Rev. Estud. Ling., Belo Horizonte, v. 28, n. 3, p. 1101-1248, 2020

Wordsmith Tools e *Sketch Engine*: um estudo analítico-comparativo para pesquisas científicas com uso de *corpora*

Wordsmith Tools and Sketch Engine: an analytical-comparative study for scientific research with corpora manipulation

Guilherme Fromm
Universidade Federal de Uberlândia (UFU), Uberlândia, Minas Gerais / Brasil.
guifromm@ufu.br
https://orcid.org/0000-0001-5654-0135

Candice Guarato Santos
Universidade Federal de Uberlândia (UFU), Uberlândia, Minas Gerais / Brasil.
candiceguaratos@gmail.com
https://orcid.org/0000-0001-5957-1842

Daniela Faria Grama
Universidade Federal de Uberlândia (UFU), Uberlândia, Minas Gerais / Brasil.
daniela_grama@hotmail.com
https://orcid.org/0000-0001-9301-3297

Neubiana Silva Veloso Beilke
Universidade Federal de Uberlândia (UFU), Uberlândia, Minas Gerais / Brasil.
neubeilke@hotmail.com
https://orcid.org/0000-0002-4432-0861

Resumo: O presente trabalho consiste na descrição e comparação de dois *softwares* de análise lexical, o *WordSmith Tools* (WST) e o *Sketch Engine* (SE). O *corpus* de estudo selecionado para a realização da análise comparativa entre os programas denomina-se LexTest, é composto por artigos, dissertações, resenhas e teses da área da Lexicologia, escritos em língua portuguesa, e apresenta 552.903 *tokens*. Os aspectos analisados

eISSN: 2237-2083
DOI: 10.17851/2237-2083.28.3.1191-1248

no WST e no SE são: interface, configuração de línguas, *upload* de *corpus*, número de *tokens* e *types*, etiquetagem do *corpus*, lista de palavras, lista de palavras-chave; acesso às linhas de concordância, entre outros. A partir das análises comparativas, foi possível concluir que o WST e o SE são eficazes no que se propõem, uma vez que, por exemplo, processam palavras-chave, permitem configurar a língua conforme o *corpus* de estudo e calculam o número de *tokens* de um *corpus*. Entretanto, dependendo do objetivo da pesquisa, um desses *softwares* pode ser mais apropriado que o outro. A escolha pela utilização de um deles é de responsabilidade do pesquisador, que poderá consultar o quadro comparativo entre o WST e o SE apresentado no final deste artigo para fundamentar a sua preferência.

Palavras-chave: Linguística de *Corpus*; *WordSmith Tools*; *Sketch Engine*.

Abstract: The present work consists of the description and comparison of two lexical analysis software, WordSmith Tools (WST) and Sketch Engine (SE). The study *corpus* selected for the comparative analysis between the programs is called LexTest which is composed of articles, dissertations, reviews and theses about Lexicology. These texts are written in Portuguese, and the corpus has 552,903 tokens. The aspects analyzed in the WST and in the SE are: interface, language settings, *corpus* upload, number of tokens and types, *corpus* tagging, word list, keyword list, access to concordance lines, and so on. Based on the comparative analyzes, it was possible to conclude that the WST and the SE are effective in their purpose, because, for example, they process the keywords and allow configuring the language according to the study *corpus* and calculate the number of tokens; however, depending on the purpose of the research, one of these programs may be more appropriate than the other. The choice of one of them will be responsibility of the researcher, who may consult the comparative table between the WST and the SE, presented at the end of this paper to substantiate his or her preference.

Keywords: *Corpus* Linguistics; WordSmith Tools; Sketch Engine.

Recebido em 17 de setembro de 2019

Aceito em 27 de novembro de 2019

1 Introdução

A necessidade de realizar descrições linguísticas é assinalada por Perini (2006, 2008) quando trata da carência de estudos descritivos amplos que se baseiem em dados linguísticos primários a partir de observações, organizações e sistematizações. O referido linguista salienta que há escassez de dados coletados e descritos para sustentar algumas

teorizações. Dessa forma, Perini (2006, 2008) enfatiza o problema de se proporem teorias que visem explicar o funcionamento de línguas sem que estejam ligadas, satisfatoriamente, a fatos linguísticos. Cabe esclarecer que o autor não despreza a teorização, mas dá ênfase ao modo como ela é empreendida; para Perini (2006, 2008), tal processo deve dar espaço à linguística descritiva, em uma perspectiva que se fundamente no uso efetivo da língua.

Ao listar alguns princípios da linguística descritiva e discorrer sobre a gramática descritiva, Perini (2006, 2008) elenca algumas características que o linguista deve possuir, a saber: se interessar pela língua como ela é, e não como ela deveria ser; partir sempre de fatos linguísticos para sua análise; não confundir suas hipóteses com os fatos; registrar e descrever como é que as pessoas realmente falam e/ou escrevem e sistematizar os fatos da língua encontrados. Observamos que tais princípios também norteiam a Linguística de *Corpus* (doravante LC), embora Perini (2006; 2008) não tenha se referido diretamente a ela nem a *corpora* eletrônicos. De qualquer modo, ele reconhece que o uso de *corpus*, de maneira geral, tem a vantagem de neutralizar os desejos do pesquisador, pois, se o linguista se valer apenas da introspecção, isso poderá influenciar subjetivamente os resultados de sua análise.

Tendo em vista a importância do uso de *corpora* em pesquisas científicas na área da Linguística, o objetivo deste artigo é realizar uma análise comparativa entre duas grandes ferramentas que auxiliam o pesquisador a lidar, em termos quantitativos, com *corpora* eletrônicos extensos: W*ordsmith Tools* (doravante WST) e S*ketch Engine* (doravante SE). Nessa perspectiva, inscrevemos este trabalho na subárea da Linguística Computacional, em virtude de focarmos nas funcionalidades do WST e do SE – frequentemente utilizados em diferentes trabalhos que lançam mão da abordagem/metodologia da LC. Vale ressaltar que, para o desenvolvimento da nossa análise comparativa, restringimo-nos à verificação de algumas funcionalidades básicas de ambos, detalhadas na seção metodológica e de descrição e análise do presente artigo, pensando na perspectiva de pesquisas que se enquadram na área da Lexicografia e da Terminografia.

Existem diversos estudos comparativos entre ambientes e ferramentas que servem ao trabalho do linguista que utiliza *corpora.* Podemos citar, por exemplo, o de Wilkens *et al*. (2012), que avaliou a eficácia de três ambientes de gestão terminológica que estão disponíveis

na internet e que propiciam a criação de um produto final, a saber: o e-Termos, o VoTec e o TermWiki. Além de descrevê-los e analisá-los, os autores apontaram fatores que consideraram essenciais para a construção de um glossário *on-line*, dentre eles, ambiente amigável, interface que permita a inclusão de elementos multimídia e ferramentas que possibilitem a identificação do perfil e da necessidade dos usuários finais e que sejam responsáveis pela elaboração de mapa conceitual e pela interação entre membros da equipe, já que Wilkens *et al*. (2012) visavam à construção colaborativa de um glossário. Os autores constataram que os ambientes analisados não se adequavam às suas necessidades e, por isso, apresentaram algumas soluções para o gerenciamento terminológico de que precisavam.

Outro trabalho que podemos mencionar é o de Gomide (2015), que discorre, de forma comparativa, sobre o *AntConc* e o pacote *TextMining-R*. Para compará-los, Gomide (2015) lança mão de um *corpus* de textos escritos por aprendizes de inglês. A autora dedica-se à análise das seguintes operações realizadas pelo *AntConc* e pela linguagem de programação de código-aberto R: lista de palavras, colocados e linhas de concordância. Ao final, Gomide (2015) afirma que o *AntConc* e o pacote *TextMining-R* não se diferem muito em termos de resultados, no entanto ela especifica qual deles é mais vantajoso quando desempenha determinada atividade e a qual público cada um atende.

Ao observarmos que análises comparativas como as mencionadas são úteis aos pesquisadores que utilizam *corpora* e que não há um estudo que tenha realizado uma análise contrastiva especificamente entre o WST e o SE, sentimo-nos motivados a produzir este artigo. Acreditamos que esta investigação poderá contribuir sobremaneira para orientar as escolhas que linguistas e pesquisadores de áreas afins precisem tomar em seus estudos e em suas práticas docentes quando o assunto envolver análise de *corpora* por meio de programas de análise lexical.

Além desta introdução, este artigo está dividido em mais quatro grandes seções: Fundamentação teórica, Metodologia, WST e SE: Descrição e análise e Considerações finais. Na próxima seção, Fundamentação teórica, abordamos conceitos básicos relativos à LC, à Lexicografia e à Terminografia e apresentamos o WST e o SE – alvos de nossa análise. Na metodologia, descrevemos os procedimentos que realizamos para analisarmos o WST e o SE com o uso de um mesmo *corpus* da área da Linguística. A seção intitulada WST e SE: Descrição

e análise tem por objetivo mostrar com detalhes o que há de comum entre ambos e também os recursos que cada um tem em particular. Nas considerações finais, apresentamos um quadro comparativo que consiste na sumarização das análises realizadas. A finalidade desse quadro é auxiliar o pesquisador que está em dúvida entre o WST e o SE, isto é, que possui dificuldades em identificar qual deles é mais adequado para desenvolver sua pesquisa linguística nas áreas da Lexicografia e da Terminografia.

2 Fundamentação teórica

Inicialmente, discorremos sobre o que é a LC, uma vez que ela é a base deste trabalho. Na sequência, explicamos brevemente o que é o WST e o SE – alvos de análise comparativa neste artigo. Em seguida, abordamos as áreas da Lexicografia e da Terminografia, relacionando-as à metodologia/abordagem da LC.

2.1 Linguística de *Corpus* (LC)

A LC está relacionada "à criação e análise de *corpora*" (BERBER SARDINHA, 2009, p. 7). De acordo com Beilke (2016), a LC é uma:

> abordagem-metodologia de princípios descritivos, que se fundamenta em dados autênticos e se relaciona com as evidências de maneira ampla. Ela permite a produção de conhecimentos ancorados na realidade linguística, pois dá primazia à observação prévia dos dados levantados, além de nos guiar para a investigação de hipóteses não premeditadas e para a descoberta e a comprovação de fatos linguísticos (BEILKE, 2016, p. 72).

Ainda em referência às características e aos princípios que definem a LC, a autora esclarece:

> Esse empirismo, um de seus pressupostos, é um meio de fundamentar a pesquisa objetivamente, em detrimento da especulação. Sob sua perspectiva, a análise dos dados permite verificar traços que se repetem, padrões de comportamento linguístico, variações recorrentes e, assim, atestar se existem regularidades sistemáticas, confirmando a hipótese de que não são aleatórias. A partir de então, torna-se possível quantificá-las, descrevê-las e analisá-las, o que contribui para esclarecer

> suposições a respeito do funcionamento linguístico e que produz conhecimentos inovadores e os mais diversificados estudos e olhares sobre a linguagem em geral (BEILKE, 2018, p. 370).

Quanto à definição de *corpus*, é uma "coleção de textos" (SINCLAIR, 1991, p. 171) e, nesse sentido, já existia antes da LC. Entretanto, dentro da LC, a definição de *corpus* vai além desse conceito básico, pois precisa ser uma coleção de textos "de ocorrências de linguagem natural" (SINCLAIR, 1991, p. 171), ou seja, autêntica e escolhida para caracterizar um estado ou uma variedade de linguagem.

Berber Sardinha (2004, p. 16-17) define *corpus* como "uma coletânea de textos reunida com um propósito definido de ser usado como base para a pesquisa linguística". O autor estabelece também que *corpus* é um artefato produzido para a pesquisa, com textos autênticos, com "porções de linguagem que são planejadas, selecionadas, organizadas, de acordo com critérios linguísticos explícitos, a fim de serem usadas como uma amostra da linguagem e armazenadas em formato legível por computador" (BERBER SARDINHA, 2004, p. 16-17).

Consideramos que as propriedades que definem *corpus* podem ser complementadas pela conceituação de Biderman (2001). A autora afirma que um *corpus* "constitui um conjunto homogêneo de amostras da língua de qualquer tipo (orais, escritos, literários, coloquiais)" (BIDERMAN, 2001, p. 79) e que "a análise dos dados linguísticos de um *corpus* deve permitir ampliar o conhecimento das estruturas linguísticas da língua que eles representam" (BIDERMAN, 2001, p. 79).

Por fim, podemos dizer que a principal finalidade de um *corpus* nos moldes exigidos pela LC, atualmente, é propiciar a análise lexical – atividade linguisticamente realizada com o auxílio de programas, ambientes e linguagens computacionais por pesquisadores que atuam nas áreas da Lexicologia, Lexicografia, Terminologia e Terminografia. Vale salientar que, embora a LC seja bastante utilizada nessas subáreas da Linguística, nada impede que ela seja útil, do ponto de vista metodológico, para outras subáreas – tema que será desenvolvido em trabalhos futuros. A seguir, apresentamos, respectivamente, o WST e o SE.

2.2 *WordSmith Tools* (WST)

Conforme Berber Sardinha (2009, p. 6), o WST surgiu em 1996 e "é um conjunto de programas integrados", definido por seu criador, Scott,

como um *software* de análise lexical. Publicado pela *Oxford University Press*, o WST oferece várias funções, dentre elas, visualização dos dados de *corpora*, criação de listas de palavras, de linhas de concordâncias em forma de visualização vertical, a partir de nódulos (palavras de buscas), e de palavras-chave, entre outras. Suas funcionalidades são mais bem apresentadas na seção intitulada WST e SE: Descrição e Análise.

2.3 *Sketch Engine* (SE)

Elaborado por Adam Kilgarriff e Pavel Rychlý e desenvolvido pela *Lexical Computing Ltd*., em 2003, o SE é um gerenciador de *corpus* e *software* de análise de textos *online*. O objetivo dessa ferramenta é o de propiciar pesquisas, por meio de *corpora*, em torno do funcionamento de diversas línguas. Para atingir tal propósito, há vários recursos, como o *tesauro*, que encontra palavras com significados semelhantes ou que aparecem em contextos similares, as listas de frequências e a compilação e gestão de *corpora*. Assim como as funcionalidades do WST, os recursos do SE são aprofundados na seção WST e SE: Descrição e análise.

2.4 Lexicografia

A Lexicografia é uma subárea da Linguística Aplicada que abrange questões de cunho teórico e prático relativas ao processo de elaboração de obras dicionarísticas. De acordo com Borba (2003), Seabra (2011) e Welker (2011), a Lexicografia é dividida em Lexicografia Teórica (Metalexicografia), que dá conta dos princípios teóricos que subsidiam a prática, e em Lexicografia Prática, que diz respeito à construção em si de dicionários.

No âmbito da Lexicografia, lidamos com dicionários que abarcam as palavras de uma língua (dicionários monolíngues) ou de mais de uma língua (dicionários bilíngues). São vários os pontos que precisam ser pensados e planejados quando discorremos sobre a elaboração de dicionários, tais como: macroestrutura, medioestrutura, microestrutura, público-alvo (em específico, as necessidades e dificuldades dele), disponibilização da obra (*on-line*, eletrônica ou impressa), proposta lexicográfica, *design*, uso de *corpora*, entre outros, o que faz com que um dicionário geral de língua seja fruto de um trabalho complexo, visto que demanda tempo, recursos financeiros e uma equipe bem organizada de profissionais/ colaboradores.

Geralmente, buscamos um dicionário para sanar dúvidas em relação à forma e ao sentido das palavras, a fim de que possamos usá-las de modo apropriado em determinados contextos de comunicação. Mais do que isso, conforme Bevilacqua e Finatto (2006), os dicionários são uma forma de registrar o nosso "patrimônio sociocultural" (BEVILACQUA; FINATTO, 2006, p. 45), na medida em que encontramos neles os signos que constituem a nossa língua. Portanto, os dicionários são mais do que obras de consulta que podem nos auxiliar sobremaneira em momentos de dúvida quanto ao uso de palavras de uma língua; na verdade, registram as escolhas lexicais de um povo, evidenciando a cultura e a identidade dele.

Quando tratamos de Lexicografia, é essencial associarmos a ela a metodologia/abordagem da LC. Segundo Biderman (2003), é importante que o trabalho de elaboração de dicionários seja pautado em *corpus*, pois, dessa forma, os lexicógrafos podem basear suas decisões, em termos de escolha de macroestrutura e de informações que devem constar na microestrutura, em usos reais e mais ou menos frequentes de uma língua.

Nessa perspectiva, o uso de programas de análise lexical tem sido cada vez mais frequente em trabalhos que envolvem a construção de obras dicionarísticas. Com o auxílio de *softwares* voltados para a análise do léxico, como o WST e o SE, é possível identificarmos as palavras que são mais ou menos usadas em um *corpus*, ou seja, a frequência delas, com quais outras palavras são usadas, com quais temas ou assuntos associam-se em maior número, em quais contextos linguísticos podem ser utilizadas, entre outras informações. A partir disso, é possível subsidiar análises sobre o uso e comportamento da língua, por exemplo, para a elaboração de um dicionário.

Com base em Fromm (2002), podemos dizer que as obras lexicográficas diferenciam-se das terminográficas por serem produtos gerais de língua, e não de uma especialidade, de uma área técnica ou científica. No tópico a seguir, abordamos especificamente sobre Terminografia.

2.5 Terminografia

A Terminografia consiste na esfera prática da Terminologia, disciplina essa que oferece base teórica para análises de termos técnico-científicos. De acordo com Fromm e Yamamoto (2013), o estudo terminológico visa selecionar as palavras que são específicas de uma área de especialidade qualquer. Também conhecida como Terminologia

Aplicada, a Terminografia é responsável pela elaboração de obras de referência, por exemplo, os dicionários terminológicos e os glossários. Teline, Almeida e Aluísio (2003) destacam a relevância da Terminologia atualmente:

> A Terminologia cumpre um importante papel no mundo moderno, repleto de inovações científico-tecnológicas, posto que esses avanços científicos e tecnológicos precisam ter nomes, e nomes apropriados. Dessa forma, o uso de repertórios terminológicos sistematizados ou harmonizados – por meio da Terminologia – contribui para tornar mais eficaz a comunicação entre especialistas, comunicação essa que se propõe, acima de tudo, a ser concisa, precisa e adequada (CABRÉ, 1996) (TELINE; ALMEIDA; ALUÍSIO, 2003, p. 1).

A perspectiva terminológica adotada para um estudo influenciará na produção das obras de consulta. Entre as correntes, podemos citar a Teoria Geral da Terminologia (TGT), desenvolvida por Eugênio Wüster, que é prescritiva, ou seja, o seu objetivo consiste em normatizar os termos, permitindo que a comunicação entre os especialistas seja a mais eficiente possível.

Outra corrente terminológica é a Teoria Comunicativa da Terminologia (TCT), proposta por Cabré (1999), que apresenta uma visão diferente do termo, isto é, o termo é visto como uma lexia que está na posição de termo, pois, conforme Krieger e Finatto (2004), as unidades lexicais obtêm estatuto terminológico no contexto das comunicações especializadas. Assim, o objetivo da TCT é descrever os usos dos termos, caracterizando-se, portanto, como uma concepção descritiva da Terminologia.

Segundo Cabré (1995), os dados coletados, o método de coleta, o tratamento dos dados e a apresentação em forma de glossários são os fatores que diferenciam a Terminografia da Lexicografia (Lexicologia Aplicada). De forma resumida, é a metodologia que distingue essas duas disciplinas que tratam do léxico. A pesquisadora comenta que, apesar das semelhanças no processo de elaboração de seus produtos, cada uma apresenta suas particularidades:

> Certamente, a Lexicografia, concebida como um ramo aplicado da Lexicologia que trata da elaboração de dicionários poderia coincidir com a Terminografia, que é o ramo da Terminologia

> Aplicada que também lida com a elaboração de dicionários especializados ou glossários terminológicos. Mas, embora o processo de trabalho de ambas as práticas convirja no desenvolvimento de dicionários, outros aspectos lhes conferem especificidade e fazem de um dicionário geral um produto diferenciado de terminologia (CABRÉ, 1995, p. 9, tradução nossa).[1]

Considerando o aspecto metodológico, a LC oferece ferramentas eficientes para a compilação e o processamento de dados terminográficos. Quanto ao gerenciamento dos dados linguísticos por meio de *corpus*, Teline, Almeida e Aluísio (2003) explicam que, para realizar "a tarefa de sistematizar/harmonizar repertórios terminológicos, é fundamental que haja ferramentas computacionais compatíveis com esse tipo de empreendimento" (TELINE; ALMEIDA; ALUÍSIO, 2003, p. 1).

Conforme Almeida e Vale (2008), as transformações nos campos da Terminologia e da LC têm mudado o método de descrever e sistematizar terminologias, principalmente a extração de termos. Para os autores, o aumento no número de compilação de *corpus* nas pesquisas no campo da Terminologia é devido à facilidade de se obter textos eletrônicos por meio da internet. Almeida e Vale (2008) explicam que, após as etapas de processamento das informações linguísticas contidas no *corpus*, os candidatos a termo podem ser identificados:

> Após serem finalizadas todas as etapas que envolvem o *corpus* (compilação, manipulação, anotação e pré-processamento), ele está pronto para ser objeto de extração semiautomática de candidatos a termos. Os candidatos constituem itens léxicos que se comportam nos seus respectivos contextos como termos, mas cuja autenticidade será validada posteriormente (ALMEIDA; VALE, 2008, p. 484).

---

[1] No original: "Ciertamente, la lexicografía, concebida como rama aplicada de la lexicología que se ocupa de la elaboración de diccionarios, podría coincidir con la terminografía, que es la rama aplicada de la terminología que se ocupa también de la elaboración de diccionarios especializados o de glosarios terminológicos. Pero aunque el proceso de trabajo de ambas prácticas converge en la elaboración de diccionarios, otros aspectos les dan especificidad y hacen que un diccionario general sea un producto diferenciado de una terminología".

Como podemos observar, por meio da LC, é possível ter acesso ao contexto do uso real do candidato a termo. Outro aspecto que podemos notar, por meio dessa citação, é que o processo é semiautomático, ou seja, o programa apresenta os dados processados, mas o pesquisador deve interpretá-los e analisá-los. Tal fato demonstra que o programa auxilia o trabalho do pesquisador, porém não realiza todo o trabalho.

Navarro (2013) apresenta outros benefícios do uso de *corpora* em estudos terminológicos. Segundo a autora, essa prática tem sido muito divulgada e tornou-se indispensável. "Suas vantagens também são amplamente conhecidas, como, por exemplo, maior facilidade, objetividade e confiabilidade na identificação de padrões lexicais" (NAVARRO, 2013, p. 195).

O processo de compilação e organização de *corpora* terminológicos necessita de uma série de procedimentos, pois é importante a organização e a confiabilidade das informações obtidas por meio das ferramentas de análise lexical, conforme explica Almeida (2010):

> A elaboração de um *corpus* para pesquisas terminológicas (sobretudo naquelas cujo objetivo é a construção de dicionários, glossários, vocabulários, ontologias, bases terminológicas, etc.) exige o cumprimento de uma série de requisitos, já que é a partir do *corpus* compilado que: a) se extraem os termos e suas eventuais formas variantes; b) se observam as colocações e as fraseologias próprias de um discurso especializado; c) se infere as relações semânticas entre os termos de maneira que seja possível a elaboração de uma eventual ontologia; d) se observa o termo em todos os seus contextos de ocorrência, sendo possível inferir traços semânticos recorrentes para redigir a definição terminológica (ALMEIDA, 2010, p. 78).

Além dessas possibilidades que o uso de *corpus* na Terminografia proporciona, podemos acrescentar a questão da frequência, que, para a LC, denota uso, e um termo recorrente na língua especializada precisa ser registrado em uma obra terminográfica. Na sequência, tratamos da metodologia utilizada neste artigo.

3 Metodologia

Dividimos esta seção em duas subseções. Na primeira, apresentamos os critérios de escolha do *corpus* de testagem, denominado,

LexTest e suas características. Na segunda seção, descrevemos os passos metodológicos que seguimos no processo de comparação entre WST e SE.

3.1 Origem, elaboração e características do LexTest

Para escolhermos um *corpus* de estudo que pudesse ser testado tanto no WST quanto no SE, estabelecemos os seguintes critérios:

- ✓ Ser composto por textos escritos em língua portuguesa;
- ✓ Ser sincrônico e contemporâneo;
- ✓ Ter aproximadamente 500 mil *tokens*;
- ✓ Estar normalizado e codificado;
- ✓ Estar disponível gratuitamente;
- ✓ Ser de fácil acesso.

Essas exigências nos levaram a optar pelo *Corpus* de Linguística. Tal *corpus* foi planejado por Fromm, um dos autores deste artigo, professor do curso de Graduação em Letras: Inglês e Literaturas de Língua Inglesa e da Pós-Graduação em Estudos Linguísticos (PPGEL) da Universidade Federal de Uberlândia (UFU), com o intuito de ser a base para a criação de verbetes bilíngues (português/inglês) na plataforma terminográfica Vocabulário Técnico *On-line* (VoTec[2]) – desenvolvida em sua tese de Doutorado (FROMM, 2007).

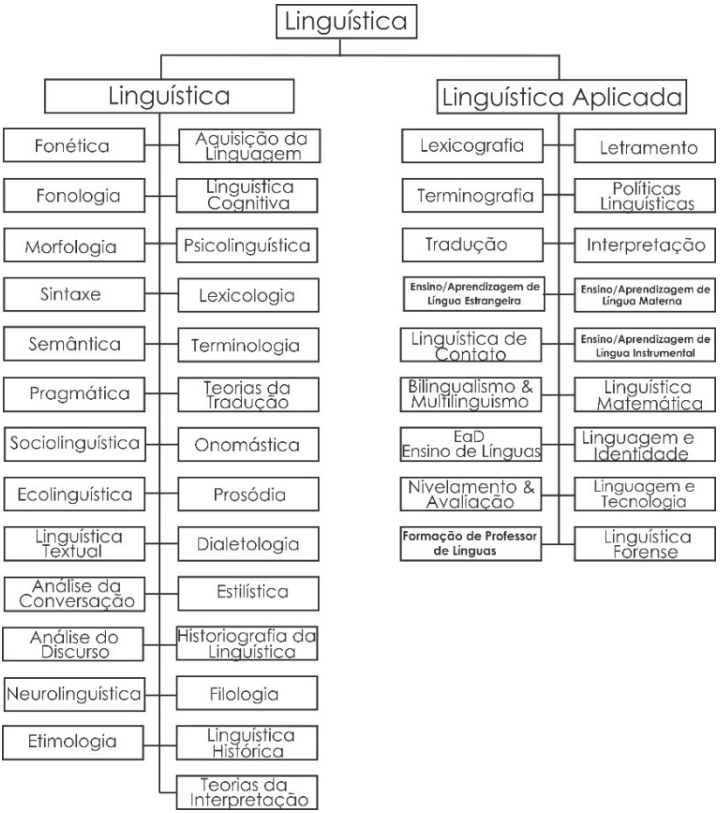
O projeto de construir um *corpus* da área de Linguística, segundo Fromm e Yamamoto (2013), ocorreu de maneira colaborativa, ou seja, alunos de graduação e de pós-graduação da Universidade, após compreenderem os princípios da LC, auxiliaram na compilação das subáreas da árvore de domínio da Linguística. Conforme os referidos autores, os textos que constituem o *Corpus* de Linguística são artigos, resenhas, dissertações e teses. Pelo fato de o principal objetivo do *Corpus* de Linguística ser a criação de verbetes bilíngues, a compilação foi realizada tanto na língua inglesa quanto na portuguesa.

No que diz respeito à elaboração da árvore de domínio da Linguística, notamos que o processo não foi simples, pois Fromm e Yamamoto (2013) descrevem que os especialistas da área não entraram em concordância em relação às subáreas pertencentes à Linguística, além

[2] Disponível em: http://pos.votec.ileel.ufu.br/. Acesso em: 15 nov. 2019.

de as informações oferecidas pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) no que diz respeito ao que é do domínio da Linguística não terem sido suficientes. Dessa forma, Yamamoto (2018) esclarece que, à medida que a compilação do *corpus* era realizada, ele ia percebendo quais campos de estudo poderiam ser caracterizados como subáreas da Linguística. Atualmente, a árvore da Linguística, encontra-se de acordo com a Figura 1 a seguir:

FIGURA 1 – Árvore de domínio da Linguística



Fonte: Fromm (2018, p. 317).

Para o recorte deste artigo, restringimo-nos a utilizar como *corpus* de testagem apenas uma subárea referente à árvore de domínio da Linguística: a Lexicologia. É válido ressaltar que Yamamoto (2018) tem o

propósito de normalizar, limpar e organizar o *Corpus* de Linguística para usá-lo em sua pesquisa de Doutorado na elaboração de um vocabulário bilíngue português/inglês da Linguística. Em virtude disso, solicitamos a ele o envio dos textos relativos à Lexicologia para que pudéssemos usá-lo como c*orpus* de testagem neste artigo. A partir desse momento, passamos a denominar o nosso *corpus* de testagem relativo à subárea da Lexicologia como LexTest.

Assim, o LexTest apresenta a tipologia descrita no Quadro 1. Para elaborá-la, baseamo-nos na tipologia de *corpus* proposta por Berber Sardinha (2004) e, posteriormente, ampliada por Teixeira (2008):

QUADRO 1 – Tipologia do LexTest

Tipologia do LexTest	
Língua	Monolíngue (português)
Tipo	Modalidade padrão
Conteúdo	Especializado
Modo	Escrito
Autoria	Diversas (falantes nativos e não nativos)
Seleção	Por amostragem (*sample corpus*)
Tamanho	552.903 *tokens* – Médio[3]
Finalidade	Testagem e análise dos programas WST e SE
Tempo	Sincrônico
Balanceamento	Não balanceado
Integralidade	Composto por textos integrais
Fechamento/Status	Estático
Nível de codificação	Não etiquetado

Fonte: Elaboração própria.

Na seção a seguir, descrevemos os procedimentos que realizamos para analisar o WST e o SE.

[3] Segundo Berber Sardinha (2004, p. 26), quanto aos parâmetros de tamanho de corpora na perspectiva da LC, até o ano de 2004, um corpus de menos de 80 mil palavras seria classificado como pequeno; um corpus de 80 a 250 mil palavras seria classificado como pequeno-médio; um corpus de 250 mil a 1 milhão de palavras seria classificado como médio; um corpus de 1 milhão a 10 milhões de palavras seria classificado como médio-grande e um corpus de 10 milhões ou mais de palavras seria classificado como grande.

3.2 WST e SE: procedimentos para realização de análises comparativas

O primeiro passo para iniciarmos a análise comparativa entre WST e SE foi adquirir o acesso a eles, uma vez que são pagos. Utilizamos a sétima versão do pacote de ferramentas do WST e um plano de assinatura de um mês, com direito ao uso de até um milhão de *tokens*, do SE.

O segundo passo consistiu em delimitarmos a nossa análise comparativa entre o WST e o SE, visto que ambos disponibilizam muitas funcionalidades ao pesquisador. Dessa forma, pensando no recorte deste artigo, optamos por verificamos apenas as seguintes questões:

- Interface;
- Configuração de línguas;
- Agilidade no *upload* de *corpus*;
- Contagem de *tokens, words e types*
- Formas de etiquetagem possíveis;
- Funções e ferramentas associadas;
- Importação e exportação de dados;
- Processamento da lista de palavras;
- Acesso às linhas de concordância, contextos e textos;
- Processamento de palavras-chave.

Na próxima seção, apresentamos a análise comparativa entre o WST e o SE.

4 WST e SE: Descrição e análise

Nesta seção, descrevemos e analisamos como o WST e o SE se comportam ao trabalharmos com o LexTest. Os principais aspectos que verificamos **são: interface, configuração,** carregamento de *corpus*, contagem de *tokens* e as ferramentas *WordList*, *Concord* e *Keywords*.

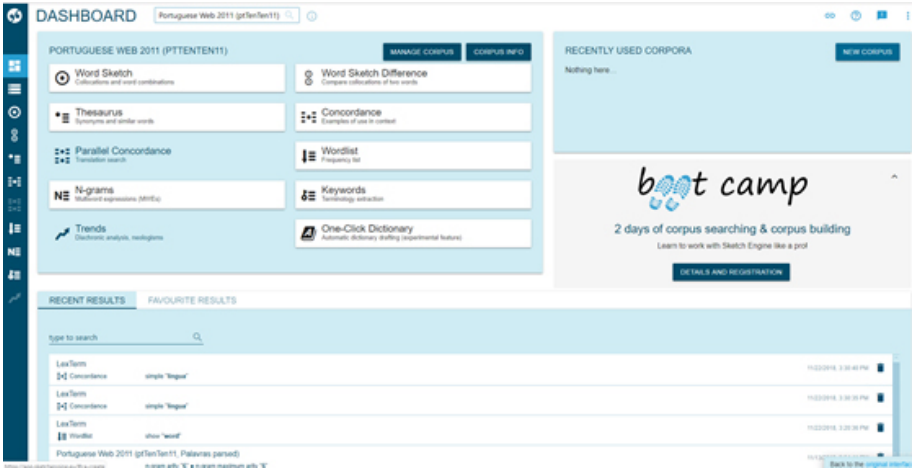
4.1 WST e SE: interface

Logo após inserirmos as informações relativas a usuário e à senha no WST e SE, vemos as telas iniciais de cada um, conforme as Figuras 2 e 3.

FIGURA 2 – Interface do WST

Fonte: Scott (2016).

FIGURA 3 – Interface do SE



Fonte: Kilgarriff e Rychlý (2003).

Consideramos que ambas as interfaces são amigáveis; isso significa que o pesquisador, mesmo em um primeiro contato com o WST e o SE, poderá ter certa familiaridade com eles, visto que foram projetados conforme os moldes do *Windows*. No que alude aos aspectos visuais, a tela inicial do WST apresenta as três ferramentas principais (*Concord*, *KeyWords* e *WordList*) em destaque na parte superior da janela e, abaixo, na lateral esquerda, as visualizamos novamente acompanhadas de outras três ferramentas, *WSConcgram*, *Chargrams* e *Utilities*, além das configurações opcionais.

A tela inicial do SE apresenta treze ferramentas que podem ser acessadas por meio de botões: *Word Sketch*, *Thesaurus*, *Parallel Concordance*, *N-grams*, *Trends*, *Word Sketch Difference*, *Concordance*, *Wordlist*, *Keywords*, *One-Click Dictionary*, *Manage Corpus*, *Corpus Info* e *New Corpus*, excetuando-se as configurações. Diante disso, a nosso ver, um usuário iniciante pode ter mais dificuldades em lidar com o SE do que com o WST. Inclusive, o próprio SE anuncia na mesma tela a venda de cursos que ensinam o usuário a mexer no programa.

Outra questão importante é que os dois foram elaborados totalmente em língua inglesa, o que pode acarretar um pouco de dificuldade para aqueles que não dominam esse idioma ou que estejam conhecendo a terminologia da LC. O manual ou a aba *Help* do WST e

do SE também estão em língua inglesa. Vale ressaltar que, no caso do WST, há vídeos[4] feitos por um falante de língua portuguesa, Wendell Dantas (2010),[5] que estão disponíveis no canal *Youtube* e que contêm conteúdo explicativo sobre como utilizar a versão 3.0 e 5.0 do programa, além do livro *Pesquisa em Linguística de Corpus com WordSmith Tools* de Berber Sardinha (2006), que também pode auxiliar bastante o usuário do WST. No *Youtube*, também encontramos um canal[6] e vários vídeos sobre o SE, porém nenhum feito por um falante de língua portuguesa.

4.2 WST e SE: configuração de língua

Antes de trabalhar com o WST e o SE, o pesquisador deve configurar a língua de ambos de acordo com o idioma do *corpus* que será processado por eles. Para isso, no WST, clicamos em *language settings*, escolhemos a língua desejada (portuguesa do Brasil) e, em seguida, clicamos em *Ok* para salvar a opção feita. No SE, observamos que a configuração de língua deve ser realizada no preenchimento de um formulário que antecede a inserção de um *corpus*. Assim, após clicar em *New Corpus*, surge o formulário com os seguintes campos: *name*, *corpus type* (*single language corpus* ou *multilingual corpus*), *language* e *description*. No nosso caso, colocamos o nome LexTest, optamos por *corpus* de uma única língua, selecionamos a língua portuguesa e, na descrição, escrevemos apenas *Corpus* de Lexicologia.

A nosso ver, a configuração de língua é bastante simples tanto no WST quanto no SE. Contudo, vale ressaltar que o fato de ser uma etapa obrigatória e que antecede a inserção de um *corpus* de estudo no SE faz com que o usuário não se esqueça de passar por ela, o que consideramos um aspecto positivo devido à importância da escolha do idioma do *corpus*.

4.3 WST e SE: *upload* de *corpus*

No WST, no momento em que o pesquisador utiliza as ferramentas *Concord* ou *WordList* é que ele solicita o carregamento de um *corpus*.

---

[4] Disponível em: https://www.youtube.com/user/CorpusLael/videos. Acesso em: 25 jul. 2019.

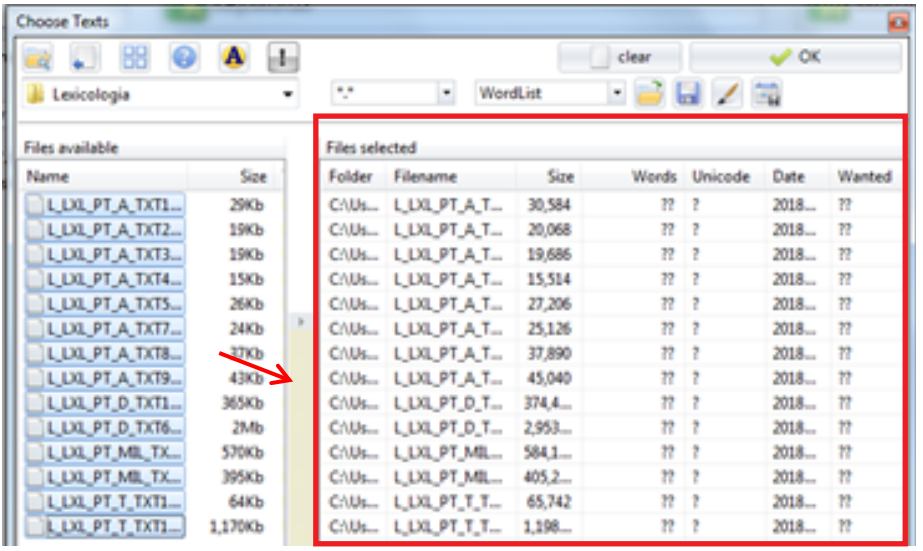
[5] Disponível em: https://www.youtube.com/watch?v=7ybj-0-y5lI. Acesso em: 14 jan. 2020.

[6] Disponível em: https://www.youtube.com/channel/UCo2fn2_SNxCikCSAFCBcWBw. Acesso em: 25 jul. 2019.

Por ser um *software* instalável, é necessário que o *corpus* já esteja no computador em que o programa está instalado. Cientes disso, clicamos em *WordList* – botão em destaque na parte superior e direita da interface do programa (visualizado na Figura 2) – depois clicamos em *File* e em *New*.

Em seguida, o WST abre uma janela, na qual clicamos na opção *Choose Texts Now*. Esse botão nos permite encontrar o local, no nosso computador, em que o LexTest está. Ao localizarmos o nosso *corpus* de estudo, selecionamos os arquivos que o compõem e o inserimos no WST. Após carregar o *corpus*, é necessário indicar quais arquivos serão efetivamente processados pelo programa. Para isso, é preciso selecionar os arquivos inseridos no lado esquerdo da janela e clicar na seta cinza ou em qualquer lugar da barra cinza; a partir de então, eles constarão do lado direito do *software*, conforme Figura 4.

FIGURA 4 – *Upload* do LexTest no WST

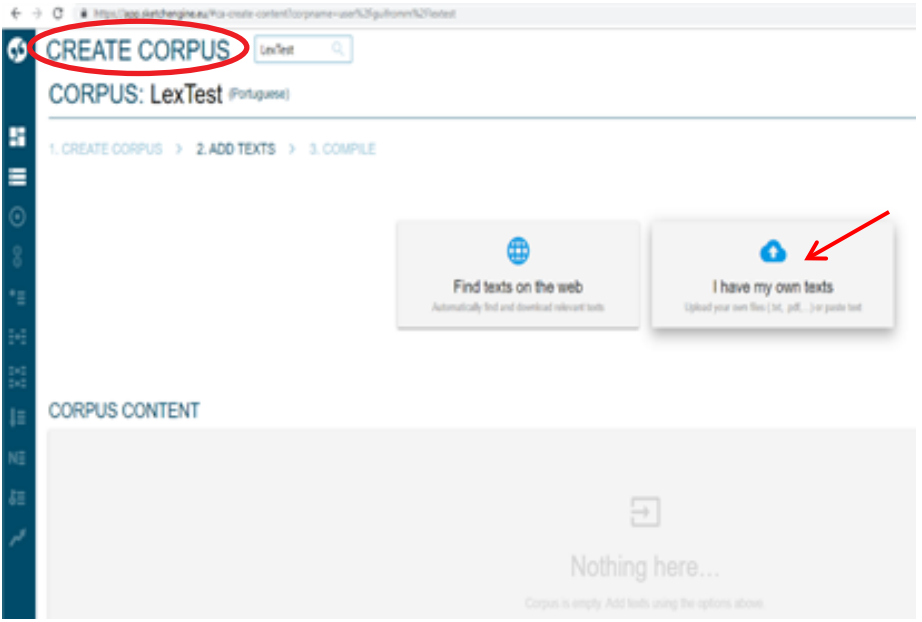


Fonte: Scott (2016).

Vale ressaltar que os arquivos do *corpus* a serem carregados no WST precisam estar salvos em arquivo TXT (bloco de notas) com a codificação *Unicode*.

No SE, após preenchermos o formulário com as informações sobre o LexTest, clicamos em *Next*, para darmos sequência ao *upload* dele. Em seguida, surgiu a tela da Figura 5.

FIGURA 5 – *Upload* do LexTest após o preenchimento do formulário



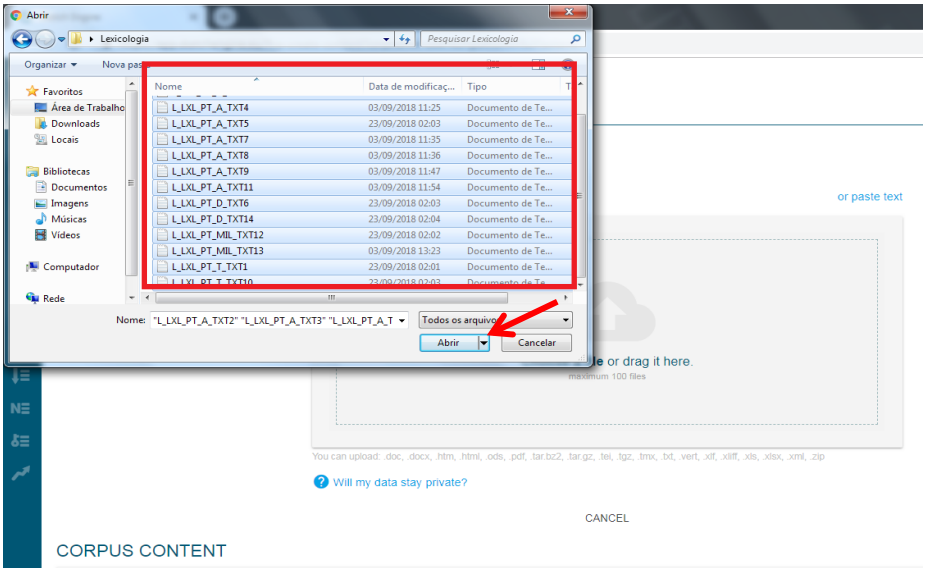
Fonte: Kilgarriff e Rychlý (2003)

Notamos que a terminologia utilizada pelo SE não é a mais adequada, uma vez que *create corpus* (criação de *corpus*), em destaque com um círculo vermelho na Figura 5, não é o mesmo que simplesmente realizar o *upload* de um *corpus* em um *software*, embora ele dê a opção para compilação automática de *corpus*, denominada *find texts on the web* (encontre textos na *web*). Como já tínhamos o nosso *corpus* de testagem, clicamos em *I have my own texts* (eu tenho os meus próprios textos).

Vale pontuar que, diferentemente do WST, os arquivos do *corpus* a serem carregados no SE podem estar salvos em vários formatos: TXT, DOC, DOCX, HTML, XML etc. No nosso caso, os arquivos estavam em TXT. Além disso, o SE também aceita tanto a codificação *Unicode* quanto a *ANSI* do arquivo em TXT.

Após clicarmos em *I have my own texts*, o SE apresenta uma tela em que aparece a opção *choose a file or drag it here* (escolha um arquivo ou arraste para aqui). Ela nos permite localizar o LexTest em nosso computador e selecionar os arquivos que o compõem, conforme Figura 6.

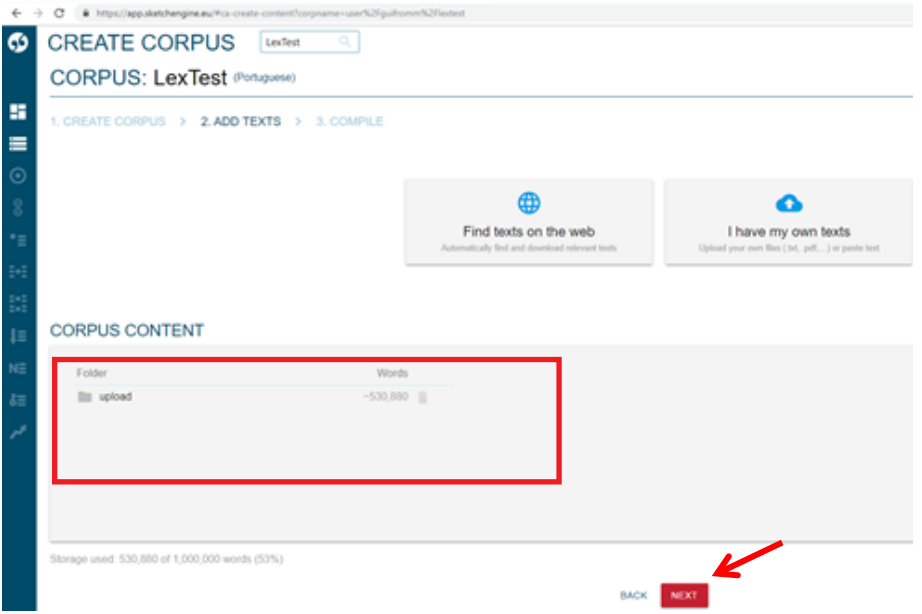
FIGURA 6 – Seleção dos arquivos do LexTest no *SE*



Fonte: Kilgarriff e Rychlý (2003)

Após selecionarmos os textos do nosso *corpus* e clicarmos na opção Abrir da Figura 6, o SE iniciou e finalizou o carregamento do LexTest. Ao finalizar o carregamento, as informações sobre esse procedimento aparecem na cor cinza, conforme a Figura 7.

FIGURA 7 – Carregamento finalizado do LexTest no SE



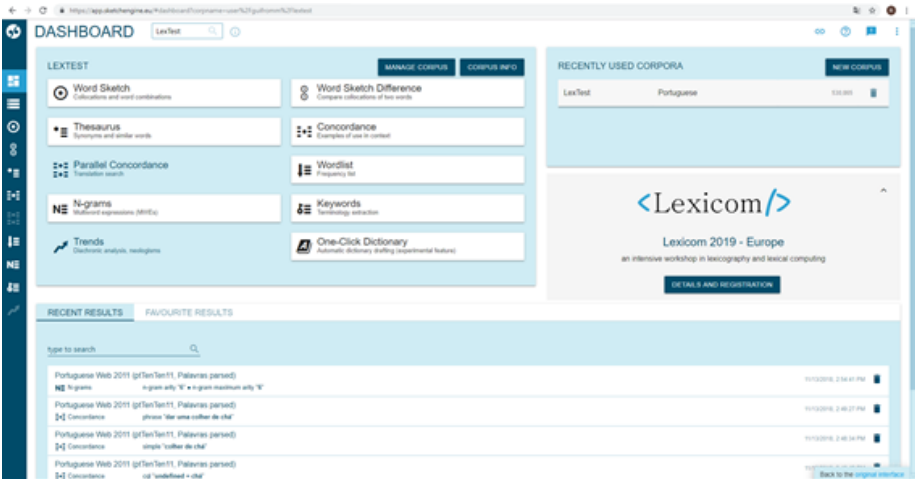
Fonte: Kilgarriff e Rychlý (2003)

Observamos que o caminho percorrido para carregar o LexTest no SE e também o tempo que o próprio sistema levou para efetuar o *upload* dos arquivos do *corpus* em questão foi bem mais lento em relação ao WST. O SE levou um minuto e quatro segundos para realizar o *upload* do LexTest, e o WST despendeu apenas dois segundos para fazer o *upload* do mesmo *corpus*.

Após clicarmos em *Next*, presente na Figura 7, aparece outra tela que oferece duas opções: *add more texts* (adicionar mais textos) ou *compile* (compilar). Optamos por *compile*, por não termos mais arquivos para inserir. Notamos que a funcionalidade *compile* alude à contabilização dos arquivos do *corpus* e ao processo de etiquetagem dele. No caso do LexTest, o processo denominado *Compile* durou 44 segundos. Finalizada a compilação, o SE oferece três opções: *add more texts* (adicionar mais textos), *recompile* (recompilar) e *corpus dashboard* (painel de controle do *corpus*).

Ao clicarmos em *corpus dashboard*, o SE nos direciona ao painel de controle do LexTest (FIGURA 8), local em que, finalmente, podemos começar a explorar o *corpus* por meio das ferramentas disponibilizadas pelo SE.

FIGURA 8 – Painel de controle do LexTest no SE



Fonte: Kilgarriff e Rychlý (2003)

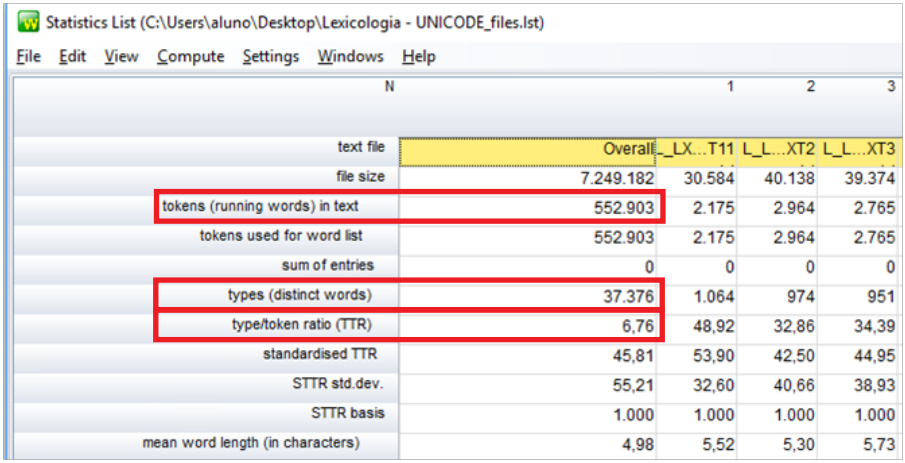
Uma grande vantagem do SE em relação ao WST é que, após o *upload* do *corpus*, ele o etiqueta automaticamente de acordo com o sistema gramatical da língua escolhida no momento da inserção do *corpus*. Para facilitar o reconhecimento das *tags* (etiquetas), o SE possui tabelas explicativas sobre as codificações utilizadas na etiquetagem na forma de legendas. Dessa forma, caso um pesquisador necessite etiquetar seu *corpus*, pode, aparentemente, ganhar tempo com o uso do SE, visto que o WST não faz esse procedimento automaticamente. O quesito etiquetagem no SE será comentado novamente mais adiante.

4.4 WST e SE: *tokens*, *words* e *types*

No WST, os arquivos do LexTest, em TXT com a codificação *Unicode*, atingiram 552.903 *tokens* (*running words*) *in text*, ou seja, palavras corridas, conforme Figura 9[7] a seguir. Segundo Berber Sardinha (2009, p. 174), *tokens*, "também chamado de '*running words*', significa o total de palavras, levando em conta as repetições, desde a primeira até a última de todos os arquivos selecionados" (BERBER SARDINHA, 2009, p. 174).

[7] Os tipos de informações da Figura 9 podem ser acessados na terceira aba denominada *statistics* (estatísticas) da ferramenta *WordList* do WST.

FIGURA 9 – Número de *tokens* do LexTest no WST

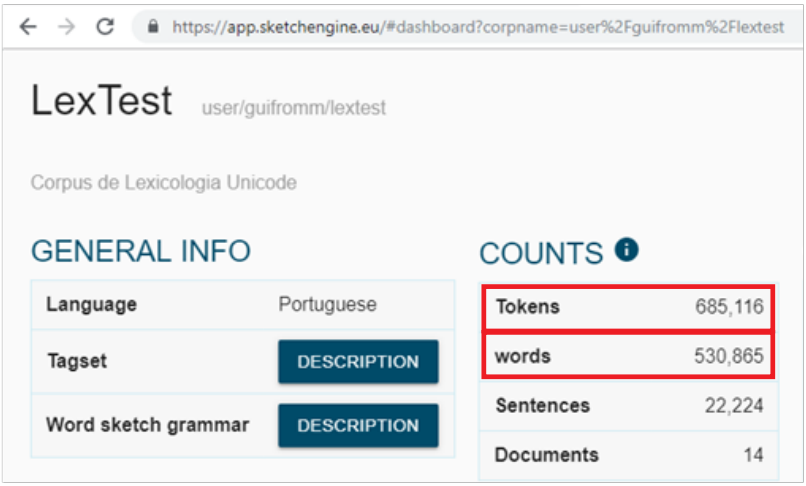


N		1	2	3
text file	Overall	_LX...T11	L_L...XT2	L_L...XT3
file size	7.249.182	30.584	40.138	39.374
tokens (running words) in text	552.903	2.175	2.964	2.765
tokens used for word list	552.903	2.175	2.964	2.765
sum of entries	0	0	0	0
types (distinct words)	37.376	1.064	974	951
type/token ratio (TTR)	6,76	48,92	32,86	34,39
standardised TTR	45,81	53,90	42,50	44,95
STTR std.dev.	55,21	32,60	40,66	38,93
STTR basis	1.000	1.000	1.000	1.000
mean word length (in characters)	4,98	5,52	5,30	5,73

Fonte: Scott (2016).

No SE, os mesmos arquivos do LexTest, com igual codificação, somaram 685.116 *tokens*, conforme Figura 10.[8]

FIGURA 10 – Número de *tokens* do LexTest no SE



Fonte: Kilgarriff e Rychlý (2003)

[8] Os tipos de informações da Figura 10 podem ser acessados na aba *Corpus Info* presente no *Dashboard* (Painel de Controle) do *corpus* inserido no SE.

Ao estranharmos essa diferença entre eles, buscamos saber o que representava *tokens* no SE e encontramos a definição presente no glossário do próprio SE. De acordo com a nossa leitura, no SE, *tokens* diz respeito não somente à quantidade de palavras de um *corpus*, mas também ao número de caracteres, como vírgulas, pontos etc., presentes nele. Diante disso, chegamos à conclusão de que *tokens* no WST não equivale a *tokens* no SE, e sim a *words* (palavras). Inclusive, o número de *words* do LexTest no SE é 530.865, conforme podemos visualizar na Figura 10, sendo um número próximo do resultado de *tokens* do LexTest apresentado no WST: 552.903.

No que alude ao número de *types*, o WST reconhece 37.376 *types* (*distinct words*), palavras distintas, no *corpus* LexTest, como podemos visualizar na Figura 9. Conforme Berber Sardinha (2009, p. 175), *types* alude ao "total de itens, formas ou vocábulos do(s) arquivo(s), sem levar em conta as repetições". Ao analisarmos o SE, não conseguimos encontrar um termo ou resultado equivalente ao número de *types*, assim constatamos que o SE não apresenta esse tipo de informação em relação a um *corpus* inserido nele.

A ausência do número de *types* no SE impede que o usuário conheça a variação lexical do *corpus* – informação disponibilizada pelo WST por meio do item *type/token ratio* (TTR) visualizado na Figura 9. Sobre tal item presente no programa WST, Berber Sardinha (2009) esclarece que:

> *Type-Token Ratio*. É o resultado da divisão do total de '*types*' pelo total de '*tokens*', multiplicado por 100. A multiplicação por 100 serve para transformar o valor em porcentagem. Esse valor significa a extensão da variação lexical do texto. Um número maior indica uma variação maior, isto é, há menos repetições de palavras (do mesmo '*type*'); um número menor aponta para uma variação menor, pois há mais repetições do mesmo '*type*'. Em suma, quanto maior o seu valor, mais palavras diferentes o texto conterá. Em contraposição, um valor baixo indicará um número alto de repetições, o que pode indicar um texto menos 'rico' ou variado do ponto de vista de seu vocabulário. Por isso, ela é interpretada como uma medida da riqueza lexical do texto (BERBER SARDINHA, 2009, p. 175).

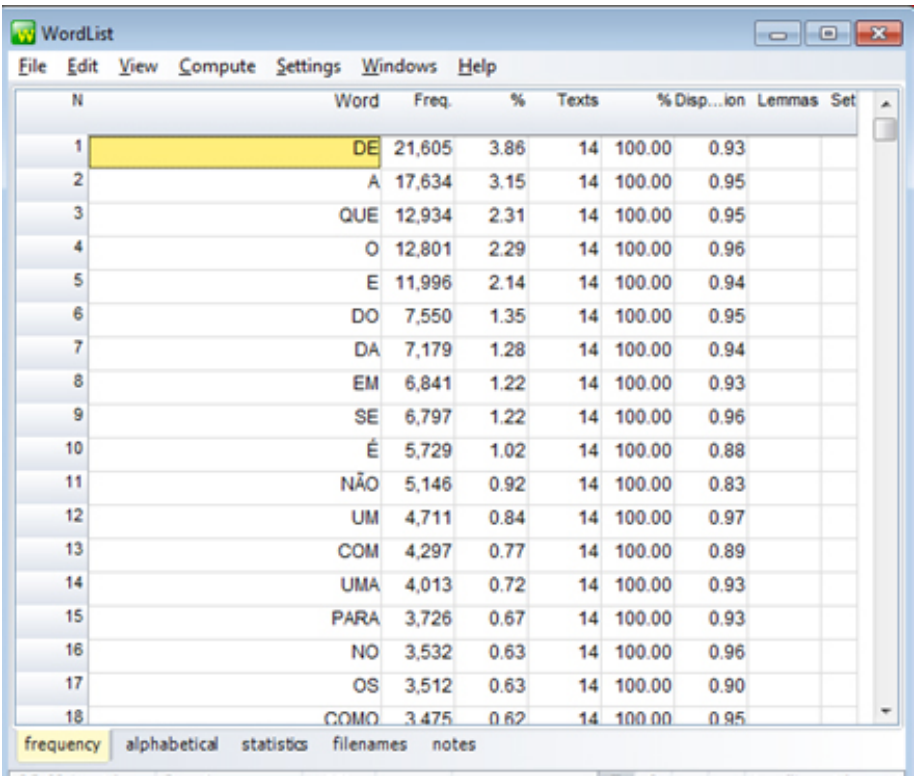
Acreditamos que o número de *types* e o cálculo que o relaciona à quantidade de *tokens* é fundamental para pesquisas que visem à análise

da riqueza lexical de um autor – como é o caso de investigações no campo da Literatura, da Tradução e da Estilística que lançam mão da metodologia/abordagem da LC. Pensando nisso, o uso do WST torna-se mais eficiente nesse ponto em comparação ao SE.

4.5 WST e SE: *WordList*

Após termos clicado na ferramenta *WordList* do WST (FIGURA 2) e realizado o processo de *upload* do LexTest no WST, geramos a *WordList* ilustrada na Figura 11.

FIGURA 11 – *WordList* parcial do LexTest no WST



N	Word	Freq.	%	Texts	%	Disp...ion	Lemmas	Set
1	DE	21,605	3.86	14	100.00	0.93		
2	A	17,634	3.15	14	100.00	0.95		
3	QUE	12,934	2.31	14	100.00	0.95		
4	O	12,801	2.29	14	100.00	0.96		
5	E	11,996	2.14	14	100.00	0.94		
6	DO	7,550	1.35	14	100.00	0.95		
7	DA	7,179	1.28	14	100.00	0.94		
8	EM	6,841	1.22	14	100.00	0.93		
9	SE	6,797	1.22	14	100.00	0.96		
10	É	5,729	1.02	14	100.00	0.88		
11	NÃO	5,146	0.92	14	100.00	0.83		
12	UM	4,711	0.84	14	100.00	0.97		
13	COM	4,297	0.77	14	100.00	0.89		
14	UMA	4,013	0.72	14	100.00	0.93		
15	PARA	3,726	0.67	14	100.00	0.93		
16	NO	3,532	0.63	14	100.00	0.96		
17	OS	3,512	0.63	14	100.00	0.90		
18	COMO	3,475	0.62	14	100.00	0.95		

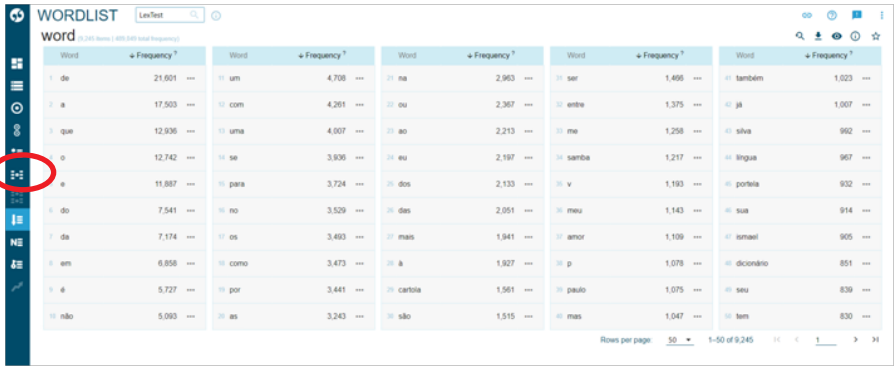
Fonte: Scott (2016).

No WST, é possível visualizar todas as palavras que compõem o LexTest juntamente com a posição delas apenas manuseando a barra de

rolagem. Além disso, conseguimos ver a lista de palavras por ordem de frequência (da maior para a menor e vice-versa) e por ordem alfabética.

No SE, no *dashboard* do LexTest, ao clicarmos em *WordList*, chegamos à tela da Figura 12.

FIGURA 12 – *WordList* parcial do LexTest no SE



Fonte: Kilgarriff e Rychlý (2003)

No SE, percebemos que a visualização da lista de palavras é feita por páginas. É possível ter acesso a, no mínimo, 10 palavras por página e a, no máximo, 200 palavras. Quando optamos por 200 palavras por página, o programa gera cinco colunas na mesma página. Assim como no WST, é possível ver essa lista por ordem de frequência e por ordem alfabética.

Apesar de o SE oferecer a opção de baixar a lista completa em PDF, XML, XLS, CSV, vemos essa questão da paginação como uma desvantagem em termos de acesso às informações de maneira ágil em relação ao WST. Vale ressaltar ainda que a exibição da *WordList* no SE por páginas dificulta o acesso ao momento em que as palavras chamadas *hápax legomena* (aquelas que ocorrem apenas uma vez no *corpus*) começam a aparecer na lista. Isso, de certa forma, torna-se prejudicial para pesquisas em que as *hápax legomena* são importantes, como a de Gonçalves (2006) e a de Grama (2016).

Notamos que é possível localizar as palavras com uma única ocorrência no SE por meio da busca avançada e do preenchimento da opção "número de ocorrência mínima e máxima: 1", porém o resultado é dado como uma lista isolada das demais palavras listadas no *corpus*.

Assim, a noção de posição das *hápax legomena* dentro do *corpus* se perde. Já no WST, para encontrarmos as *hápax legomena*, basta descermos a barra de rolagem até encontrarmos as palavras com uma única ocorrência na lista, procedimento simples, que mantém o vínculo das *hápax legomena* com o restante do *corpus* e a posição delas na ordem geral das ocorrências. Vale ressaltar que, em versões anteriores do WST, conseguíamos acessar a lista de palavras da menos frequente para mais frequente ao clicarmos no topo da lista, o que nos dava acesso às *hápax legomena* com apenas um clique. Na atual versão do WST, observamos que, ao clicarmos no topo da lista, o programa a organiza pela ordem de menor frequência, considerando também a ordem alfabética das palavras.

Outro ponto importante é que o WST é um programa que funciona sem internet, mas apresenta problemas caso o pesquisador faça alterações no *corpus* de estudo (mudança de local das pastas em que o *corpus* está armazenado, renomeação dessas pastas ou dos arquivos de texto do *corpus*, alteração de palavras nos textos que constituem o *corpus* etc.) e depois queira acessar uma *WordList* que tenha salvado antes disso, por exemplo. Observamos que, após o pesquisador começar a trabalhar no WST, ele deve evitar alterar as pastas nas quais o *corpus* está armazenado. Caso isso seja necessário, será preciso recomeçar a análise no programa. Além disso, se o computador utilizado estragar, provavelmente, todas as análises já realizadas e salvas na máquina estarão perdidas.

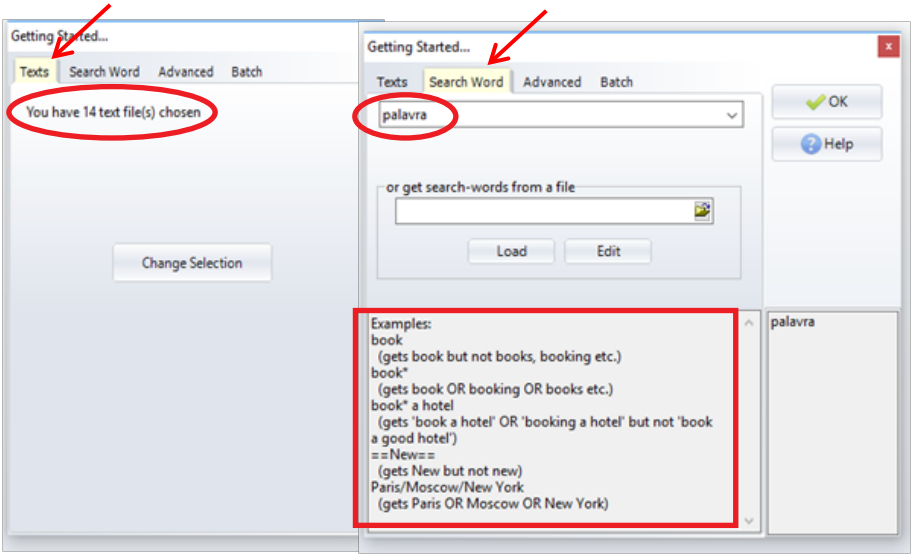
O SE, por ser uma plataforma virtual, exige a conexão com a internet, que pode ser lenta ou falha, por exemplo, dificultando, assim, o processo. No entanto, a vantagem é que, uma vez que o *corpus* é carregado, ele pode ser acessado em qualquer computador que possua internet. Dessa forma, o pesquisador não fica limitado a trabalhar em uma máquina específica como no caso do WST.

Acrescentamos que tanto o WST quanto o SE permitem o acesso às linhas de concordâncias através da *WordList* de maneira simples e rápida. No WST, basta selecionar determinada palavra da *WordList*, clicar na aba *Compute* e, depois, em *concordance*. No SE, é preciso clicar com o botão esquerdo do *mouse* em cima dos três pontinhos localizados no canto direito, ao lado do número referente à frequência de uma palavra da *WordList*, e acessar a opção *concordance*.

4.6 WST e SE: *Concord*

A ferramenta *Concord* do WST exige que o pesquisador carregue o *corpus* mesmo se ele já tiver feito esse procedimento para gerar uma *WordList*. Assim, após clicar em *Concord*, ilustrada na Figura 2, é preciso clicar em *File* e em *New*. Em seguida, o WST abre uma caixa de diálogo e, na aba *texts*, é possível carregar o *corpus* ao clicarmos em *Choose Texts Files* (escolha os arquivos de texto). Logo depois, o programa apresenta uma mensagem informando quantos arquivos foram carregados como na Figura 13 – imagem que ilustra o carregamento dos arquivos do LexTest.

FIGURA 13 – Aba *Texts* e *Search Word* da *Concord* no WST

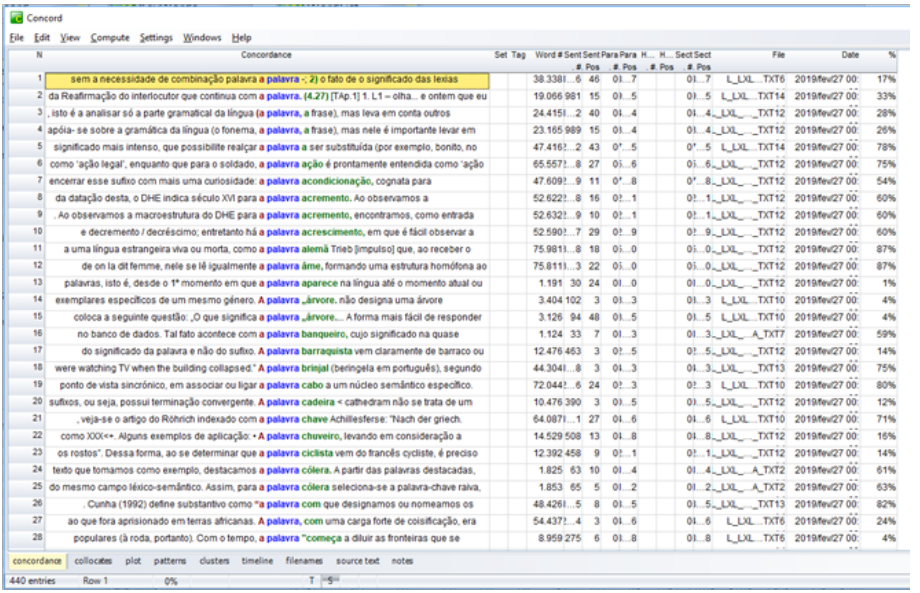


Fonte: Scott (2016).

Na próxima aba, *Search Word* (Busque palavra), visualizada também na Figura 13, há campos que permitem o pesquisador realizar diferentes buscas. Alguns modos de fazer as buscas são exemplificados pelo próprio programa em *Examples*. No primeiro campo em branco, destacado com um círculo vermelho na Figura 13, é possível digitar uma palavra específica que desejamos que o programa faça as linhas de concordâncias. Nesse caso, usar a ferramenta *Concord* é muito útil quando o pesquisador já sabe qual palavra, parte da palavra (radical ou desinência), expressão ou fraseologismo quer buscar no *corpus*. No caso

do LexTest, buscamos as linhas de concordância do item “palavra”. A Figura 14 ilustra parcialmente o resultado dessa busca.

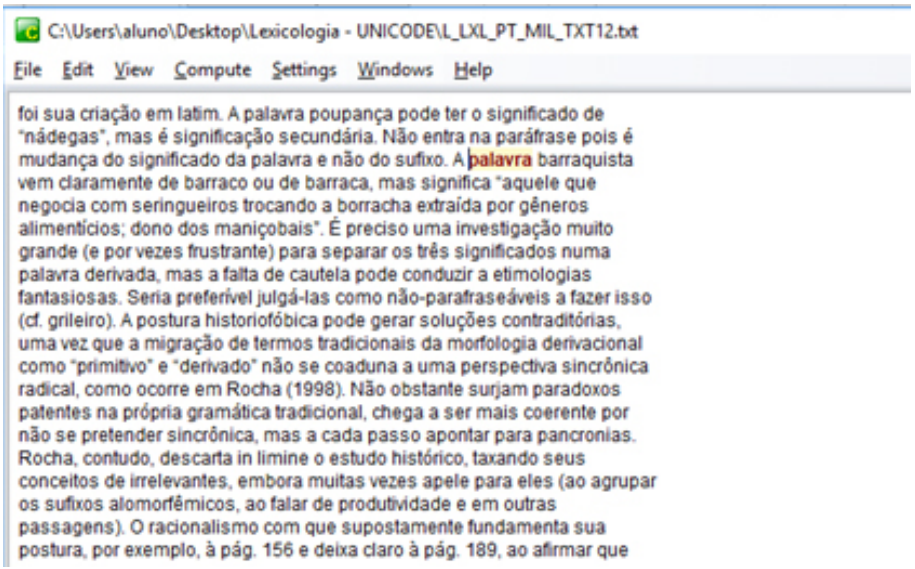
FIGURA 14 – Linhas de concordância parciais do substantivo “palavra” no WST



Fonte: Scott (2016).

Ao clicar duas vezes em cima de uma linha de concordância, é possível acessar o contexto linguístico referente à linha (FIGURA 15), o que é bastante produtivo para diversas pesquisas.

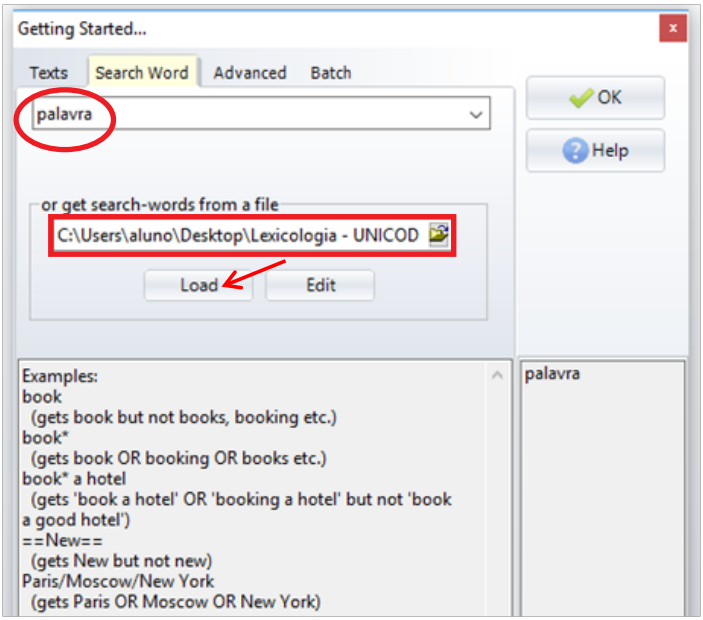
FIGURA 15 – Contexto linguístico de uma linha de concordância no WST



Fonte: Scott (2016).

O outro campo em branco da Figura 13, denominado *or get search-words from a file* (obter a busca de palavras de um arquivo), permite que a busca seja restringida a um único arquivo do *corpus*. Nesse caso, é preciso selecionar o arquivo e clicar em *Load* (carregar), conforme Figura 16.

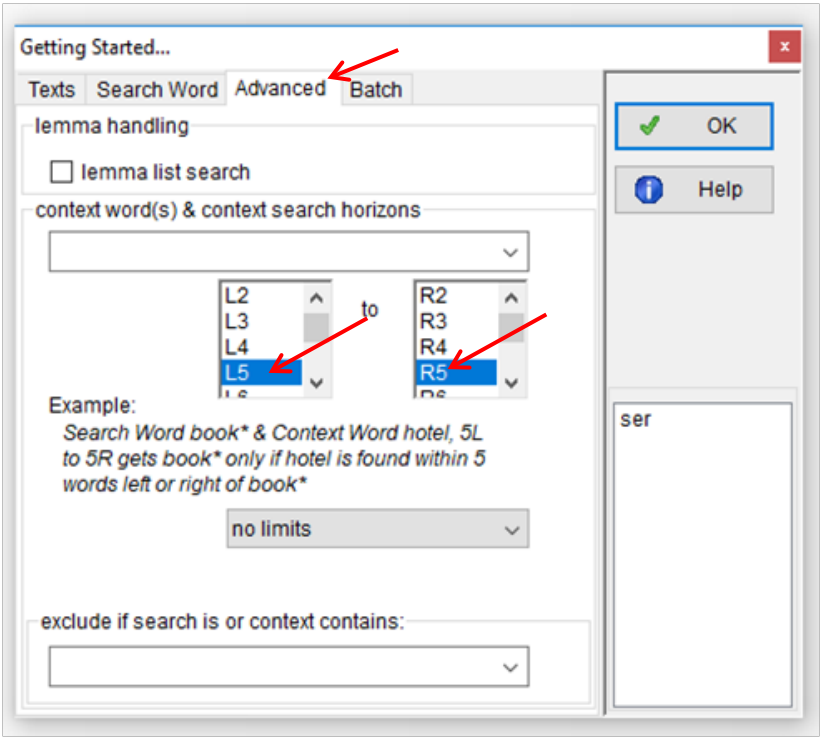
FIGURA 16 – *Get search-words from a file* da *Concord* no WST



Fonte: Scott (2016).

Na aba *Advanced*, podemos realizar buscas mais avançadas, definindo quantas palavras serão destacadas à direita ou à esquerda da palavra principal de busca, conforme podemos visualizar nas setas indicadas na Figura 17.

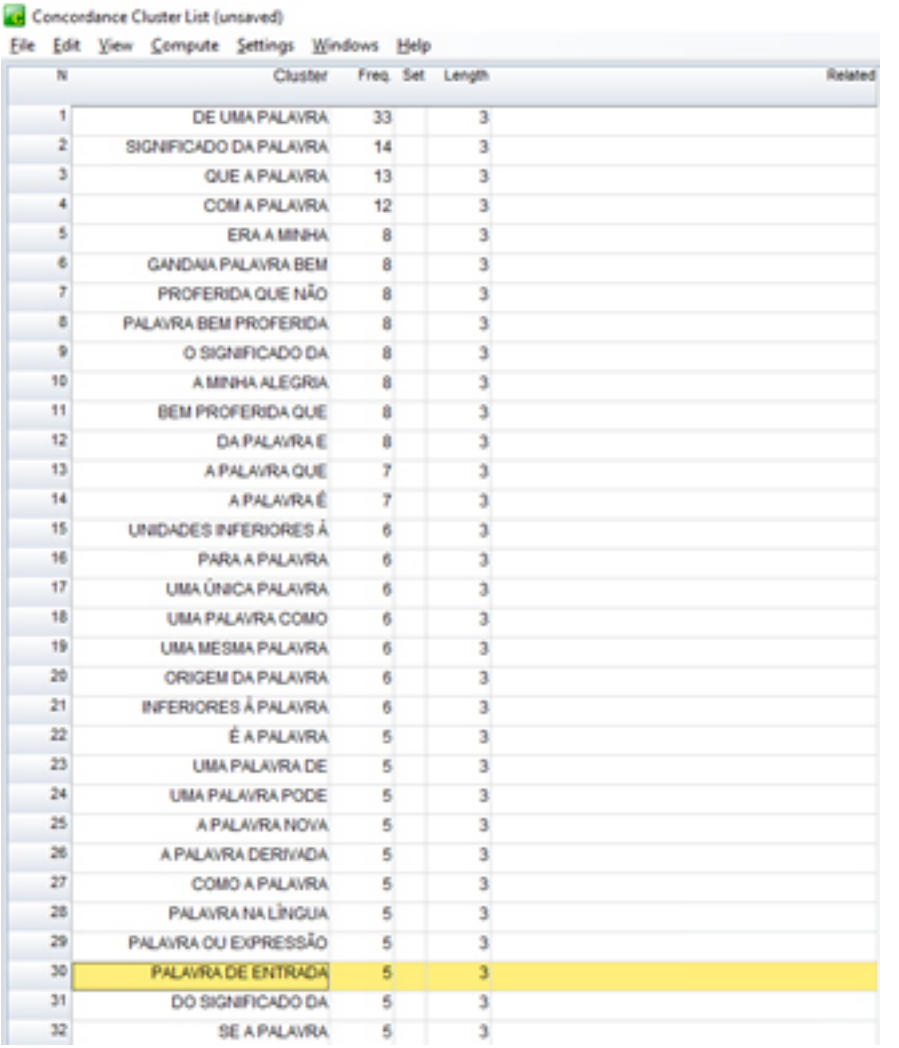
FIGURA 17 – Aba *Advanced* da *Concord* no WST



Fonte: Scott (2016).

Outra questão que devemos mencionar é que no WST não é possível acessar as linhas de concordância de um *cluster* (agrupamento de palavras) a partir da lista de *clusters* oferecida pelo programa. Isso, com certeza, dificulta análises que envolvem multipalavras, como as realizadas por Grama (2016). Na Figura 18, ilustramos uma lista de *clusters* que geramos a partir das linhas de concordância de "palavra". Nela, destacamos o *cluster* "palavra de entrada".

FIGURA 18 – Lista parcial de *clusters* no WST com destaque para "palavra de entrada"



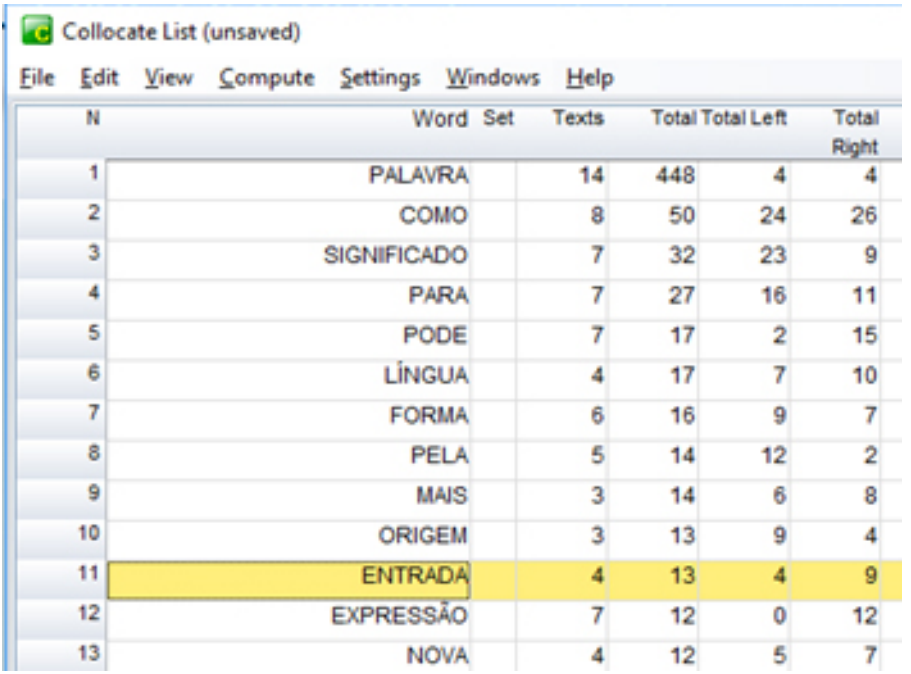
N	Cluster	Freq.	Set	Length	Related
1	DE UMA PALAVRA	33		3	
2	SIGNIFICADO DA PALAVRA	14		3	
3	QUE A PALAVRA	13		3	
4	COM A PALAVRA	12		3	
5	ERA A MINHA	8		3	
6	GANDAIA PALAVRA BEM	8		3	
7	PROFERIDA QUE NÃO	8		3	
8	PALAVRA BEM PROFERIDA	8		3	
9	O SIGNIFICADO DA	8		3	
10	A MINHA ALEGRIA	8		3	
11	BEM PROFERIDA QUE	8		3	
12	DA PALAVRA E	8		3	
13	A PALAVRA QUE	7		3	
14	A PALAVRA É	7		3	
15	UNIDADES INFERIORES À	6		3	
16	PARA A PALAVRA	6		3	
17	UMA ÚNICA PALAVRA	6		3	
18	UMA PALAVRA COMO	6		3	
19	UMA MESMA PALAVRA	6		3	
20	ORIGEM DA PALAVRA	6		3	
21	INFERIORES À PALAVRA	6		3	
22	É A PALAVRA	5		3	
23	UMA PALAVRA DE	5		3	
24	UMA PALAVRA PODE	5		3	
25	A PALAVRA NOVA	5		3	
26	A PALAVRA DERIVADA	5		3	
27	COMO A PALAVRA	5		3	
28	PALAVRA NA LÍNGUA	5		3	
29	PALAVRA OU EXPRESSÃO	5		3	
30	PALAVRA DE ENTRADA	5		3	
31	DO SIGNIFICADO DA	5		3	
32	SE A PALAVRA	5		3	

Fonte: Scott (2016).

Diante disso, observamos que, para acessar as linhas de um conjunto de palavras específico (no nosso caso, "palavra de entrada"), é preciso gerar as linhas de concordância de uma das palavras que compõem o *cluster* (escolhemos a palavra "palavra") e, em seguida, clicar

na aba *collocates* (colocados). Tal aba nos fornece uma lista de palavras que co-ocorrem com a palavra que fizemos as linhas de concordância. Assim, podemos escolher outra palavra do *cluster* (que esteja à direita ou à esquerda) para produzirmos, enfim, linhas de concordância que tenham mais chances de conter determinado agrupamento que visamos. A seguir, na Figura 19, ilustramos a lista de colocados que o WST gerou após termos feito as linhas de concordância de "palavra".

FIGURA 19 – Lista parcial de *collocates* de "palavra" na *Concord* do WST

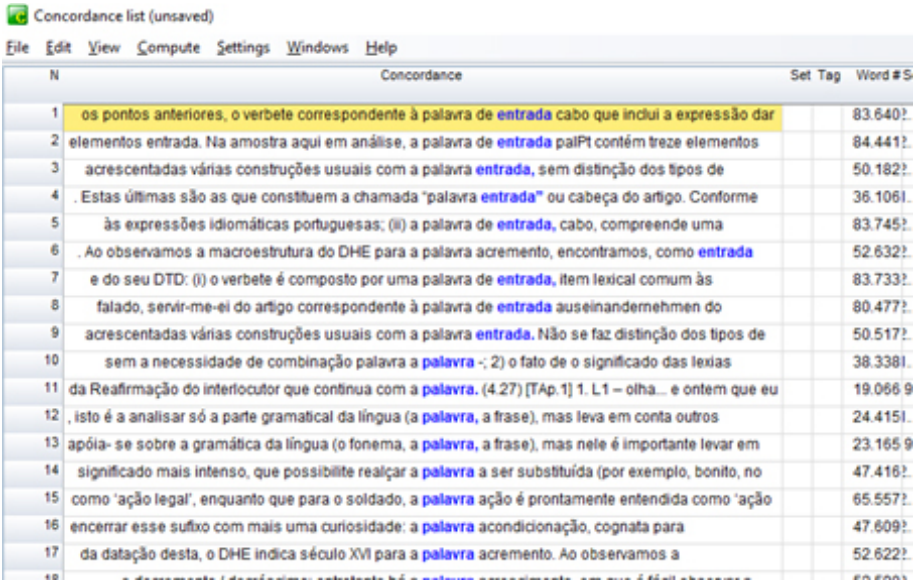


N	Word	Set	Texts	Total	Total Left	Total Right
1	PALAVRA		14	448	4	4
2	COMO		8	50	24	26
3	SIGNIFICADO		7	32	23	9
4	PARA		7	27	16	11
5	PODE		7	17	2	15
6	LÍNGUA		4	17	7	10
7	FORMA		6	16	9	7
8	PELA		5	14	12	2
9	MAIS		3	14	6	8
10	ORIGEM		3	13	9	4
11	ENTRADA		4	13	4	9
12	EXPRESSÃO		7	12	0	12
13	NOVA		4	12	5	7

Fonte: Scott (2016).

A linha que destacamos na Figura 19 apresenta a palavra "entrada" e a informação de que há nove ocorrências em que ela está à direita de "palavra". Ao clicarmos no número nove, o programa nos oferece as linhas de concordância da Figura 20, em que, finalmente, podemos visualizar as concordâncias do conjunto "palavra de entrada":

FIGURA 20 – Linhas de concordância parciais de "palavra de entrada" a partir da aba *collocates* na *Concord* do WST



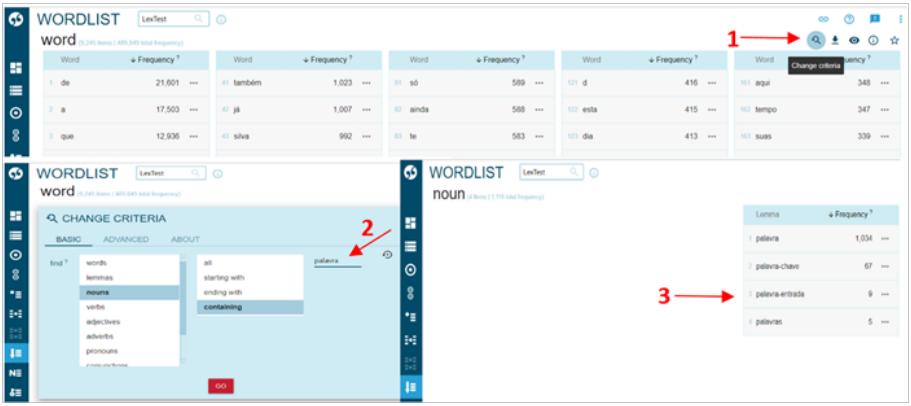
Fonte: Scott (2016).

No SE, a ferramenta *Concordance* encontra-se na lateral, mais precisamente no sexto símbolo circulado na cor vermelha na Figura 12. Diferentemente do WST, não é necessário carregar o *corpus* novamente para utilizar a *Concordance*.

Uma questão fundamental diz respeito ao fato de o próprio SE ter, à disposição do usuário, um vídeo que explica sobre a *Concordance* e os modos básicos de busca possíveis por meio dela. Nesse ponto, consideramos o SE mais didático e moderno em relação ao WST. Entretanto, na aba *Advanced* da *Concordance* no SE, não encontramos nenhum vídeo explicativo, o que nos causou certa frustração em virtude de o modo avançado ser mais complicado do que o básico. De qualquer maneira, prosseguimos com os testes.

A Figura 21 ilustra o nosso pedido de exibir todas as linhas de concordância do LexTest com ocorrências do substantivo "palavra" ao clicarmos em *nouns* na primeira coluna e em *containing* na segunda. Na sequência, vemos os resultados, em específico, as nove ocorrências de palavra + entrada.

FIGURA 21 – Aba *Advanced* na *Concord* do *SE*



Fonte: Kilgarriff e Rychlý (2003)

É perceptível que o SE realiza automaticamente a lematização das palavras, conforme mostra a Figura 22, em que vemos o singular e o plural do item "palavra" nas linhas de concordância.

FIGURA 22 – *Advanced* – Linhas de concordância e lematização na *Concord* do *SE*

Fonte: Kilgarriff e Rychlý (2003)

Nesse aspecto, o WST fica em desvantagem em relação ao SE, pois o processo de lematização no WST é feito manualmente, o que

demanda mais tempo do pesquisador, além de aumentar a chance de haver erros.

No que alude à visualização do contexto linguístico de uma linha de concordância, o SE mostra apenas um trecho de, no máximo, onze linhas, conforme Figura 23.

FIGURA 23 – Contexto linguístico de uma linha de concordância no SE

Fonte: Kilgarriff e Rychlý (2003)

Avaliamos que essa amostra de 11 linhas pode não ser suficiente para uma pesquisa que exija uma análise mais aprofundada dos elementos contextuais que cercam determinada palavra em uma ocorrência, principalmente se pensarmos em trabalhos que envolvam a área da Literatura, por exemplo.

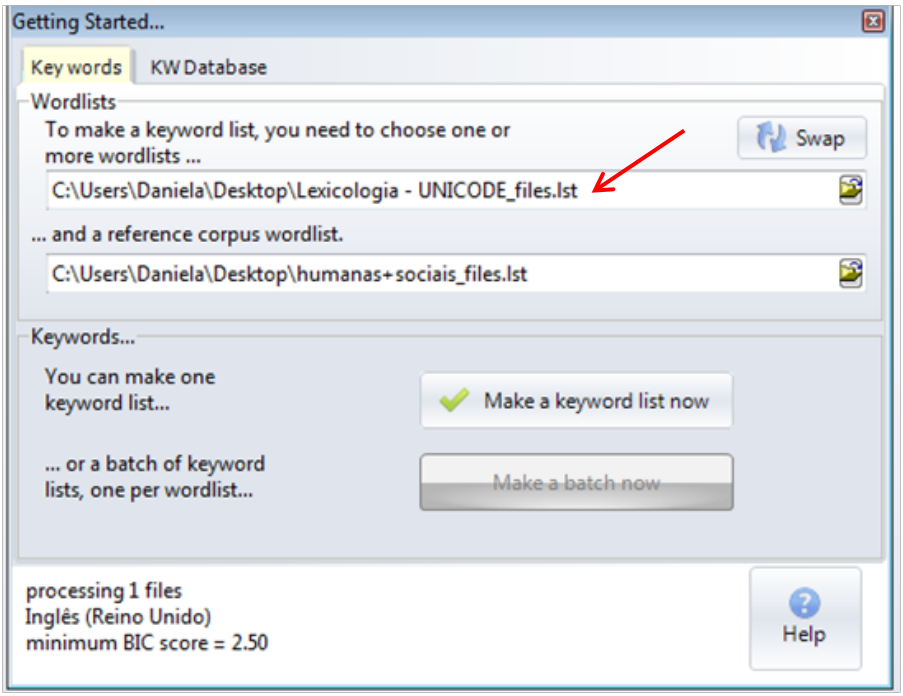
4.7 WST e SE: *KeyWords*

Ao acessarmos a ferramenta *Keywords* do WST, visualizada na Figura 2, clicamos em *File* e em *New*. Em seguida, o programa abre a janela da Figura 24 a seguir, em que consta a *WordList* que fizemos do nosso *corpus* e um campo, em branco, para que acrescentemos a *WordList* de um *corpus* de referência (que deve ser feita e salva anteriormente). A respeito do *corpus* de referência, Berber Sardinha (2004) sugere que ele seja no mínimo duas, três ou cinco vezes maior do que o *corpus* de estudo:

> O tamanho do *corpus* de referência influencia a quantidade de palavras-chave obtidas. Os tamanhos críticos de corpora de referência são 2, 3 e 5 vezes o tamanho do *corpus* de estudo. *Corpora* de referência com essas dimensões retornam significativamente mais palavras-chave do que *corpora* de tamanhos menores (BERBER SARDINHA, 2004, p. 100).

Tendo isso em vista, o *corpus* de referência que escolhemos para ser contrastado com o LexTest é o *corpus* da área de humanas e de sociais do Lácio *Web*, uma vez que juntos cumprem o critério de ser cinco vezes maior que o LexTest. No caso, o LexTest apresenta 552.903 *tokens*, e o *corpus* de referência que utilizamos possui 2.973.094 *tokens*.

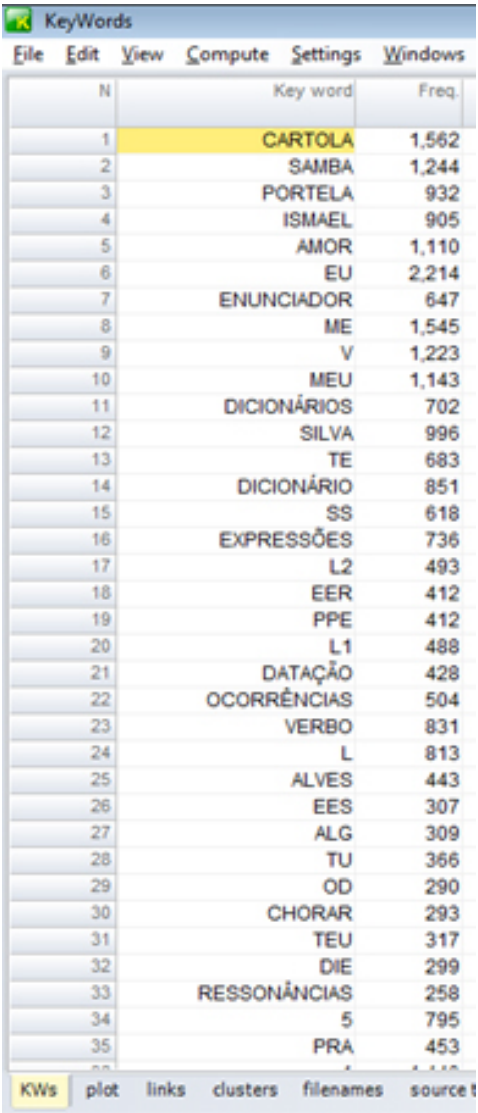
FIGURA 24 – Inserção do *corpus* de referência no WST



Fonte: Scott (2016).

Na Figura 25, há a lista parcial de palavras-chave que realizamos no WST com o LexTest.

FIGURA 25 – Resultado parcial da *Keywords* no WST



N	Key word	Freq.
1	CARTOLA	1,562
2	SAMBA	1,244
3	PORTELA	932
4	ISMAEL	905
5	AMOR	1,110
6	EU	2,214
7	ENUNCIADOR	647
8	ME	1,545
9	V	1,223
10	MEU	1,143
11	DICIONÁRIOS	702
12	SILVA	996
13	TE	683
14	DICIONÁRIO	851
15	SS	618
16	EXPRESSÕES	736
17	L2	493
18	EER	412
19	PPE	412
20	L1	488
21	DATAÇÃO	428
22	OCORRÊNCIAS	504
23	VERBO	831
24	L	813
25	ALVES	443
26	EES	307
27	ALG	309
28	TU	366
29	OD	290
30	CHORAR	293
31	TEU	317
32	DIE	299
33	RESSONÂNCIAS	258
34	5	795
35	PRA	453

Fonte: Scott (2016).

No SE, não é necessário ter um *corpus* de referência, uma vez que a própria plataforma oferece vários *corpora* de referência prontos para serem utilizados pelo pesquisador, o que é positivo na medida em que

poupa tempo e trabalho. No entanto, a informação sobre o tamanho dos *corpora* de referência disponibilizados pelo SE não consta no momento em que o usuário vai escolher qual *corpus* de referência usará.

Para nos certificarmos sobre o tamanho do *corpus* de referência que escolhemos no SE, no nosso caso, do *Brazilian Portuguese Corpus* (*Corpus* Brasileiro), tivemos de selecioná-lo no *dashboard*, no campo *Recently used corpora* (*Corpora* recentemente usados), e, depois, clicar em *Corpus info* (Informações do *corpus*). Nesse aspecto, consideramos que a ausência de uma sequência lógica ou intuitiva para saber o tamanho do *corpus* de referência escolhido no SE é prejudicial ao usuário, na medida em que ele precisa retornar ao *dashboard* e fazer todo esse processo que descrevemos.

No SE, também testamos a possibilidade de realizarmos o *download* do *corpus* de referência que utilizamos. Para isso, clicamos em *dashboard* > *manage corpus* > *download*, mas apareceu uma mensagem que esclarecia a necessidade de contatar um administrador para realizar o *download*. Pensando no recorte deste artigo, o fato de não podermos efetuar o *download* do *corpus* de referência oferecido pelo SE não foi benéfico, já que pensamos em comparar os resultados da *Keywords* entre o WST e o SE, lançando mão de um mesmo material de análise.

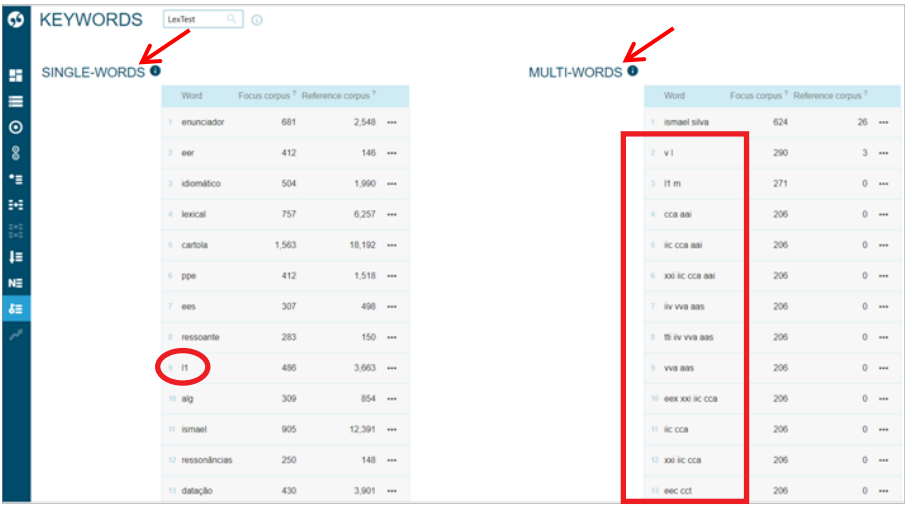
Diante disso, pensamos em inserir no SE o *corpus* de referência (das áreas de humanas e sociais do Lácio *Web*) que utilizamos no WST. Ao colocarmos tal ideia em prática, descobrimos que, no SE, caso o pesquisador queira usar um *corpus* de referência que não seja disponibilizado pelo próprio SE, é preciso inseri-lo como um novo *corpus* (*dasboard* > *new corpus*), mas, vale pontuar que, para isso, é necessário ter uma franquia de dados que suporte o tamanho do *corpus* de referência desejado, o que demanda recursos financeiros. No nosso caso, não conseguimos finalizar o procedimento justamente por termos excedido o nosso limite de dados.

Tendo isso em vista, chegamos à conclusão de que é importante que o pesquisador pense na questão do limite de dados ao optar pelo SE. Se ele necessita inserir um *corpus* de referência no SE, talvez seja melhor lançar mão do WST, pois pagará apenas pela licença (taxa única) do WST, e não pela quantidade de dados (plano mensal) que carregará no *software* como ocorre no SE.

Há algumas formas de gerar as *Keywords* no SE. O primeiro modo que testamos foi o básico, que é sem filtro. Após clicarmos em

Dashboard e, na sequência, em *Keywords*, chegamos a uma tela na qual clicamos na aba *Basic* e, depois, em *Go*. Nessa situação, o *SE* escolhe automaticamente o *corpus* de referência, por isso não conseguimos saber qual *corpus* o programa utilizou como referência nem se a língua desse *corpus* era compatível com o nosso *corpus* de análise, no nosso caso, português brasileiro. Ademais, salientamos o fato de que também não conseguimos saber qual (is) variedade (s) (português vernacular, falado, acadêmico etc.) compunham o *corpus* de referência. O resultado parcial da *Keywords* aparece na Figura 26:

FIGURA 26 – Resultado parcial da *Keywords* automática do LexTest no SE



Fonte: Kilgarriff e Rychlý (2003)

Como podemos observar, o SE gerou duas listas – uma de *single-words* (unipalavras) e outra de *multi-words* (multipalavras). Para nós, essa divisão, por um lado, pode facilitar o trabalho daqueles que querem focar apenas nas uni ou multipalavras, mas, por outro lado, pode não ser bem-vinda para aqueles que não têm interesse em privilegiar uma ou outra, uma vez que tal separação nos impede de saber a real posição da palavra, seja ela uni ou multi, em termos de importância na lista das palavras-chave.

Notamos que, dentre as primeiras palavras-chave, aparecem números, por exemplo, na posição nove da coluna da esquerda (*single-words*), e algarismos romanos, nas posições de dois a 13 da coluna da direita (*multi-words*). Acreditamos que esse tipo de resultado seja

decorrente do processo de compilação e processamento (etiquetagem, lematização) automático de *corpora* que o SE realiza. Assim, o que, em primeiro momento, parece ser vantajoso – etiquetagem e lematização automática de *corpus* – procedimento obrigatório ao carregar um *corpus* no SE – na verdade, pode acarretar problemas como esse.

Outra maneira de gerarmos as *Keywords* é por meio do modo avançado. Percorremos o mesmo caminho descrito anteriormente, com a diferença de que, ao invés de clicarmos na aba *Basic*, clicamos na *Advanced*. Ao iniciar o processo de geração da lista de palavras-chave no modo *Advanced*, o SE apresenta opções para a escolha do *corpus* de referência que será utilizado no processamento, porém observamos que uma das opções é o nosso próprio *corpus* de testagem, o LexTest. Esse fato é contrário aos parâmetros postulados por Berber Sardinha (2004), que sugere que o *corpus* de referência seja, no mínimo, duas, três ou cinco vezes maior do que o *corpus* de estudo.

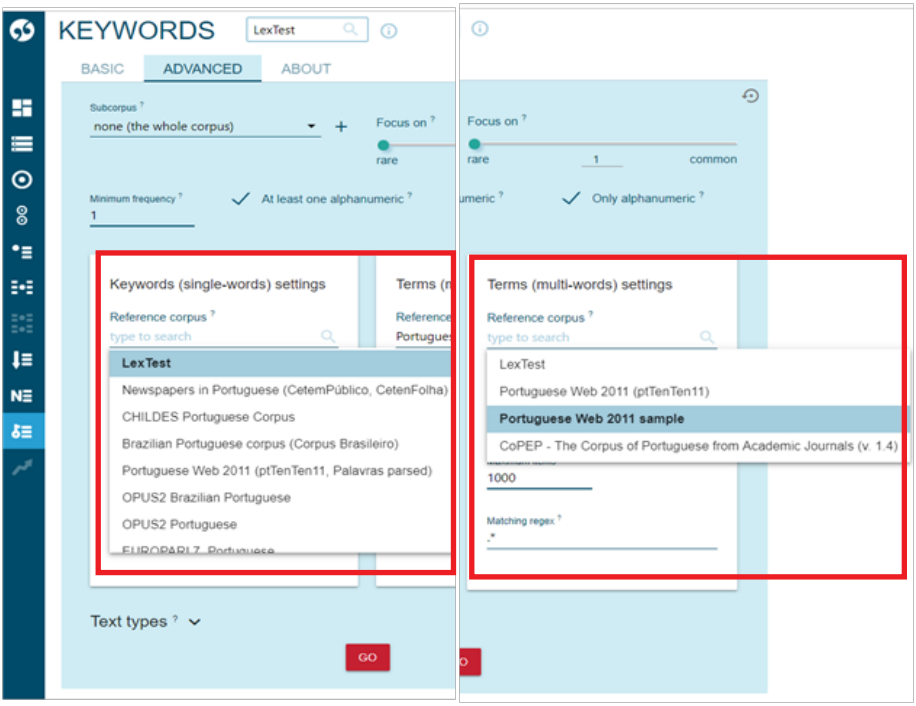
O próprio conceito de lista de palavras-chave está relacionado a um cálculo contrastivo entre a proporção que as palavras ocorrem em um *corpus* investigado e em um *corpus* de referência. Ademais, Berber Sardinha (2004) explica o motivo pelo qual um *corpus* de referência não deve ser constituído por textos do *corpus* de estudo:

> O *corpus* de referência não deve conter o *corpus* de estudo, pelo menos não deliberadamente e por completo. Há duas razões para isso. A primeira refere-se aos valores absolutos: devido à soma das frequências, as mais salientes no *corpus* de estudo tendem a obscurecer, e portanto, a deixar de indicar palavras-chave. Por exemplo, se no *corpus* de estudo a palavra *casa* tem frequência 10, e no *corpus* de referência l, a diferença será grande (10) e possivelmente significativa, ou seja, a palavra *casa* tem chances de ser chave. Mas se o *corpus* de estudo for adicionado ao de referência, as frequências passam a ser 10 no *corpus* de estudo e 11 no de referência, ou seja, uma diferença de apenas 1, o que diminui as chances de a palavra ser chave. A segunda razão diz respeito às frequências relativas: a soma pouco altera a diferença entre as porcentagens, e é, portanto, desnecessário unir os *corpora* (caso o *corpus* de referência seja 5 vezes maior no mínimo; ver discussão abaixo acerca do valor crítico de 5). Tomando-se o exemplo anterior, se o *corpus* de estudo possuir 100 itens, a frequência 10 de *casa* seria 10% (10/100), e se o *corpus* de referência tiver 500 itens, a frequência l seria equivalente a 0,2% (l/5OO). Juntando os *corpora*, a frequência no *corpus* de referência passa a ser de 11, ou l,

> 8% (11/600), ou seja, a palavra *casa* ainda continua com propensão a ser chave. Com palavras de frequências menos discrepantes, a diferença também pouco altera a propensão à chavicidade. Por exemplo, se em vez de 10, *casa* tiver frequência 1 no *corpus* de estudo, as porcentagens antes da união dos *corpora* seriam 1% no *corpus* de estudo (1/100) e 0,2% no de referência (1/500); depois da união, a frequência no *corpus* de referência passaria a 0,3% (2/600), pouco aumentando as chances de chavicidade da palavra *casa* (BERBER SARDINHA, 2004, p. 100).

Além disso, o SE não permitiu que usássemos o mesmo *corpus* de referência para gerarmos as *single-words* e as *multi-words*, conforme ilustra a Figura 27.

FIGURA 27 – Opções de *corpus* de referência para *single-words* e *multi-words* no SE

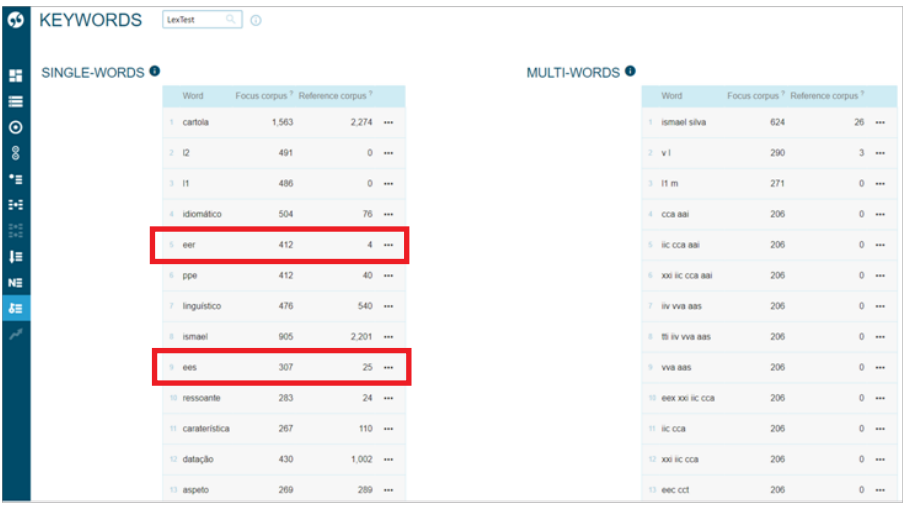


Fonte: Kilgarriff e Rychlý (2003)

Inferimos que essa diferenciação quantos aos *corpora* de referência entre as duas listas de palavras-chave pode acarretar problemas para a análise dos dados de uma pesquisa, já que, em se tratando do mesmo conjunto de dados em análise, é necessário utilizar o mesmo tratamento, ou seja, o mesmo *corpus* de referência. Conforme já mencionamos, a separação das listas de resultados atrapalha a análise, pois ocasiona a perda da real posição das palavras-chave em relação ao todo. Além do mais, em decorrência dos cálculos de chavicidade serem feitos com base em *corpora* de referência distintos para uni e multipalavras, o critério para determinação do índice de chavicidade fica comprometido.

Mesmo assim, prosseguimos com o teste. Para as *single-words*, optamos pelo *corpus* de referência *Brazilian Portuguese corpus* (*Corpus* Brasileiro) e, para as *multi-words*, escolhemos o *Portuguese Web 2011 sample*. Ao realizarmos as *KeyWords* com tais *corpora* de referência, sugeridos pelo SE, por meio da aba *Advanced*, chegamos à tela da Figura 28.

FIGURA 28 – Resultado parcial da *Keywords* no modo *Advanced* no SE



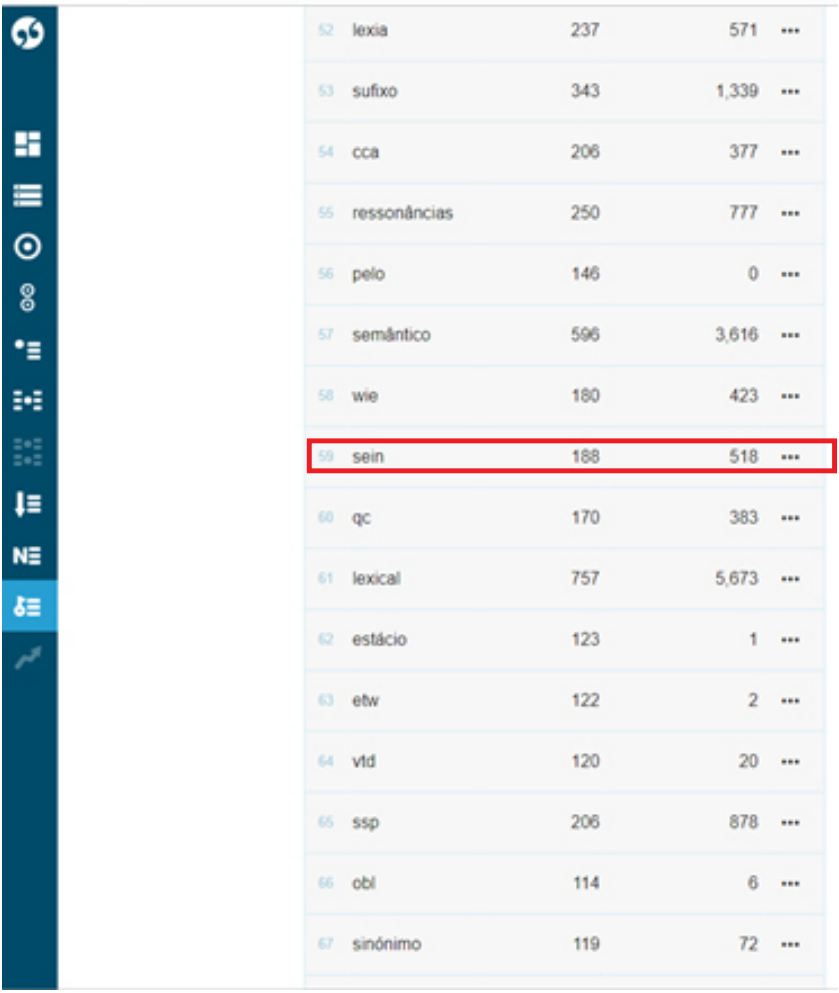
Fonte: Kilgarriff e Rychlý (2003)

Em primeiro lugar, percebemos que, na lista da Figura 28, há a presença de itens que não reconhecemos qualitativamente como palavras da língua portuguesa, como as que destacamos com setas (“eer” e “ees”). Uma hipótese que levantamos é a de que não foi feita a limpeza dos

corpora de referência no processo de compilação deles, processo que, inclusive, pode ter sido realizado por meio de extrator automático.

Em segundo lugar, observamos que, embora o LexTest tenha sido etiquetado automaticamente pelo SE na língua portuguesa, alguns itens, supostamente, da língua alemã foram listados dentre as palavras mais frequentes da lista de palavras-chave gerada. Na Figura 29, destacamos a palavra *sein* como exemplo.

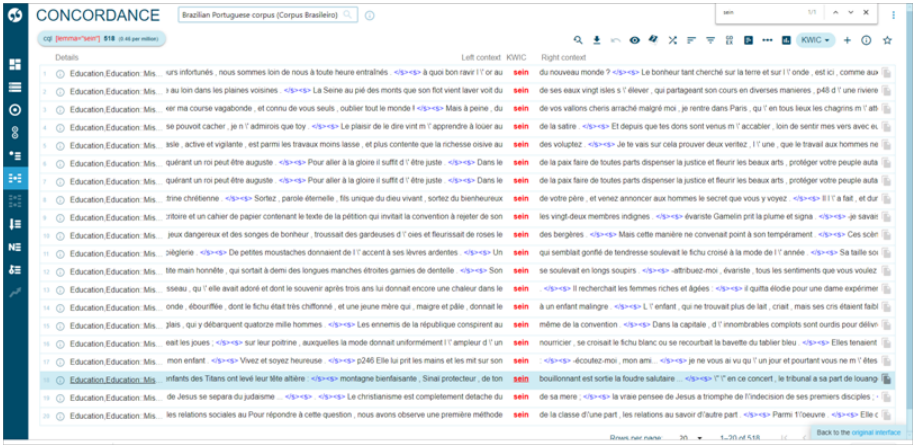
FIGURA 29 – Resultado parcial da *Keywords* no SE com destaque para a palavra *sein*



Fonte: Kilgarriff e Rychlý (2003)

O que parece ser o verbo em alemão *sein* (ser), com 518 ocorrências no *corpus* de referência, na realidade, diz respeito a uma palavra da língua francesa, conforme mostram as linhas de concordância da Figura 30.

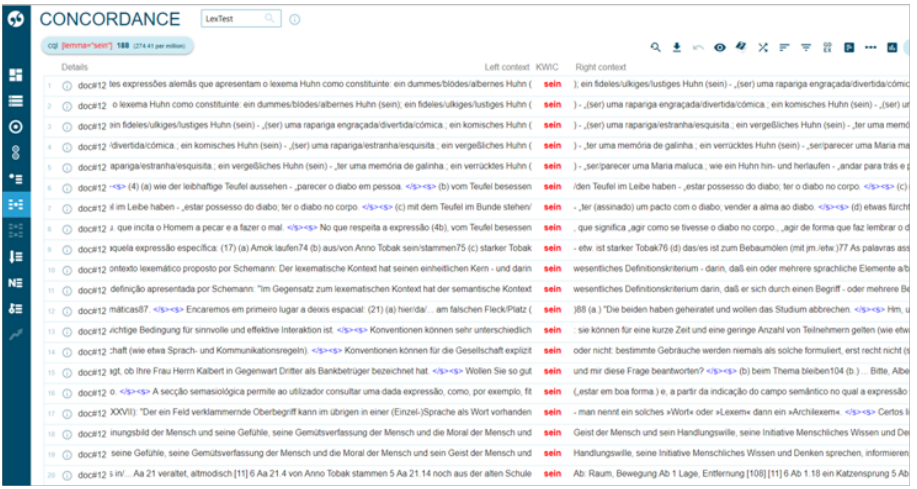
FIGURA 30 – Linhas de concordância parciais de *sein no corpus* de referência no SE



Fonte: Kilgarriff e Rychlý (2003)

No LexTest, ou *focus corpus* (*corpus* de estudo), o verbo em alemão *sein* ocorreu 188 vezes. Na Figura 31, mostramos as linhas de concordância dele.

FIGURA 31 – Linhas de concordância parciais de *sein* do LexTest no SE



Fonte: Kilgarriff e Rychlý (2003)

Por termos obtido tal resultado, questionamo-nos por qual motivo os *corpora* de referência sugeridos pelo SE possuem palavras que não são do idioma português, uma vez que os *corpora* de referência são de língua portuguesa. Ademais, do nosso ponto de vista, o SE deveria ter reconhecido os itens que não pertencem à língua portuguesa no processo de etiquetagem automática do nosso *corpus* de estudo.

A respeito da etiquetagem automática e do processamento eletrônico de *corpora* efetuados pelo SE, Davies (2019) avalia que, para anotar um *corpus*, é necessário que alguém saiba realmente o idioma em questão, no nosso caso, o português. O referido pesquisador, com base nos problemas de precisão do SE identificados por ele, acredita que não há um responsável que realmente domine a língua portuguesa. Segundo Davies (2019), "eles simples e cegamente rodaram o *tagger* nos *corpora* e depois os colocaram *on-line*, com pouca ou nenhuma tentativa de consertar as coisas. Rápido, mas não muito útil" (DAVIES, 2019).[9]

[9] No original: "They simply blindly ran the tagger on the corpora and then placed them online, with little or no attempt to fix things. Quick, but not very helpful".

Davies (2019) realizou um teste com lemas do português e apontou alguns problemas resultantes do *tagging* e da lematização do SE, segundo ele, "imprecisos". A seguir, apresentamos um excerto da análise realizada por ele:

> Dando uma olhada apenas nos verbos, descobrimos que mais do que pelo menos 46 desses 68 'verbos' frequentes não são realmente verbos (e esses são supostamente 'verbos' comuns – ocorrendo 1000 vezes ou mais). Alguns deles são formas de verbos portugueses (saiu, saímos, selecionaram, sabíamos, sorocaba), mas, na verdade, não são lemas (ou seja, a entrada que se encontraria em um dicionário). Alguns destes pelo menos terminam em -r, o que sugere que eles podem ser verbos portugueses em algum universo alternativo (secalhar, sanduichar, sinistrar, sair, siar, sapar, futebol), mas eles não são realmente palavras neste universo. E outros claramente nunca poderiam ser verbos (pelo menos em português, a língua do *corpus*): sensei, sibutramina, simpática, sm, sábados, semiárido, sobrevivência, simples, prata, amostra. Se fôssemos mais adiante na lista – palavras que ocorrem 100-200 vezes, por exemplo – descobriríamos que quase 90% de todas as entradas são problemáticas. [...] Mas mesmo com esses 'verbos' muito frequentes (que ocorrem 1000-2000 vezes cada um no *corpus*), os dados são extremamente confusos (DAVIES, 2019).[10]

Davies (2019) conclui que, se vão ser criados dados de frequência de palavras, eles precisam ser revisados cuidadosamente. Isso significa que deve haver verificação dos contextos, correção de lemas, dentre

[10] No original: "Taking a look at just the verbs, we find that more than at least 46 of these 68 frequent "verbs" aren't really verbs at all (and these are supposedly common «verbs» – occurring 1000 times or more). Some of them are forms of Portuguese verbs (saíu, saimos, selecionaram, sabiamos, sorocaba), but they are not actually lemmas (i.e. the entry that one would find in a dictionary). Some of these at least end in an -r, which would suggest that they might be Portuguese verbs in some alternate universe (secalhar, sanduichar, sinistrar, saír, siar, sapar, soccer), but they are not actually words in this universe. And others clearly could never be verbs (at least in Portuguese, the language of the corpus): sensei, sibutramina, simpatica, sm, sabados, semiárido, sobrevivencia, simple, silver, sample. If we were to go further down the list – words that occur 100-200 times, for example – we would find that nearly 90% of all of the entries are problematic.n[...] But even with these very frequent "verbs" (which occur 1000-2000 times each in the corpus), the data is extremely messy."

outras ações, com base, pelo menos, em um conhecimento rudimentar da língua trabalhada. Para ele, essa revisão parece não ter sido feita no SE em virtude da quantidade de problemas detectados em seu teste.

Por fim, observamos que é possível favoritar e fazer o *download* da *KeyWord list* no SE. O SE possui a opção de salvar a *KeyWord list*, porém, dependendo da opção de saída do arquivo, o SE não salva a lista completa, por exemplo, ao salvarmos em arquivo PDF, obtivemos somente o *download* da primeira página de resultados. Já na opção XLS (planilha Excel), conseguimos salvar a lista completa, que vai até 1001 palavras. Também não localizamos uma opção para aplicar *stop list* nem fazer a limpeza da *KeyWord list*. A seguir, tecemos as nossas considerações finais.

Considerações finais

Podemos dizer que tanto o WST quanto o SE são eficientes no que se propõem, mas é preciso que o pesquisador, com base nos objetivos de sua pesquisa, saiba escolher qual deles atende melhor às suas necessidades. Por isso, elaboramos o Quadro 2, que contém um resumo da análise comparativa que desenvolvemos ao longo deste artigo. O símbolo de adição (+) está presente nos itens que consideramos mais positivos em cada *software* e o de subtração (-) consta nos itens que percebemos que podem ser aperfeiçoados em cada um. No campo observações, esclarecemos o motivo pelo qual atribuímos os símbolos de adição e/ou de subtração a determinado quesito e, em alguns casos, comentamos adicionalmente algum ponto que pode ser melhorado no WST ou no SE. Acreditamos que esse resumo possa auxiliar pesquisadores a tomarem uma decisão no que alude ao uso de um desses dois programas de análise lexical.

QUADRO 2 – Resumo da análise comparativa entre WST e SE

Itens de análise comparativo-descritiva	WST	SE	Observações
***Software* instalável em computador pessoal**	+	–	O SE demanda acesso à internet e, cada vez mais, recursos financeiros dependendo do tamanho do *corpus* de estudo e de referência.
Software on-line	–	+	O WST precisa ser instalado em um computador pessoal, o que dificulta o processamento de *corpus* nele em qualquer máquina.
Interface amigável	+	+	É possível que o usuário tenha certa familiaridade com ambos, visto que foram projetados conforme os moldes do *Windows*. Contudo, vale lembrar que o SE requer que o usuário tenha noções básicas prévias da lógica da Linguística de *Corpus* ou que ele curse os treinamentos oferecidos pelo próprio SE.
Quantidade de informações na interface	+	–	O WST apresenta seis bot**ões, enquanto o SE apresenta 13, o que o torna mais complexo.**
Idioma de interação com o *software*	–	–	Ambos foram elaborados apenas em língua inglesa.
Terminologia adotada pelo *software*	+	–	O SE denomina como *compile* o processo de etiquetagem e lematização automática do *corpus* e como *create corpus* o que, na verdade, é o procedimento de *upload* de um *corpus*.
Configuração de língua de acordo com o *corpus* de estudo	+	+	Ambos os *softwares* apresentam esse recurso. É importante salientar que, no SE, a configuração de língua faz parte do procedimento de inserção do *corpus* de estudo, e isso faz com que o usuário não se esqueça dela, o que é positivo e **não ocorre no WST.**
Flexibilidade no quesito extensão e codificação dos arquivos que compõem o *corpus* de estudo	–	+	O WST aceita apenas arquivos TXT e com codificação *Unicode*.
Fornecimento de *corpus* de estudo	–	+	O WST não fornece um *corpus* de estudo pronto para análise.
Fornecimento de *corpus* de referência	–	–	O WST não fornece um *corpus* de referência pronto para uso. O SE não fornece o mesmo *corpus* de referência para processar *uni* e *multi-words*.
Agilidade no carregamento do *corpus*	+	–	O SE demandou um minuto e quatro segundos para carregar o LexTest, enquanto o WST levou apenas dois segundos.

Armazenamento do *corpus*	–	+	O WST não armazena o *corpus*. O armazenamento do *corpus* deve ser realizado no computador pessoal. As pastas e arquivos onde os textos estão armazenados não podem sofrer alterações, pois isso prejudica o acesso aos resultados já salvos pelo pesquisador, como lista de palavras e linhas de concordância.
Etiquetamento do *corpus*	–	+	O WST não realiza etiquetamento automático.
Tokens	+	+	Ambos os *softwares* apresentam o número de *tokens*.
Types	+	–	O SE não disponibiliza o número de *types*.
Visualização da lista de palavras	+	–	A visualização da lista de palavras no SE, por meio de paginação, não propicia agilidade ao pesquisador.
Visualização de resultados	+	–	A visualização de resultados divididos em *single* e *multi words* no SE faz com que a real posição das palavras seja perdida. Não há outro modo de configurar os resultados.
Acesso ao contexto linguístico das linhas de concordância	+	–	A visualização do contexto linguístico é limitada no SE – conseguimos ver o máximo de 11 linhas para os contextos verificados.
Acesso às linhas de concordância dos *clusters*	–	+	O WST não dá acesso às linhas de concordância dos *clusters*, sendo necessário lançar mão da aba *collocates*.
Processamento de palavras-chave	+	+	Ambos calculam e apresentam as palavras-chave do *corpus* de estudo.
Explicações sobre como usar a ferramenta/ dicas de uso ferramenta	–	–	Tanto o WST quanto o SE podem melhorar na quantidade de informações que disponibilizam sobre o uso das ferramentas e no que diz respeito ao idioma em que as explicações são propagadas.

Fonte: Elaboração própria.

Cabe destacar que, por questões de viabilidade, não conseguimos analisar e comparar todas as ferramentas e funcionalidades que constam no WST e no SE. Apenas testamos as principais: *Word List*, *KeyWords List* e *Concordance*, o que significa que novas pesquisas contrastivas precisam ser desenvolvidas para apurar todo o potencial de funcionamento dos dois *softwares*.

Além disso, tanto o WST quanto o SE passam por constantes atualizações, em especial, o SE, que é uma plataforma *on-line*, logo

a análise comparativa que efetuamos diz respeito a um estágio de desenvolvimento específico dos *softwares* e está restrita ao *corpus* de testagem LexTest. Também é importante mencionar que tivemos mais dificuldade em trabalhar com o SE do que com o WST, talvez porque já somos usuários do WST desde 2014 e do SE apenas desde 2017. Nesse sentido, vale lembrar que é necessário que o pesquisador esteja constantemente utilizando as ferramentas para que não se esqueça de como elas funcionam e para que adquira certa familiaridade com elas.

Considerando nosso processo de aprendizagem para executar os testes e as tarefas descritas neste artigo, sendo iniciantes no SE, podemos dizer que enfrentamos dificuldades para descobrir como realizar alguns procedimentos. Em algumas situações, como a configuração de língua, a localização das legendas das etiquetas e das informações a respeito dos *corpora* de referência, dentre outras, inferimos onde poderiam estar localizadas algumas funções no SE porque já possuíamos conhecimentos prévios em LC.

A nosso ver, o WST é organizado de maneira lógica, objetiva e intuitiva, o que facilita a localização rápida das principais funções do programa feita pelo usuário; já o SE não destaca os principais recursos, dentre os diversos que disponibiliza ao pesquisador, o que dificulta a usabilidade da ferramenta. Observamos ainda que o SE oferece cursos *on-line* para que o consulente aprenda a usar seus recursos e se torne um usuário capacitado.

A partir dessa experiência, acreditamos que, embora o SE faça o processamento automático de algumas etapas da pesquisa em LC, ele exige uma considerável curva de aprendizagem, dificuldades que não foram tão sobressalentes quando de nossas experiências iniciais com o WST. Em outras palavras, acreditamos que pode haver uma curva de aprendizagem mais longa no SE do que no WST.

Agradecimentos

Agradecemos ao Programa de Pós-Graduação em Estudos Linguísticos (PPGEL) do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU) e ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Declaração de autoria

A ideia de analisar, de maneira comparativa, o WST e o SE partiu de Fromm. Ele conduziu todo o processo de análise e escrita deste artigo, além de ter subsidiado financeiramente esta pesquisa, na medida em que proporcionou o acesso aos referidos *softwares* da LC. Fromm e Santos elaboraram o resumo e o *abstract*. Fromm, Beilke e Grama produziram a seção introdutória e metodológica. Em relação à fundamentação teórica, Beilke escreveu a parte relativa à LC, Santos apresentou o WST e o SE, Grama discorreu sobre a Lexicografia e Santos sobre a Terminografia. A produção escrita da seção de descrição e análises foi elaborada concomitantemente à análise comparativa dos programas WST e SE por Santos, Grama e Beilke em reuniões que ocorreram no laboratório de informática da UFU. As considerações finais foram formuladas pelos quatro autores deste artigo. Por fim, a revisão deste texto ficou sob a responsabilidade de Fromm e Grama.

Referências

ALMEIDA, G. M. B.; VALE, O. A. Do texto ao termo: interação entre Terminologia, Morfologia e Linguística de Corpus na extração semi-automática de termos. *In*: ISQUERDO, A. N.; FINATTO, M. J. B. (org.). *As ciências do Léxico*: Lexicologia, Lexicografia e Terminologia. Campo Grande: Editora da UFMS, 2008. v. IV, p. 483-499.

ALMEIDA, G. M. B. Fazer Terminologia é fazer Linguística. *In*: PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. (org.). *Linguagens especializadas em corpora*: modos de dizer e interfaces de pesquisa. Porto Alegre: Editora da PUCRS, 2010. v. 1, p. 72-90.

BEILKE, N. S. V. *Pommersche Korpora*: Uma proposta metodológica para compilação de corpora dialetais. 2016. 286 f. Dissertação (Mestrado em Estudos Linguísticos) – Instituto de Letras e Linguística, Universidade Federal de Uberlândia, Uberlândia, 2016. Disponível em: https://repositorio.ufu.br/handle/123456789/18022. Acesso em: 27 set. 2019.

BEILKE, N. S. V. Pommersche korpora: um conjunto de corpora dialetais da variedade brasileira do pomerano. *In*: FINATTO, M. J. B.; REBECHI, R. R.; SARMENTO, S.; BOCORNY, A. E. P. (org.). *Linguística de Corpus*: Perspectivas. Porto Alegre: Instituto de Letras – UFRGS, 2018.

p. 365-398. Disponível em: https://lume.ufrgs.br/handle/10183/177640. Acesso em: 15 nov. 2019.

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

BERBER SARDINHA, T. Linguística forense. *In*: _____. *Pesquisa em Linguística de Corpus com WordSmith Tools*. Campinas: Mercado das Letras, 2009.

BEVILACQUA, C. R.; FINATTO, M. J. B. Lexicografia e Terminologia: Alguns contrapontos fundamentais. *Alfa*, São Paulo, v. 50, n. 2, p. 43-54, 2006. Disponível em: http://seer.fclar.unesp.br/alfa/article/view/1410/1111. Acesso em: 31 jul. 2019.

BIDERMAN, M. T. C. *Teoria linguística*. São Paulo: Martins Fontes, 2001.

BIDERMANN, M. T. C. Análise de dois dicionários gerais do português brasileiro contemporâneo: o Aurélio e o Houaiss. *Filologia e Linguística Portuguesa*, São Paulo, n. 5, p. 85-116, 2003. DOI: https://doi.org/10.11606/issn.2176-9419.v0i5p85-116. Disponível em: http://www.revistas.usp.br/flp/article/view/59701. Acesso em: 31 jul. 2019.

BORBA, F. S. *Organização de Dicionários*: Uma introdução à Lexicografia. São Paulo: UNESP, 2003.

CABRÉ, M. T. La terminología hoy: concepciones, tendências y aplicaciones. *Ciência da Informação*, Brasília, v. 24, n. 3, p. 1-15, 1995. Disponível em: http://revista.ibict.br/ciinf/article/view/567/568. Acesso em: 29 de jul. 2019.

CABRÉ, M. T. *Terminology***:** theory, methods, and applications. Philadelphia, PA: John Benjamins, 1999. DOI: https://doi.org/10.1075/tlrp.1

DANTAS, W. Disponível em: https://www.youtube.com/user/CorpusLael/videos. 2010. Acesso em: 25 jul. 2019.

DAVIES, M. *O Corpus do Português*. 2019. Disponível em: https://www.corpusdoportugues.org/compare_larger.asp. Acesso em: 22 jul. 2019.

FROMM, G. *Proposta para um modelo de glossário de informática para tradutores*. 2002. 82 f. Dissertação (Mestrado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2002. Disponível em: http://www.ileel.ufu.br/guifromm/wp-content/uploads/2014/05/dissertacao.pdf. Acesso em: 2 abr. 2019.

FROMM, G. *VoTec*: a construção de vocabulários eletrônicos para aprendizes de tradução. 2007. 214 f. Tese (Doutorado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007. Disponível em: http://www.teses.usp.br/teses/disponiveis/8/8147/tde-08072008-150855/pt-br.php. Acesso em: 2 abr. 2019.

FROMM, G. Vocabulário de Linguística: treinamento em Terminografia Bilíngue, uso de corpora e Ambiente de Gestão Terminológica. *In*: ISQUERDO, A. N.; DAL CORNO, G. O. M. (org.). *As ciências do léxico*: lexicologia, lexicografia, terminologia. Campo Grande: Ed. UFMS, 2018. v. 7, p. 309-328.

FROMM, G; YAMAMOTO, M. I. Terminologia, Terminografia, Tradução e Linguística de Corpus: a criação de um vocabulário bilíngue sobre Linguística. *In*: TAGNIN, S.; BEVILACQUA, C. (org.). *Corpora na Terminologia*. São Paulo: Hub Editorial, 2013. p. 129-152.

GOMIDE, A. R. Contrastando duas ferramentas para análise de corpus de aprendizes: AntConc e Pacote tm. *In*: CONGRESSO NACIONAL UNIVERSIDADE, EAD E SOFTWARE LIVRE, 2015, Belo Horizonte. *Anais*... Belo Horizonte: Faculdade de Letras da UFMG, 2015. p. 1-5. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/ueadsl/article/view/8659/7604. Acesso em: 10 mar. 2019.

GONÇALVES, L. B. *Dubliners sob a lupa da Linguística de Corpus*: Uma contribuição para a análise e a avaliação da tradução literária. 2006. 328 f. Tese (Doutorado em Estudos Linguísticos e Literários em Inglês) – Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2006. Disponível em: http://www.teses.usp.br/teses/disponiveis/8/8147/tde-08112007-154609/pt-br.php. Acesso em: 31 jul. 2019.

GRAMA, D. F. *Uma análise lexicográfica dos elementos coesivos sequenciais do português para a elaboração de uma proposta de definição*: um estudo com base em corpus. 2016. 371 f. Dissertação (Mestrado em Estudos Linguísticos) – Instituto de Letras e Linguística, Universidade Federal de Uberlândia, Uberlândia, 2016. Disponível em: https://repositorio.ufu.br/handle/123456789/18084. Acesso em: 27 set. 2019.

KILGARRIFF, A.; RYCHLÝ, P. *Sketch Engine*. East Sussex: Lexical Computing Limited, 2003. Disponível em: http://www.sketchengine.eu. Acesso em: 31 jul. 2019.

KRIEGER, M. G.; FINATTO, M. J. B. *Introdução à Terminologia*. São Paulo: Contexto, 2004.

NAVARRO, S. Corpora e variantes culturais: um estudo de caso da hotelaria. *In*: TAGNIN, S.; BEVILACQUA, C. (org.). *Corpora na Terminologia*. São Paulo: HUB, 2013. v. 1, p. 115-130.

PERINI, M. A. *Princípios de linguística descritiva*: introdução ao pensamento gramatical. São Paulo: Parábola Editorial, 2006.

PERINI, M. A. *Estudos de gramática descritiva*: as valências verbais. São Paulo: Parábola Editorial, 2008.

SCOTT, M. *WordSmith Tools*. Versão 7. Stroud: Lexical Analysis Software, 2016.

SEABRA, M. C. T. C. de. Questões teóricas genéricas. *In*: XATARA, C.; BELIVACQUA, C. R.; HUMBLÉ, P. R. M. (org.). *Dicionários na teoria e na prática*: como e para quem são feitos. São Paulo: Parábola Editorial, 2011. p. 29-37.

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

TEIXEIRA, E. D. *A Linguística de Corpus a serviço do tradutor*: proposta de um dicionário de culinária voltado para a produção textual. 2008. 439 f. Tese (Doutorado em Estudos Linguísticos e Literários em Inglês) – Universidade de São Paulo, São Paulo, 2008. Disponível em: http://www.teses.usp.br/teses/disponiveis/8/8147/tde-16022009-141747/pt-br.php. Acesso em: 02 fev. 2016.

TELINE, M. F.; ALMEIDA, G. M. B.; ALUÍSIO, S. M. Extração manual e automática de Terminologia: comparando abordagens e critérios. *In*: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TIL), 1., 2003, São Carlos. *Anais* [...] São Carlos: UFSCAR, 2003. p. 1-12. Disponível em: http://www.nilc.icmc.usp.br/til/til2003/oral/Teline_Almeida_Aluisio_37.pdf. Acesso em: 31 jul. 2019.

WELKER, H. A. Questões de lexicografia pedagógica. *In*: XATARA, C.; BELIVACQUA, C. R.; HUMBLÉ, P. R. M. (org.). *Dicionários na teoria e na prática*: como e para quem são feitos. São Paulo: Parábola Editorial, 2011. p. 103-113.

WILKENS, R.; PEREIRA BOCORNY, A. E.; KRAUSE KILIAN, C.; VILLAVICENCIO, A. Ambientes web de gestão terminológica para a criação de produtos terminológicos on-line. *Debate Terminológico*, Porto Alegre, n. 8, p. 16-22, 2012. Disponível em: https://seer.ufrgs.br/riterm/article/view/29877/18474. Acesso em: 31 jul. 2019.

YAMAMOTO, M. I. Vocabulário bilíngue Português/Inglês de Linguística Geral. *Revista Philologus*, Rio de Janeiro, ano 24, n. 70, p. 272-297, jan./abr. 2018. Disponível em: http://www.filologia.org.br/x_sinefil/completos/vocabulario_bilingue_MARCIO.pdf. Acesso em: 31 jul. 2019.