a linguagem funciona. A análise de um cóorpus pode revelar, e frequentemente revela, fatos a respeito de uma língua que nunca se pensou em procurar (KENNEDY, 1998) – o que torna a língua um objeto de estudo sem precedentes.

Para o ensino de línguas estrangeiras, a LC pode fornecer insumos relevantes no que se refere à frequência de palavras, às colocações, ao estudo de ocorrência e coocorrência de determinados itens. Um córpus de aprendizes, por exemplo, possibilita identificar pontos que apresentam mais dificuldades ao aprendiz. O professor ou o pesquisador tem a seu dispor ferramentas computacionais que revelam em poucos segundos informações preciosas. Assim, a pesquisa baseada em córpus pode ser desenvolvida de modo a lançar luz sobre certos usos, o que favorece uma tomada de consciência. Corpora da ordem de milhões de ocorrências retratam a língua em movimento, tal como é usada pelos falantes de determinada comunidade linguística. O professor ou o pesquisador passa a poder contar com a observação direta dos fenômenos linguísticos, o que garante certo nível de confiança ao trabalho, uma vez que os eventos são retratados tal como ocorrem e não como se acredita que possam ocorrer. Hunston (2010, p. 137) atesta que o córpus tem um impacto direto na atividade profissional do professor de língua estrangeira de duas maneiras: em primeiro lugar, modifica a maneira pela qual a língua é percebida, a partir das descrições linguísticas e, em segundo lugar, pode ser explorado para produzir material de ensino, formando uma base para o planejamento de novos conteúdos e metodologia.

Biber, Conrad e Reppen (2004) destacam que uma das grandes vantagens de uma abordagem baseada em córpus é que ela proporciona um alcance e uma fidedignidade antes impossível. Vale ressaltar que as análises baseadas em córpus não estão limitadas meramente a uma análise quantitativa estanque. É essencial que o trabalho com córpus possibilite a inclusão de análises de cunho qualitativo a respeito dos padrões quantitativos discutidos na pesquisa.

3 Metodologia

A presente pesquisa tem cunho exploratório, uma vez que tem como objetivo principal o aprimoramento de ideias que envolvem levantamento bibliográfico e análise de exemplos que instiguem a compreender o assunto em tela. Embora possa ser rotulada como exploratória, a pesquisa tem, também, características descritivas.

Utilizamos o *software VocabProfile* (VP), versão 4 para traçar o perfil lexical dos artigos tanto na SEW quanto na *Wikipedia*. Como explica Silva (2011), o VP é *software* de tratamento linguístico que divide o texto em várias faixas de frequência lexical e fornece como resultado um perfil lexical em termos quantitativos. O *software* analisa as palavras do texto nele inserido por meio da comparação com seu próprio banco de dados, ou seja, tomando como referência as listas de palavras pré-carregadas, é executado o trabalho de comparação entre a palavra arquivada no VP e as palavras do texto inserido para a pesquisa.

Ainda segundo Silva (2011), o resultado é a identificação das palavras e a indicação da faixa de frequência à qual a palavra pertence: K1 (as primeiras 1.000 palavras mais frequentes da língua inglesa), K2 (as próximas 1.000 palavras mais frequentes), AWL (as palavras mais comuns encontradas em textos acadêmicos) e OFF (todas as palavras que não pertencem às faixas anteriores).

Com relação aos procedimentos metodológicos, por meio do processo de copiar ("ctr+c") e colar ("ctr+v"), salvamos e organizamos os 35 melhores artigos da SEW (na opinião do editor do site), no formato texto simples. Cada artigo foi salvo em um arquivo independente.

O mesmo procedimento foi adotado na recolha dos 35 artigos da *Wikipedia*. Utilizando o título do artigo da SEW, procuramos seu par na *Wikipedia*, para então criar outros arquivos. Por exemplo, o artigo intitulado *Jupiter* foi pesquisado nas duas enciclopédias. Ao final, constituímos um córpus com 70 artigos em língua inglesa.

Os dados foram tabulados em uma planilha eletrônica do MS Excel 2016, com duas abas, uma para cada enciclopédia. Após a tabulação dos dados oriundos do *VocabProfile*, os cálculos utilizados foram "soma" e "desvio-padrão", funções de fórmulas da própria planilha. Optamos por incluir também o cálculo do desvio-padrão, que é uma medida de dispersão, ou seja, uma medida de variabilidade dos dados de uma distribuição de frequências. Em outras palavras, o desvio-padrão

possibilita que sejam medidos os valores para cima ou para baixo da média. Consequentemente, o desvio-padrão pode ser usado para descrever o grau de dispersão na distribuição da frequência.

Todos os dados obtidos e tabulados podem ser consultados nos Apêndices A e B.

4 Resultados e análise dos dados

Nesta seção, será estudado o perfil lexical dos artigos das duas enciclopédias. Na sequência, uma exposição do número de *types* (palavras consideradas isoladamente, sem suas repetições no texto) e *tokens* (palavras consideradas com as repetições) presentes na amostra. Depois, uma comparação de forma mais pontual com base no texto enciclopédico integral de um verbete para ilustrar uma análise linguístico-estatística. Por fim, será apresentado um excerto oriundo do texto enciclopédico das duas enciclopédias.

Nos 35 artigos da SEW selecionados para a análise amostral, percebe-se que, na média, 75,07% do conteúdo lexical pertence ao grupo das primeiras 1.000 palavras mais frequentes do inglês, ou seja, encontram-se na faixa K1. Como pode ser visualizado na Tabela 1, tendo em mente que o valor do desvio-padrão é 2,27, podemos afirmar que existe uma flutuação para mais e para menos na amostra. Isso quer dizer que, na maioria dos casos, a variação dos artigos encontra-se entre 77,34 e 72,8. No que se refere à *Wikipedia*, levando-se em conta o desvio-padrão, vemos uma variação na faixa K1 que vai de 66,01 a 77,33. Graças à variação do perfil lexical na faixa K1 nas duas enciclopédias, podemos notar que ambas acabam apresentando índices equivalentes nesse nível – o que atesta que vários textos das duas enciclopédias compartilham palavras do mesmo grupo de frequência. Dito de outra forma, vários textos da SEW, do ponto de vista lexical, não fazem jus ao título de "simples".

TABELA 1 – Perfil lexical (%) dos artigos das duas enciclopédias colaborativas

	SEW	DP	W	DP
K1	75,07	2,27	71,67	5,66
K2	5,74	1,78	5,66	1,41
AWL	2,54	1,31	4,29	2,07
OFF	16,73	2,19	18,37	2,38

Fonte: Dados da presente pesquisa.

Nota: SEW: *Simple English Wikipedia*; W: *Wikipedia*; DP: Desvio-padrão.

Ainda consultando a Tabela 1, percebemos que as palavras pertencentes à faixa K2 apresentam uma porcentagem de uso nos textos praticamente idêntico nas duas enciclopédias. Do ponto de vista lexical, a SEW não se mostra como mais simples em comparação com a *Wikipedia*.

Efetivamente, observamos uma baixa utilização do vocabulário acadêmico na SEW. Porém, ao avaliarmos a variação do desvio-padrão, mais uma vez, do ponto de vista lexical, a SEW acaba se mostrando praticamente no mesmo nível da *Wikipedia*, sem se diferenciar pela simplicidade de seu vocabulário.

Por último, os nomes próprios e as palavras menos comuns são abarcadas na faixa OFF. Não é percebida diferença que justifique o nome de simples para a SEW. O desvio-padrão indica que as duas enciclopédias mantêm uma porcentagem muito próxima no que se refere à porcentagem de nomes próprios e palavras menos comuns em seus textos.

Frente ao exposto, do ponto de vista lexical, a diferença entre o perfil lexical dos textos da SEW e dos textos da *Wikipedia* é mínima. Reiteramos que essa diferença é tão pequena que prejudica o título de "simples" da SEW.

A causa da ausência de diferença entre ambas enciclopédias pode residir no fato de os autores não serem especialistas na elaboração de material didático para aprendizes, tendo em mente o conteúdo dos verbetes da SEW. Outra possível explicação pode estar no julgamento e nos critérios subjetivos para a seleção dos artigos pelo editor da SEW. Quais seriam as razões que tornaram os 35 artigos da SEW os melhores? Paradoxalmente, a escolha pode ter se pautado muito mais pela presença de vocabulário menos simples em detrimento do vocabulário fundamental recomendado nas listas de palavras.

Com relação ao número de *types* e *tokens* presentes nas duas enciclopédias, podemos afirmar que na *Wikipedia*, os artigos são mais extensos – conforme Tabela 2. Alguns artigos têm um tamanho duas vezes maior na *Wikipedia* quando comparados à SEW. No entanto, houve casos nos quais o artigo da SEW era maior que seu similar da *Wikipedia*.

TABELA 2 – Número médio de types e tokens presentes nas duas enciclopédias

	SEW	DP	W	DP
Types	394,80	175,67	687,03	429,94
Tokens	1.583,40	1.054,24	3.433,66	2.922,83

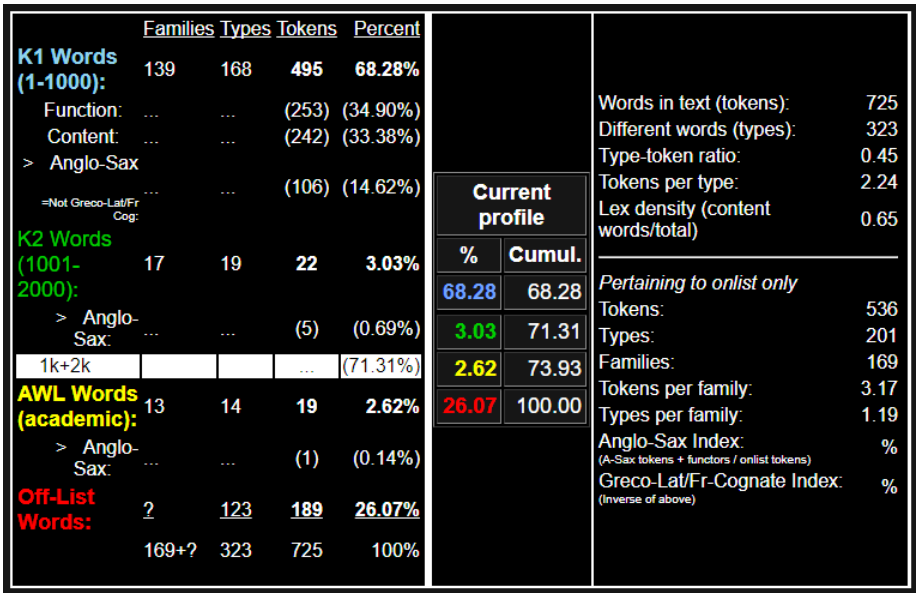
Fonte: Dados da presente pesquisa.

Nota: SEW: *Simple English Wikipedia*; W: *Wikipedia*; DP: Desvio-padrão.

De fato, os textos da SEW analisados trazem em média 1.583 palavras, com um desvio-padrão de 1.054, conforme dados da Tabela 2. Essa constatação possibilita concluir que os textos da amostra analisada seguem a sugestão de que contenham por volta de 2.000 palavras (SIMPLE ENGLISH WIKIPEDIA, 2016). Os dados obtidos levam ao entendimento que as 2.000 palavras fazem referência aos *tokens*, número total de palavras do texto. Percebemos que os textos da *Wikipedia* costumam ser mais extensos.

A fim de proceder uma comparação de forma mais pontual, selecionamos o verbete BRAZIL e seu texto enciclopédico integral para ilustrar uma análise linguístico-estatística. O verbete em questão não faz parte da amostra analisada neste estudo, como pode ser visto no Apêndice. No entanto, sua escolha justifica-se porque servirá para atestar se as inferências quanto à qualidade dos melhores artigos encontram amparo em um texto que não está na categoria dos melhores artigos. A Figura 1 apresenta o resultado da análise obtida pelo processamento do texto da SEW no *VocabProfile*:

FIGURA 1 – Perfil lexical do texto enciclopédico do verbete BRAZIL na SEW



	Families	Types	Tokens	Percent
K1 Words (1-1000):	139	168	**495**	**68.28%**
Function:	...	...	(253)	(34.90%)
Content:	...	...	(242)	(33.38%)
> Anglo-Sax				
=Not Greco-Lat/Fr Cog:	...	...	(106)	(14.62%)
K2 Words (1001-2000):	17	19	**22**	**3.03%**
> Anglo-Sax:	...	...	(5)	(0.69%)
1k+2k			...	(71.31%)
AWL Words (academic):	13	14	**19**	**2.62%**
> Anglo-Sax:	...	...	(1)	(0.14%)
Off-List Words:	?	123	**189**	**26.07%**
	169+?	323	725	100%

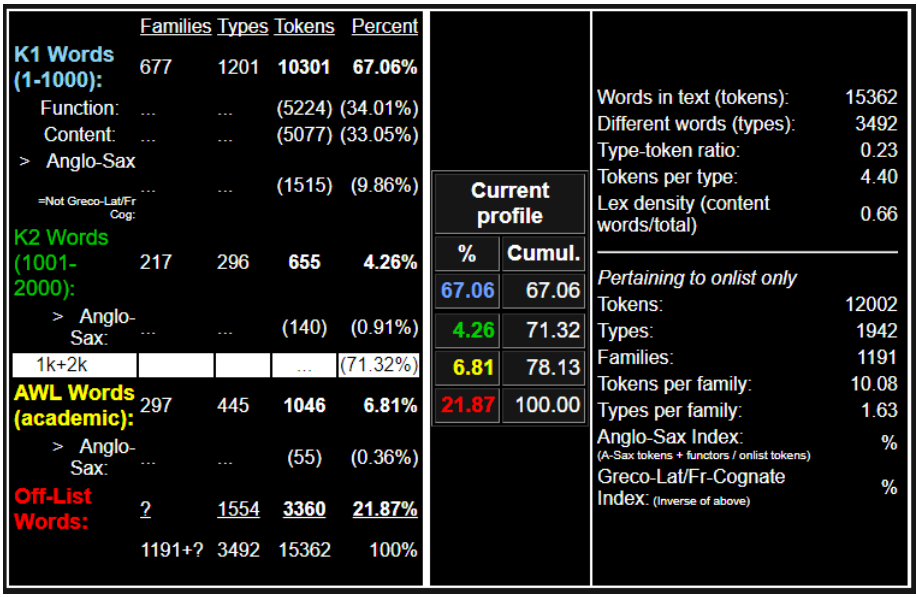
Current profile	
%	Cumul.
68.28	68.28
3.03	71.31
2.62	73.93
26.07	100.00

Words in text (tokens):	725
Different words (types):	323
Type-token ratio:	0.45
Tokens per type:	2.24
Lex density (content words/total)	0.65
Pertaining to onlist only	
Tokens:	536
Types:	201
Families:	169
Tokens per family:	3.17
Types per family:	1.19
Anglo-Sax Index: (A-Sax tokens + functors / onlist tokens)	%
Greco-Lat/Fr-Cognate Index: (Inverse of above)	%

Fonte: VocabProfile (2017).

Na SEW, como indica a Figura 1, percebemos o seguinte perfil lexical: K1: 68,28%; K2: 3,03%; AWL: 2,62% e OFF: 26,07%. O número total de palavras (*tokens*) no texto é de 725. Existem 323 palavras diferentes (*types*) nesse texto. Os dados linguístico-estatísticos apresentados podem servir para chamar a atenção do aluno para determinadas palavras, como ressalta Rodrigues (2006), o que possibilita uma maior discussão, reflexão e, consequente, retenção.

FIGURA 2 – Perfil lexical do texto enciclopédico do verbete BRAZIL na Wikipedia



	Families	Types	Tokens	Percent
K1 Words (1-1000):	677	1201	**10301**	**67.06%**
Function:	...	...	(5224)	(34.01%)
Content:	...	...	(5077)	(33.05%)
> Anglo-Sax =Not Greco-Lat/Fr Cog:	...	...	(1515)	(9.86%)
K2 Words (1001-2000):	217	296	**655**	**4.26%**
> Anglo-Sax:	...	...	(140)	(0.91%)
1k+2k			...	(71.32%)
AWL Words (academic):	297	445	**1046**	**6.81%**
> Anglo-Sax:	...	...	(55)	(0.36%)
Off-List Words:	?	1554	**3360**	**21.87%**
	1191+?	3492	15362	100%

Current profile	
%	Cumul.
67.06	67.06
4.26	71.32
6.81	78.13
21.87	100.00

Words in text (tokens):	15362
Different words (types):	3492
Type-token ratio:	0.23
Tokens per type:	4.40
Lex density (content words/total)	0.66
Pertaining to onlist only	
Tokens:	12002
Types:	1942
Families:	1191
Tokens per family:	10.08
Types per family:	1.63
Anglo-Sax Index: (A-Sax tokens + functors / onlist tokens)	%
Greco-Lat/Fr-Cognate Index: (inverse of above)	%

Fonte: VocabProfile (2017).

Por sua vez, na Wikipedia, como indica a Figura 2, identificamos o seguinte perfil lexical: K1: 67,06%; K2: 4,26%; AWL: 6,81% e OFF: 21,87%. O número total de palavras (*tokens*) no texto é de 15.362. Existem 3.492 palavras diferentes (*types*) nesse texto.

Recorrendo às informações tanto da Figura 1 quanto da Figura 2, no caso do conteúdo lexical do verbete BRAZIL, a soma das faixas K1 e K2 da SEW resulta no seguinte índice acumulado: 71,31%. A mesma somatória das faixas K1 e K2 da *Wikipedia* resulta no índice acumulado de 71,32%. Depreendemos que as enciclopédias encontram-se no mesmo patamar em termos lexicais, já que foi identificado esse empate técnico. Essas descobertas ilustram a ideia de Kennedy (1998) de que a análise de um córpus pode revelar fatos a respeito da língua que nunca se pensou em procurar.

Por limitação de espaço, a inserção do texto integral neste trabalho não é viável. Reproduzimos abaixo, então, um excerto do texto com as informações constantes na seção intitulada *Geography* do verbete BRAZIL na SEW:

> Brazil has the world's largest rainforest, the Amazon Rainforest. It makes up 40% of the country's land area. Brazil also has other types of land, including a type of savanna called cerrado, and a dry plant region named caatinga.
> The most important cities are Brasília (the capital), Belém, Belo Horizonte, Curitiba, Florianópolis, Fortaleza, Goiânia, Manaus, Porto Alegre, Recife, Rio de Janeiro, Salvador, São Paulo (the biggest city) and Vitória. Other cities are at list of largest cities in Brazil. Brazil is divided into 26 states plus the Federal District in five regions (north, south, northeast, southeast and centrewest):
> North: Acre, Amazonas, Rondônia, Roraima, Pará, Amapá, Tocantins
> Northeast: Maranhão, Pernambuco, Ceará, Piauí, Rio Grande do Norte, Paraíba, Alagoas, Sergipe, Bahia
> Centre-West: Goiás, Mato Grosso, Mato Grosso do Sul, Distrito Federal/ Federal District
> Southeast: São Paulo, Rio de Janeiro, Espírito Santo, Minas Gerais
> South: Paraná, Santa Catarina and Rio Grande do Sul
> The country is the fifth largest in the world by area. It is known for its many rainforests and jungles. It is next to every country in South America except Chile and Ecuador. (SIMPLE ENGLISH WIKIPEDIA, 2017b).

A leitura dessa seção revela a presença marcante de nomes próprios e uma preocupação em apresentar informações acerca da geografia brasileira. Retomando Leffa (2000), é possível que o aluno aprenda palavras novas nesse contexto significativo, que pode se dar por relações intratextuais, nas quais o significado da palavra desconhecida pode ser inferenciado dentro do próprio texto.

Na sequência, um excerto do texto com as informações constantes na mesma seção intitulada Geography do verbete BRAZIL na *Wikipedia*:

> Brazil occupies a large area along the eastern coast of South America and includes much of the continent's interior, sharing land borders with Uruguay to the south; Argentina and Paraguay to the southwest; Bolivia and Peru to the west; Colombia to the northwest; and Venezuela, Guyana, Suriname and France (French overseas region of French Guiana) to the north. It shares a border with every South American country except Ecuador and Chile. It also encompasses a number of oceanic archipelagos, such as Fernando de Noronha, Rocas Atoll, Saint Peter and Paul Rocks,

and Trindade and Martim Vaz. Its size, relief, climate, and natural resources make Brazil geographically diverse. Including its Atlantic islands, Brazil lies between latitudes 6°N and 34°S, and longitudes 28° and 74°W.

Brazil is the fifth largest country in the world, and third largest in the Americas, with a total area of 8,515,767.049 km2 (3,287,956 sq mi),[156] including 55,455 km2 (21,411 sq mi) of water.[15] It spans four time zones; from UTC−5 comprising the state of Acre and the westernmost portion of Amazonas, to UTC−4 in the western states, to UTC−3 in the eastern states (the national time) and UTC−2 in the Atlantic islands. Brazil is the only country in the world that has the equator and the Tropic of Capricorn running through it. It is also the only country to have contiguous territory both inside and outside the tropics. Brazilian topography is also diverse and includes hills, mountains, plains, highlands, and scrublands. Much of the terrain lies between 200 metres (660 ft) and 800 metres (2,600 ft) in elevation. The main upland area occupies most of the southern half of the country. The northwestern parts of the plateau consist of broad, rolling terrain broken by low, rounded hills.

The southeastern section is more rugged, with a complex mass of ridges and mountain ranges reaching elevations of up to 1,200 metres (3,900 ft). These ranges include the Mantiqueira and Espinhaço mountains and the Serra do Mar.[158] In the north, the Guiana Highlands form a major drainage divide, separating rivers that flow South into the Amazon Basin from rivers that empty into the Orinoco River system, in Venezuela, to the north. The highest point in Brazil is the Pico da Neblina at 2,994 metres (9,823 ft), and the lowest is the Atlantic Ocean. Brazil has a dense and complex system of rivers, one of the world's most extensive, with eight major drainage basins, all of which drain into the Atlantic. Major rivers include the Amazon (the world's second-longest river and the largest in terms of volume of water), the Paraná and its major tributary the Iguaçu (which includes the Iguazu Falls), the Negro, São Francisco, Xingu, Madeira and Tapajós rivers.

- Geography of Brazil
- Trindade and Martin Vaz is a volcanic archipelago off the coast of the Brazil.
- Serra dos Órgãos, part of the Serra do Mar.
- Chapada Diamantina, in the Chapada Diamantina National Park, Bahia.

- *Iguazu Falls, Paraná, is the largest waterfalls system in the world.*
- *Pico da Neblina, Amazonas, the highest mountain in Brazil.*
- *Cavern in Bonito, Mato Grosso do Sul.* (WIKIPEDIA, 2017b).

Nesse caso, considerando apenas as informações do excerto por ora analisado, podemos perceber que se trata de uma seção mais longa. O presente excerto pode ser dividido em diferentes grupos, como lembra Nation (2000). Essa divisão possibilita a criação de exercícios (orais e escritos) de forma a estudar e fixar cada grupo de palavras. O aporte dos recursos computacionais apresenta-se como uma ferramenta importante nessa perspectiva de ensino.

Haja vista a facilidade de acesso aos verbetes das duas enciclopédias, entendemos que ambas podem servir como base para a aquisição ou prática de vocabulário, dentro ou fora da sala de aula. Entendemos que o texto enciclopédico configura-se como um instrumento de *input* (entrada de informações) estimulante para o aprendizado da língua inglesa, já que reflete o uso de vocabulário e de estruturas gramaticais – podendo ser selecionado em função dos assuntos preferidos dos estudantes.

5 Considerações finais

Iniciamos o presente trabalho abordando questões relevantes para o ensino de língua inglesa com a Lexicologia e a Linguística de Córpus. Acreditávamos que essas linhas de estudo seriam fundamentais para a reflexão entre o vocabulário e o texto enciclopédico no ensino de língua inglesa. Essa pesquisa procurou lançar luz sobre uma estratégia de ensino de vocabulário, com foco no texto enciclopédico eletrônico. O contato do aprendiz com o léxico frequente pode contribuir no para ampliar seu repertório lexical.

Ao longo do trabalho, apresentamos o texto enciclopédico como um recurso didático para o enriquecimento e prática de vocabulário em língua inglesa; procedemos uma análise do perfil lexical de 35 artigos da *Simple English Wikipedia* e 35 artigos análogos da *Wikipedia*; comparamos os artigos nas duas enciclopédias sob um viés quantitativo e checamos se os artigos adaptados da enciclopédia destinada aos aprendizes realmente empregam vocabulário mais elementar em seus textos.

Os resultados obtidos indicam que, do conteúdo lexical dos textos da SEW, 80,81% das palavras dos textos encontram-se nas faixas K1 e K2 – o vocabulário fundamental. Da mesma forma, os resultados revelam que do conteúdo lexical dos textos da *Wikipedia*, 77,33% das palavras dos textos encontram-se nas faixas K1 e K2. Como estamos tratando de médias, merece atenção a variação indicada pelo desvio-padrão. A variação em todas as categorias (K1, K2, AWL e OFF) mostra que a diferença entre as duas enciclopédias é muito pequena – o que, se não as torna muito semelhantes, consegue torná-las muito pouco distintas, em uma visão léxico-quantitativa.

Apesar de haver uma preocupação com vocabulário adaptado e mais simples e várias sugestões de listas, a análise da amostra de textos do presente estudo indica que, levando em consideração apenas o perfil lexical, a SEW não se justifica. Não existe uma diferença expressiva com relação à qualidade das palavras utilizadas entre a SEW e *Wikipedia*. A utilização das listas *Basic English* 850, *Basic English* 1500, *Voice of America Special English Word Book* e o Inglês Simplificado da *European Association of Aerospace Manufacturers* parece não ter garantido uma acuidade na seleção do vocabulário empregado nos textos.

Com base na amostra de nosso estudo, parece ser equivocado o termo "*Simple English*". Portanto, nossa hipótese inicial de o conteúdo lexical da *Simple English Wikipedia* ser diferente do encontrado na *Wikipedia* não se confirmou.

A despeito da não diferenciação em termos da qualidade lexical, ou seja, da grande afinidade entre a *Simple English Wikipedia* e a Wikipedia, a leitura de seus textos pode ser benéfica em uma dupla perspectiva: além de expor os aprendizes a um grande número de vocabulário comum na língua inglesa, também pode potencialmente ser usada dentro ou fora da sala de aula.

Em tempo, ressaltamos que o item que pode diferenciar as duas enciclopédias analisadas seja a estrutura gramatical, que não foi contemplada nesse trabalho.

Destacamos que a leitura aqui proposta contempla apenas uma das quatro habilidades comunicativas da língua inglesa, nomeadamente a leitura. Não obstante, o vocabulário adquirido/praticado implicará a autonomia lexical necessária para a consecução das demais habilidades, seja na recepção ou na produção linguística.

A presente pesquisa traz uma contribuição para o enriquecimento das discussões relacionadas ao vocabulário fundamental, utilização de software de análise linguística e, neste caso, do texto enciclopédico, conhecimentos que são relevantes para pesquisadores, professores em atuação, professores em formação e para formadores de professores de língua inglesa.

Referências

ALUISIO, S. M; ALMEIDA, G. M. B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. *Calidoscópio*, São Paulo, v. 4, n. 3, p. 155-177, 2006.

BERBER SARDINHA, T. *Linguística de Corpus.* Barueri: Manole, 2004.

BERBER SARDINHA, T. Linguística de Corpus. In: GONÇALVES, A. V.; GÓIS, M. L. S. (Org.). *Ciências da linguagem*: o fazer científico? Campinas: Mercado de Letras, 2012. v. 1, p. 321-347.

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus Linguistics*: investigating language structure and use. Cambridge: Cambridge University Press, 2004.

HUNSTON, S. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press, 2010.

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

LEFFA, V. J. Aspectos externos e internos da aquisição lexical. In: LEFFA, V. J. (Org.). *As palavras e sua companhia*: o léxico na aprendizagem. Pelotas: Educat, 2000. p.17-46.

NATION, P. *Learning vocabulary in another language.* Cambridge: Cambridge University Press, 2001. Doi: https://doi.org/10.1017/CBO9781139524759

NATION, P. *Como estruturar o aprendizado de vocabulário.* Tradução de Cristiane Arruda. São Paulo: Special Book Services, 2003.

NATION, P. Principles guiding vocabulary learning through extensive reading. *Reading in a foreign language*, Honolulu, v. 27, n. 1, p. 136-145, abr. 2015.

OLIVEIRA, L. C.; SILVA, E. B. Impacto da leitura intensiva em língua inglesa no repertório lexical: uma análise quantitativa. *Domínios de lingu@gem*, Uberlândia, v. 10, n. 1, p. 380-406, jan./mar. 2016.

REY-DEBOVE, J. *Étude linguistique et sémiotique des dictionnaires français contemporains*. Paris: Hachette, 1971. Doi: https://doi.org/10.1515/9783111323459

RODRIGUES, D. F. Um olhar crítico sobre o ensino de vocabulário em contextos de inglês como língua estrangeira. *Trabalhos de Linguística Aplicada*, Campinas, v. 45, n. 1, p. 55-73, jan./jun. 2006. Doi: https://doi.org/10.1590/S0103-18132006000100004

SILVA, E. B. VocabProfile: uma ferramenta linguístico-estatística para a aula de língua inglesa. *Domínios de lingu@gem,* Uberlândia, v. 5, n. 1, p. 144-159, 2011.

SILVA, E. B. *Identificação e análise do vocabulário acadêmico em língua inglesa presente em textos acadêmico-científicos*. 2015. 289 f. Tese (Doutorado em Estudos Linguísticos) – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, 2015.

SIMPLE ENGLISH WIKIPEDIA. *How to write Simple English pages*. 2016. Disponível em: <https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages>. Acesso em: 21 ago. 2017.

SIMPLE ENGLISH WIKIPEDIA. *Simple English Wikipedia*. 2017a. Disponível em: <https://simple.wikipedia.org/wiki/Wikipedia:Simple_English_Wikipedia>. Acesso em: 21 ago. 2017.

SIMPLE ENGLISH WIKIPEDIA. *Brazil*. 2017b. Disponível em: <https://simple.wikipedia.org/wiki/Brazil>. Acesso em: 21 ago. 2017

WELKER, H. A. *Dicionários:* uma pequena introdução à Lexicografia. 2. ed. rev. e amp. Brasília: Thesaurus, 2005.

WIKIPEDIA. *Welcome to Wikipedia*. 2017a. Disponível em: <https://en.wikipedia.org/wiki/Main_Page>. Acesso em: 21 ago. 2017.

WIKIPEDIA. *Brazil*. 2017b. Disponível em: <https://en.wikipedia.org/wiki/Brazil>. Acesso em: 21 ago. 2017.

APÊNDICE A – Perfil lexical dos textos dos verbetes da *Wikipedia*

	Verbetes	K1	K2	AWL	OFF	*Types*	*Tokens*
01	*Hanami*	72,47	5,18	2,02	20,33	391	1383
02	Geisha	70,37	6,4	5,24	17,99	1284	5818
03	*Kamikaze*	75,03	4,55	2,07	18,35	437	1419
04	*Evolution*	67,84	4	11,49	16,67	1381	8313
05	*Violin*	68,78	6,24	8,38	16,59	2097	15508
06	*Ana Ivanović*	70,08	5,89	2,72	21,3	655	4795
07	Daniela Hantuchová	68,21	5,58	2,13	24,08	542	5388
08	*American Airlines Flight 11*	68,5	6,41	4,86	20,22	591	2214
09	*Anna Kournikov*	68,01	5,77	4,29	21,93	988	5168
10	*Jessica Alba*	77,65	3,02	2,21	17,11	357	1124
11	*Powderfinger*	72,35	5,66	3,26	18,72	322	1220
12	Baseball Uniform	69,68	7,11	5,76	17,45	392	1533
13	*Red Hot Chili Peppers*	73,13	5,91	2,32	18,64	364	1720
14	*Gothic Architecture*	66,38	4,96	6,39	22,26	1049	5653
15	*Crich Tramway Village*	72,32	3,57	4,46	19,64	57	90
16	*Ipswich Town F.C.*	75	7,43	2,45	15,12	252	971
17	Bobby Robson	70,23	7,05	4,51	18,21	901	4387
18	*Bloc Party*	70,65	5,2	4,79	19,36	525	2204
19	*Tropical Storm Barry*	65,74	6,21	4,88	23,17	339	1323
20	*Victoria Line*	75,98	4,72	5,07	14,23	473	1507
21	*Hermann Göring*	71,54	4,12	4,4	19,94	1531	7479
22	Jupiter	79,61	2,74	1,14	16,51	382	1613
23	*Portman Road*	73,9	6,02	3,26	16,81	195	663
24	*Blackpool Tramway*	72,34	5,06	4,41	18,19	901	4901
25	*Billy Graham*	71,2	4,75	3,57	20,52	1011	4349
26	*Yellow (song)*	71,61	6,02	3,61	18,76	1373	6803
27	Tropical Storm Gabrielle	68,07	8,78	5,04	18,11	352	1203
28	*Kingsway Tramway Subway*	78,58	3,61	2,95	14,86	364	1415
29	*Hurricane Vince*	75,05	6,02	2,45	16,47	549	3301
30	*Epping Ongar Railway*	71,02	7,06	6,17	15,76	806	3427
31	*City of Manchester Stadium*	69,26	6,09	6,4	18,25	744	3220
32	1910 Cuba hurricane	72,07	8,52	4,63	14,78	535	1730
33	*Dan Kelly*	75,06	6,7	1,01	17,23	524	2618
34	*Tropical Depression*	71,06	7,52	5,81	15,61	619	2266
35	*Saturn (Planet)*	69,82	4,25	6,06	19,87	763	3452

Fonte: *Wikipedia*.

APÊNDICE B – Perfil lexical dos textos dos verbetes da SEW

	Verbetes	K1	K2	AWL	OFF	*Types*	*Tokens*
01	Hanami	73,74	5,19	1,98	19,08	345	1060
02	*Geisha*	77,37	4,24	2,51	15,87	455	1606
03	*Kamikaze*	74,07	4,96	1,63	19,35	328	992
04	*Evolution*	74,72	4,86	7,36	13,95	1211	6188
05	*Violin*	75,99	8,16	0,6	15,25	341	1267
06	Ana Ivanović	76,95	4,42	1,76	16,86	551	2312
07	*Daniela Hantuchová*	70,96	6,73	1,44	20,87	229	823
08	*American Airlines Flight 11*	72,48	7,36	2,18	17,98	311	903
09	*Anna Kournikov*	74,66	5,2	1,73	18,41	278	895
10	*Jessica Alba*	77,56	2,96	2,22	17,26	354	1117
11	Powderfinger	72,3	5,67	3,27	18,76	322	1217
12	*Baseball Uniform*	73,42	5,71	4,03	16,84	403	1921
13	*Red Hot Chili Peppers*	73,08	5,92	2,32	18,67	364	1716
14	*Gothic Architecture*	74,61	5,58	1,39	18,42	655	4222
15	*Crich Tramway Village*	74,27	4,55	2	19,18	294	889
16	Ipswich Town F.C.	75,09	7,51	2,47	14,93	253	963
17	*Bobby Robson*	69,76	7,78	4,19	18,26	288	1092
18	*Bloc Party*	75,13	5,14	2,78	16,94	369	1769
19	*Tropical Storm Barry*	73,09	6,48	2,8	17,63	557	2733
20	*Victoria Line*	78,16	5,16	3,66	13,03	460	1569
21	Hermann Göring	79,02	2,55	1,4	17,03	269	945
22	*Jupiter*	79,44	2,68	1,42	16,46	376	1528
23	*Portman Road*	74,5	6,57	2,79	16,14	225	842
24	*Blackpool Tramway*	75	3,86	2,18	18,96	305	1154
25	*Billy Graham*	75,74	3,57	1,63	19,05	539	1882
26	Yellow (song)	73,37	8,72	2,99	14,91	408	1307
27	*Tropical Storm Gabrielle*	73,53	9,29	2,74	14,44	308	1031
28	*Kingsway Tramway Subway*	78,31	3,61	2,95	15,12	365	1409
29	*Hurricane Vince*	74,41	7,75	2,09	17,75	289	945
30	*Epping Ongar Railway*	79,84	5,66	2,97	11,53	340	1220
31	City of Manchester Stadium	74,08	6,11	3,1	16,72	512	1963
32	*1910 Cuba hurricane*	77,22	9,41	1,18	12,19	362	935
33	*Dan Kelly*	74,98	7,11	0,77	17,14	524	2378
34	*Tropical Depression*	74,85	6,38	5,8	12,96	176	450
35	*Saturn (Planet)*	75,64	4,05	2,73	17,58	452	2176

Fonte: *Simple English Wikipedia*.