LINGUÍSTICA DE CORPUS: POSSIBILIDADES E AVANÇOS

Cárla Callegaro Corrêa Kader*
Marcos Gustavo Richter**

Resumo

Este trabalho tem por objetivo apresentar os aspectos histórico-teóricos da Linguística de *Corpus*, enquanto metodologia de pesquisa, além dos programas WordSmith Tools 5.0 e AntConc 3.2.1, considerados ferramentas de análise para os estudos linguísticos. Este estudo é uma revisão bibliográfica que mostra como os programas podem ser usados para investigar questões de áreas centrais da Linguística e da Linguística Aplicada. A consideração final que se pode fazer, a partir deste recorte teórico, é que os programas colocam, à disposição do pesquisador, ferramentas que podem ser úteis nas etapas fundamentais de um projeto de pesquisa, especialmente nas fases de levantamento dos dados, triagem e análise de casos relevantes.

Palavras-chave: Linguística de *Corpus*. WordSmith Tools 5.0. AntConc 3.2.1.

Introdução

A Linguística de *Corpus* (LC) é um campo que se dedica à criação e análise de *corpora* (plural latim de *corpus*), ou seja, conjunto de textos e transcrições de fala armazenadas em arquivos de computador. A Linguística de *Corpus* vem mudando a maneira como se investiga a linguagem, nos seus mais diversos níveis, colocando à disposição do analista quantidades de dados antes inacessíveis. Um dos grandes agentes dessa mudança é a informática; sem ela, a Linguística de *Corpus* contemporânea não poderia existir (BERBER SARDINHA, 2009). Assim, o linguista de *corpus* depende de programas de computador para lidar com *corpora*. Dentre os *softwares* que existem para auxiliar o linguista de *corpus*, um deles se destaca: WordSmith Tools. Esse programa foi criado em 1996 por Mike Scott, da Universidade de Liverpool, Reino Unido. Hoje, o programa possui um grande número de usuários no mundo todo e uma versão mais atualizada, a saber: WordSmith Tools 5.0.

No Brasil, os cursos, oficinas e palestras sobre seu funcionamento se multiplicam, chegando a ser considerado como um facilitador e fator de divulgação da Linguística de *Corpus* neste país (idem, 2009).

Já o Programa AntConc 3.2.1, criado por Laurence Anthony da Universidade de Waseda (Japão), é um concordanciador utilizado para listar as ocorrências de uma determinada palavra ou frase em uma quantidade definida de contextos. De forma geral, os concordanciadores também executam outras funções, como listar palavras em um texto ou *corpus,* extrair palavras-chave e colocados. O AntConc é um *software* livre para os sistemas

* Professora do Instituto Federal Farroupilha, Campus de São Vicente do Sul (IFFARROUPILHA), Mestre em estudos Linguísticos pela Universidade Federal de Santa Maria (UFSM), doutoranda em Estudos Linguísticos pela Universidade Federal de Santa Maria (UFSM). carlackader@gmail.com

** Professor Doutor em Estudos Linguísticos, Professor do Curso de Pós-Graduação em Letras da Universidade Federal de Santa Maria (UFSM). richtermg@gmail.com

Windows, Mac OS X e Linux. Tal como o WordSmith Tools, ele tem ferramentas para analisar *word clusters, n-grams*, colocados, frequência de palavras e palavras-chave.

Com base na funcionalidade dos dois programas, trazemos a visão de Scott (1998, p. 12) quanto às ferramentas computacionais, afirmando que "elas são úteis porque permitem que certas ações sejam realizadas facilmente, e esta facilidade significa realizar trabalhos mais complexos". Essas ferramentas possibilitam a remodelagem de um conjunto de dados em uma nova forma a fim de se identificar padrões.

No âmbito deste estudo, abordaremos as ferramentas computacionais, não no intuito de compará-las, mas de apresentá-las como mais uma opção na área da pesquisa em Linguística de *Corpus*, além de, inicialmente, traçarmos alguns aspectos histórico-teóricos dessa metodologia de pesquisa.

1. Aspectos Histórico-Teóricos da Linguística de *Corpus*

A Linguística de *Corpus* envolve a Linguística Computacional (manuseio virtual de dados) e a Estatística, uma vez que trabalham com uma amostra da língua, constituindo, em certa medida, um novo método de pesquisa.

Há autores que consideram a Linguística de *Corpus* uma abordagem, como Berber Sardinha (2000), e outros como uma metodologia, como Rocha (2000). Leech (1992) refere-se a ela como uma nova abordagem filosófica e não apenas como uma nova metodologia.

Segundo Vasilévski (2007), a Linguística de *Corpus* compreende uma base metodológica incontestável e que ao ser tratada por metodologia, tem a possibilidade de ampliar o seu campo de atuação.

Por se caracterizar pela utilização de um *corpus* linguístico, a Linguística de *Corpus* pode ser entendida como Linguística de Massa (VASILÉVSKI, 2007), expressão, segundo autora, demarcadora da dimensão de seus componentes. Mas o que seria um *corpus*? Segundo Sanchez (1995, p. 8-9), a expressão *corpus* refere-se a um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise.

Portanto, a partir da observação dos dados do *corpus*, desenvolve-se uma interpretação do fenômeno da língua estudado. Essa observação é feita com o auxílio de programas de computador, que manuseiam dados textuais, e de recursos matemáticos que permitem interpretar dados e extrair significados para serem aplicados a um todo, a partir da análise de uma parte, ou seja, de uma amostra (VASILÉVSKI, 2007).

A Linguística de *Corpus* é uma metodologia relativamente nova para os estudos da língua e pode ser aplicada em muitas áreas linguísticas, tais como: sintaxe, semântica, fonética e sociolinguística, dentre outras.

Os estudos linguísticos envolvendo *corpus* datam de 1897, embora alguns linguistas, inclusive estruturalistas, tenham utilizado um método com base em *corpus* em pesquisas nos anos 1940 (McENERY; WILSON, 1997).

Já a expressão Linguística de *Corpus* é nova. Segundo Leech (1992), a expressão Linguística de *Corpus* apareceu em 1984 como título de um livro de Aarts e Meijs. Anterior a esta data, trabalhar com *corpus* relativamente significativo em volume era inviável em função das dificuldades de manipulação (tempo, equipe de trabalho e dinheiro).

Esse problema não durou muito, pois o advento do computador veio a resolver o gerenciamento e

estocagem de dados, viabilizando a Linguística de *Corpus* e tornando-a ferramenta aplicável a pesquisa em língua.

De acordo com Vasilévski (2007), além de acelerar o manuseio de grandes quantidades de dados (busca, armazenamento, classificação e disposição, dentre outros), o computador pôs a matemática a favor da linguística, de forma que cálculos simples e complexos puderam ser realizados rapidamente com grandes quantidades de dados, sem margem para erros. Em consequência, a estatística passou a ser aplicada à Linguística de *Corpus* e concedeu a essa metodologia maior teor científico.

Leech (1992) argumenta que o volume de dados não é o principal aspecto referente à Linguística de *Corpus*, mas o estudo da língua por meio de textos autênticos, com ênfase no desempenho linguístico (sob a forma de discurso natural escrito ou falado).

Leech (1992) destaca alguns aspectos que devem ser observados quando se trabalha com a Linguística de *Corpus*:

a) por maior que seja um *corpus*, ele é apenas uma amostra da língua em uso;
b) os dados a analisar não devem ser escolhidos de acordo com as preferências do pesquisador, e sim aleatoriamente, e nenhum deles pode ser considerado irrelevante para a pesquisa;
c) teorias ou modelos podem ser criados para explicar os dados encontrados (a partir da intuição ou experiência do investigador, por exemplo), mas os valores quantitativos do modelo devem ser obtidos dos dados do *corpus*;
d) a precisão do modelo pode ser testada em outro *corpus*;
e) a princípio, a qualidade de um modelo pode ser medida e comparada com a de outros modelos (essa interação é importante para que os modelos de desempenho linguístico sejam progressivamente aperfeiçoados) e diferentes modelos podem ser testados com o mesmo *corpus*, de forma que a superioridade de um modelo em relação a outro possa ser demonstrada.

Segundo Berber Sardinha (2000), a Linguística de *Corpus* ocupa-se da coleta e exploração de *corpora*, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística e como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador.

Berber Sardinha (2000) destaca que um dos primeiros trabalhos relacionados à Linguística de *Corpus* foi a identificação das palavras mais frequentes da língua inglesa, feita por Thorndike há mais de 75 anos atrás (THORNDIKE, 1921). O levantamento foi feito manualmente em um *corpus* de aproximadamente 4,5 milhões de palavras e, quando publicado, impulsionou mudanças no ensino de língua materna e estrangeira, tanto nos Estados Unidos quanto na Europa. As abordagens baseadas no controle do vocabulário, nas quais os alunos têm contato em primeiro lugar com as palavras mais frequentes, estão relacionadas aos estudos como o de Thorndike. Quase 25 anos mais tarde, Thorndike revisou seu levantamento inicial e, tomando como base um *corpus* maior (18 milhões de palavras), publicou uma obra listando as 30 mil palavras mais comuns da língua inglesa. Em 1953, veio o *General Service List of English Words* de Michael West (West, 1953) com uma descrição do léxico inglês. A pesquisa de West dá detalhes do que seriam as 2 mil palavras mais frequentes do inglês e baseou-se no trabalho de Thorndike e Lorge.

De acordo com Berber Sardinha (2000), foi um *corpus* não computadorizado que deu origem aos *corpora* atuais, o SEU (*Survey of English Usage*), compilado por Randolf Quirk e sua equipe, em Londres, a partir de 1953.

O SEU foi planejado para ter o tamanho de 1 milhão de palavras, a composição do *corpus* também foi influente ao definir um número fixo de textos (200) e uma quantidade de palavras igual para cada texto

(5000). Segundo o autor, o *Survey* foi organizado em fichas de papel, cada uma contendo uma palavra do *corpus* inserida em 17 linhas de texto. As palavras foram analisadas gramaticalmente, com cada ficha pertencendo a uma categoria gramatical. O conjunto de categorias resultante serviu de base para o desenvolvimento dos etiquetadores computadorizados atuais, que fazem a identificação de traços gramaticais automaticamente. A transformação completa do *Survey* em *corpus* eletrônico foi atingida anos depois (1989), mas a sua parte falada foi computadorizada antes e ficou conhecida como o *London-Lund Corpus* (BERBER SARDINHA, 2000).

Segundo Granger (1998), o que prepondera na literatura atual é a descrição de linguagem, bem como o interesse no emprego de *corpora* na sala de aula e na investigação da linguagem de alunos de língua.

Atualmente, a Linguística de *Corpus* é de grande influência na pesquisa linguística, por exemplo, na Grã-Bretanha, um dos centros mais desenvolvidos, várias universidades dedicam-se à pesquisa baseada em *corpus* para a descrição de diferentes aspectos relacionados à linguagem.

Em instituições britânicas há interesse voltado à teorização, à criação de *corpora* e à materiais de apoio em diversas áreas. Em países escandinavos (Noruega, Suécia e Dinamarca), existem centros dedicados à Linguística de *Corpus* há vários anos.

Segundo Berber Sardinha (2000), fora da Europa, a Linguística de *Corpus* não está tão desenvolvida, mas já possui centros nos quais a pesquisa está instalada. Nos Estados Unidos, embora haja facilidade de obtenção de recursos de informática, a Linguística de *Corpus* tem uma presença modesta, em função da influência da linguística gerativa-transformacional. Em contrapartida, os Estados Unidos destaca-se na pesquisa em Processamento de Linguagem Natural (PLN), tanto em nível acadêmico quanto industrial (as empresas de informática investem pesado na pesquisa linguística com fins comerciais).

2. *Corpora* de Aprendizes

De acordo com McEnery et al (2006), as *corpora* de aprendizes compreendem a escrita e o discurso de aprendizes adquirindo uma segunda língua (L2). Os dados coletados são as produções desses aprendizes que são posteriormente analisadas.

O *Corpus* Internacional de Aprendizes de Inglês (ICLE) é o *corpus* de aprendiz mais conhecido atualmente. Este *corpus* contém aproximadamente três milhões de ensaios escritos por aprendizes de inglês de nível avançado. Estes aprendizes são estudantes universitários de língua estrangeira no terceiro ou quarto ano de curso e são procedentes de quatorze línguas maternas diferentes (francês, alemão, holandês, espanhol, sueco, finlandês, russo, italiano, hebreu, japonês, chinês, entre outros).

O projeto ICLE é sediado pela Universidade Católica de Louvain, Bélgica, e dirigido pela professora Drª. Sylviane Granger, especialista em *Corpora* de Inglês de Aprendizes (*Learner Corpora*). O projeto, de âmbito internacional, conta com um grupo de pesquisadores da Bélgica e com outros pesquisadores de outros países. Granger é responsável pela formação, organização e informatização de um *sub-corpus* local.

No Brasil, o grupo é coordenado pelo professor Dr. Tony Berber Sardinha da Pontifícia Universidade Católica de São Paulo. O objetivo da equipe no Brasil é fornecer (coletar, organizar e informatizar) o *sub-corpus* referente à produção em língua inglesa de alunos brasileiros (falantes de português). Em outras palavras, este grupo está oficialmente responsável pela montagem de *sub-corpora* no Brasil.

Na visão de McEnery et al (2006), para permitir o contraste entre a escrita de aprendizes de diferentes línguas maternas e o inglês, o ICLE pode ser usado em combinação com o *Louvain Corpus of native English Essays* (LOCNESS). Este *corpus* está disponível para

pesquisas na área da Linguística, mas não pode ser usado para fins comerciais.

Outro *corpus* relacionado aos *corpora* de aprendiz é o *Longman Learners' Corpus* que contém dez milhões de textos escritos por estudantes de Inglês. Estes estudantes pertencem a níveis diferentes de proficiência linguística e têm línguas maternas diferentes (vinte línguas maternas diferentes).

Para obtenção dos textos, foram delegadas tarefas diferentes com ou sem o uso do dicionário, cada ensaio dos alunos foi classificado de acordo com o conhecimento linguístico de primeira língua e o nível de proficiência da segunda.

Tal *corpus*, segundo McEnery et al (2006), oferece informação valiosa sobre os erros cometidos pelos aprendizes, bem como o conhecimento de língua estrangeira que já possuem e está publicamente disponível para fins comerciais.

O *Cambridge Learner Corpus* (CLC) apresenta uma ampla amostra de textos escritos por aprendizes de inglês de diferentes lugares do mundo. Ele contém vinte milhões de palavras e está em constante expansão. Os textos apresentados são procedentes de estudantes de diferentes lugares do mundo que prestam o exame de inglês *Cambridge ESOL*. O *corpus* contém 50,000 textos de aprendizes de 150 países com 100 línguas maternas diferentes (MCENERY et al, 2006). Atualmente, somente os autores, consultores e escritores que trabalham para a editora Cambridge podem ter acesso a este *corpus*.

Há ainda corpora de aprendizes que compreendem estudantes procedentes de uma língua materna apenas, tais como: HKUST (com dez milhões de palavras compostas por ensaios e provas prestadas por estudantes chineses), CLEC (com um milhão de palavras procedentes de textos produzidos por chineses com grau de escolaridade diferente), entre outros.

Na Universidade de São Paulo, foi criado um *corpus* de aprendiz chamado CoMAprend (TAGNIN, 2006), contendo as redações dos aprendizes de língua estrangeira das cinco áreas do Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas (alemão, inglês, espanhol, italiano e francês).

Tribble (1990, 1997) foi um dos pioneiros no uso de pequenos *corpora* de aprendizes. Tribble investigou o uso de verbos relacionados à fala em um *corpus* de aprendizes de cerca de 54 mil palavras, a maioria extraída do *Longman Corpus of Learner English*. O autor sugere que os exemplos obtidos a partir das linhas de concordância derivadas do *corpus* de aprendizes podem ser explorados por eles mesmos em sala de aula, de maneira que sejam utilizados para auxiliá-los na reestruturação de sentenças, incentivando-os a empregar uma variedade maior de vocabulário.

No entanto, os *corpora* de aprendizes não são, por enquanto, disponibilizados para professores de inglês, pois não são comercializados. Se fossem disponibilizados, eles seriam restritos a um contexto de ensino específico. Mas, os professores de língua estrangeira podem compilar seus próprios *corpora* de pequena extensão e refletir melhor sobre seu contexto de ensino.

3. Pesquisa em Linguística de Corpus com WordSmith Tools 5.0 e AntConc 3.2.1

3.1 O programa WordSmith Tools 5.0

O programa *WordSmith Tools* é um conjunto de programas integrados destinado à análise linguística, desenvolvido por Mike Scott e comercializado pela *Oxford University Press*, estando em sua quinta versão atualmente.

O programa é empregado em pesquisas lexicográficas realizadas pela editora que o comercializa e em análises diversas desenvolvidas por linguistas, professores e estudantes de línguas em escala mundial (SCOTT, 1998).

Especificamente, esse *software* permite fazer análises baseadas na frequência e na co-ocorrência de palavras em *corpora*. Além disso, ele permite pré-processar os arquivos do *corpus* (retirar partes indesejadas de cada texto, organizar o conjunto de arquivos, inserir e remover etiquetas etc), antes da análise propriamente dita. A intenção do programa é servir como ferramenta que permita a consecução de tarefas relacionadas às análises de *corpora*.

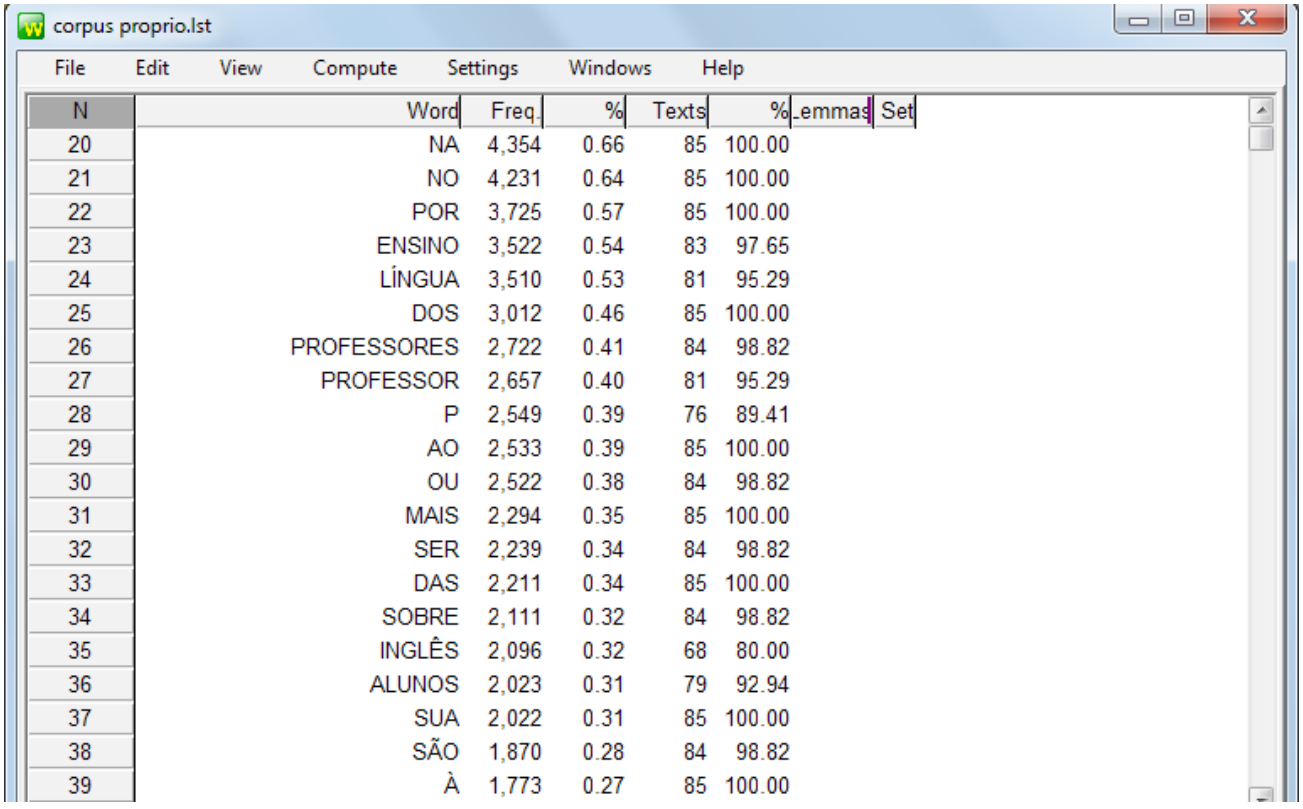
Para investigar textos com o auxílio do programa, é preciso que estes estejam em formato digital, preferencialmente TXT. Nos arquivos TXT, os usuários podem optar por utilizar cabeçalhos ou outros tipos de marcação desde que os mesmos ocorram entre os sinais de menor e maior (< e >). Por definição padrão, o programa ignora automaticamente tudo que estiver entre estes sinais. Outra possibilidade é de marcar os textos com etiquetas específicas (como em páginas HTML) que indiquem o início e o fim da mesma.

O WordSmith Tools não foi feito para efetuar análises de projetos específicos; ele disponibiliza uma série de opções de ferramentas, algumas mais gerais, outras mais restritas (BERBER SARDINHA, 2009). Esse programa permite o pré-processamento, a organização de dados e a análise propriamente dita de *corpora* ou textos isolados. Ele também oferece ferramentas para consecução de tarefas essenciais, como listas de palavras (através do programa *Wordlist*) e de concordâncias (por meio do *Concord*).

O programa possui três ferramentas e quatro utilitários.

As ferramentas são (BERBER SARDINHA, 2009):

- *WordList*: produz listas de palavras contendo todas as palavras do arquivo ou arquivos selecionados, elencadas em conjunto com suas frequências absolutas e percentuais. O programa é pré-definido para produzir, a cada vez, duas listas de palavras, uma ordenada alfabeticamente (identificada pela letra A entre parênteses) e outra classificada por ordem de frequência das palavras (com a palavra mais frequente encabeçando a lista). Cada uma dessas listas é apresentada em uma janela diferente, e, juntamente com as duas janelas correspondentes à lista alfabética (A) e à lista de frequência (F), o programa oferece uma terceira janela (S) na qual aparecem estatísticas relativas aos dados usados para produção das listas. Assim, para cada vez que o *WordList* é chamado para fazer uma lista de palavras, três janelas são produzidas: uma contendo uma lista de palavras ordenada por ordem alfabética, outra com uma lista classificada pela frequência das palavras, e uma terceira janela com estatísticas simples a respeito dos dados. Além disso, compara listas criando listas de consistência, onde é informado em quantas listas cada palavra aparece. As listas de palavras do *WordSmith Tools* podem ser de dois tipos: com palavras individuais ou com agrupamentos de palavras (*clusters*). Há três procedimentos básicos disponíveis no *WordList*, a saber: criar uma lista apenas, para um ou mais arquivos selecionados; criar várias listas, uma para cada arquivo e criar um arquivo de índice. Cada uma dessas opções tem uma finalidade. A lista simples é o procedimento básico, para quando o analista precisa criar apenas uma lista. Ela pode conter palavras individuais ou agrupamentos *(clusters*). As listas produzidas em lote possuem o mesmo formato da lista única, mas são produzidas em conjunto, uma para cada arquivo. A finalidade é tornar mais rápida a produção das listas. Ela pode conter palavras individuais ou agrupamentos (*clusters*). O arquivo de índice é um tipo especial de arquivo, diferente das listas tradicionais, que engloba, além da frequência de cada palavra, uma estatística de associação para pares de palavras, uma linha de concordância e a possibilidade de fazer concordâncias completas de modo mais rápido. Conforme figura abaixo.


corpus proprio.lst

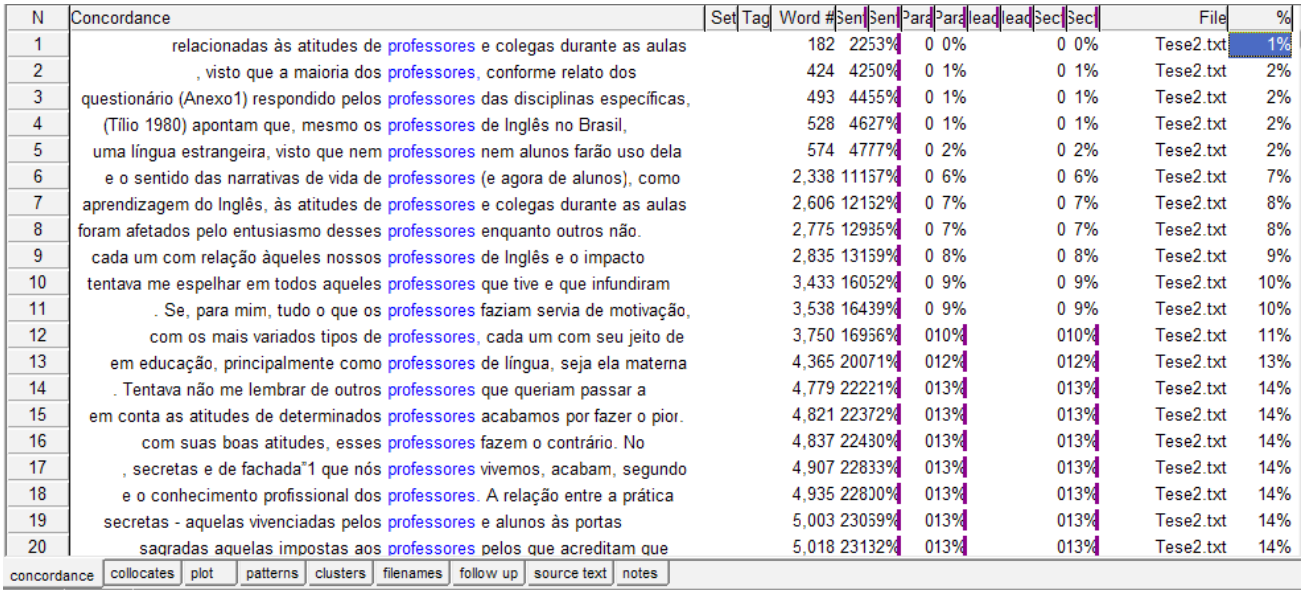
File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%_emmas	Set
20	NA	4,354	0.66	85	100.00	
21	NO	4,231	0.64	85	100.00	
22	POR	3,725	0.57	85	100.00	
23	ENSINO	3,522	0.54	83	97.65	
24	LÍNGUA	3,510	0.53	81	95.29	
25	DOS	3,012	0.46	85	100.00	
26	PROFESSORES	2,722	0.41	84	98.82	
27	PROFESSOR	2,657	0.40	81	95.29	
28	P	2,549	0.39	76	89.41	
29	AO	2,533	0.39	85	100.00	
30	OU	2,522	0.38	84	98.82	
31	MAIS	2,294	0.35	85	100.00	
32	SER	2,239	0.34	84	98.82	
33	DAS	2,211	0.34	85	100.00	
34	SOBRE	2,111	0.32	84	98.82	
35	INGLÊS	2,096	0.32	68	80.00	
36	ALUNOS	2,023	0.31	79	92.94	
37	SUA	2,022	0.31	85	100.00	
38	SÃO	1,870	0.28	84	98.82	
39	À	1,773	0.27	85	100.00	

Figura 1 - *WordList* do *corpus* próprio, composto por artigos acadêmicos de profissionais da área de Letras.

•*Concord*: realiza concordâncias, ou listagens de uma palavra específica (o 'nódulo'), juntamente com parte do texto onde ocorreu. Oferece também listas de colocados, isto é, palavras que ocorreram perto do nódulo. O sucesso da busca no *Concord* depende da especificação correta do termo de busca. O *Concord* é acionado de duas maneiras: clicando em *Tools/Concord* no *Controller* ou clicando em uma palavra da lista de palavras (produzida pelo *WordList*), ou em uma palavra de uma lista de palavras-chave (produzida pelo *KeyWords*), ou ainda em uma palavra de um arquivo índice (*index file*). Há vários tipos de concordância possíveis, de acordo com a posição do item de busca na listagem. A mais comum é a KWIC, sigla de *Key Word* in *Context*, ou palavra-chave no contexto, na qual a palavra de busca aparece centralizada e ladeada por porções contínuas do texto de origem. As concordâncias são instrumentos reconhecidamente indispensáveis no estudo da colocação e da padronização lexical e, por isso, fundamental na investigação de *corpora*.

No *WordSmith Tools*, o *Concord* pode ser usado separadamente, para concordâncias avulsas, ou em conjunto com as ferramentas *WordList* e *KeyWords*, chamados a partir desses programas. Para tanto, basta selecionar um item de uma lista de palavras ou palavras-chave e clicar no botão C na barra de tarefas do *WordList* ou *KeyWords*; o *Concord* é chamado e uma concordância do item selecionado é produzida, a partir dos textos selecionados quando da produção da lista de palavras ou palavras-chave. Essa chamada automática não funciona caso a lista tenha sido salva e os textos de onde foi feita não estiverem mais nas pastas originais. Nesse caso, a concordância é retornada vazia. Conforme figura abaixo.



N	Concordance	Set	Tag	Word #	Sent	Sent	Para	Para	lead	lead	Sec	Sec	File	%
1	relacionadas às atitudes de professores e colegas durante as aulas			182	2	253%	0	0%		0	0%		Tese2.txt	1%
2	, visto que a maioria dos professores, conforme relato dos			424	4	250%	0	1%		0	1%		Tese2.txt	2%
3	questionário (Anexo1) respondido pelos professores das disciplinas específicas,			493	4	455%	0	1%		0	1%		Tese2.txt	2%
4	(Tílio 1980) apontam que, mesmo os professores de Inglês no Brasil,			528	4	627%	0	1%		0	1%		Tese2.txt	2%
5	uma língua estrangeira, visto que nem professores nem alunos farão uso dela			574	4	777%	0	2%		0	2%		Tese2.txt	2%
6	e o sentido das narrativas de vida de professores (e agora de alunos), como			2,338	11	157%	0	6%		0	6%		Tese2.txt	7%
7	aprendizagem do Inglês, às atitudes de professores e colegas durante as aulas			2,606	12	152%	0	7%		0	7%		Tese2.txt	8%
8	foram afetados pelo entusiasmo desses professores enquanto outros não.			2,775	12	935%	0	7%		0	7%		Tese2.txt	8%
9	cada um com relação àqueles nossos professores de Inglês e o impacto			2,835	13	159%	0	8%		0	8%		Tese2.txt	9%
10	tentava me espelhar em todos aqueles professores que tive e que infundiram			3,433	16	052%	0	9%		0	9%		Tese2.txt	10%
11	. Se, para mim, tudo o que os professores faziam servia de motivação,			3,538	16	439%	0	9%		0	9%		Tese2.txt	10%
12	com os mais variados tipos de professores, cada um com seu jeito de			3,750	16	956%	0	10%		0	10%		Tese2.txt	11%
13	em educação, principalmente como professores de língua, seja ela materna			4,365	20	071%	0	12%		0	12%		Tese2.txt	13%
14	. Tentava não me lembrar de outros professores que queriam passar a			4,779	22	221%	0	13%		0	13%		Tese2.txt	14%
15	em conta as atitudes de determinados professores acabamos por fazer o pior.			4,821	22	372%	0	13%		0	13%		Tese2.txt	14%
16	com suas boas atitudes, esses professores fazem o contrário. No			4,837	22	430%	0	13%		0	13%		Tese2.txt	14%
17	, secretas e de fachada"1 que nós professores vivemos, acabam, segundo			4,907	22	833%	0	13%		0	13%		Tese2.txt	14%
18	e o conhecimento profissional dos professores. A relação entre a prática			4,935	22	800%	0	13%		0	13%		Tese2.txt	14%
19	secretas - aquelas vivenciadas pelos professores e alunos às portas			5,003	23	059%	0	13%		0	13%		Tese2.txt	14%
20	sagradas aquelas impostas aos professores pelos que acreditam que			5,018	23	132%	0	13%		0	13%		Tese2.txt	14%

concordance | collocates | plot | patterns | clusters | filenames | follow up | source text | notes

Figura 2 - Exemplo de linhas de concordância de *corpus* próprio, composto por artigos acadêmicos de profissionais da área de Letras.

•*KeyWords:* extrai palavras de uma lista cujas frequências são estatisticamente diferentes (maiores ou menores) do que as frequências das mesmas palavras num outro *corpus* (de referência). Calcula também palavras-chave, que são chave em vários textos. Palavras-chave não são o mesmo que palavras "importantes". O programa não identifica obrigatoriamente palavras-chave encontradas em artigos científicos no campo "palavras-chave", por exemplo. As palavras-chave podem ser de dois tipos: positivas e negativas. Dizemos palavras-chave positivas quando sua frequência é relativamente mais alta no *corpus* do estudo do que no *corpus* de referência. E negativas, quando sua frequência é relativamente mais alta no *corpus* de referência do que no de estudo. As palavras-chave positivas e negativas são exibidas separadamente na tela de resultados. As positivas aparecem no começo da lista, em cor amarela. As negativas surgem no final da lista, em cor diferente. As palavras-chave são úteis na análise linguística para fins diversos, tais como:

•identificar a temática de um *corpus* ou de um texto;
•descrever a organização interna de textos;
•localizar marcas indicativas de posicionamento ideológico;
•traçar um perfil lexical de um autor ou de outros indivíduos.

O programa *KeyWords* também possibilita contar a quantidade de vezes que algumas palavras foram chave em várias listas. Palavras que foram chave em um número determinado de listas são chamadas palavras-chave.

Os passos seguidos pelo programa *KeyWords* para obtenção de palavras-chave-chave são, resumidamente, os seguintes (BERBER SARDINHA, 2004):

a)Abrir um conjunto de listas de palavras-chave;
b)Identificar as palavras-chave em comum entre a primeira lista e a segunda;
c)Se as palavras em comum tiverem uma frequência igual ou superior ao valor mínimo estabelecido pelo usuário, pode-se copiá-las para uma lista de palavras-chave-chave;
d)Pode-se repetir os passos b e c para comparar cada lista com todas as outras.

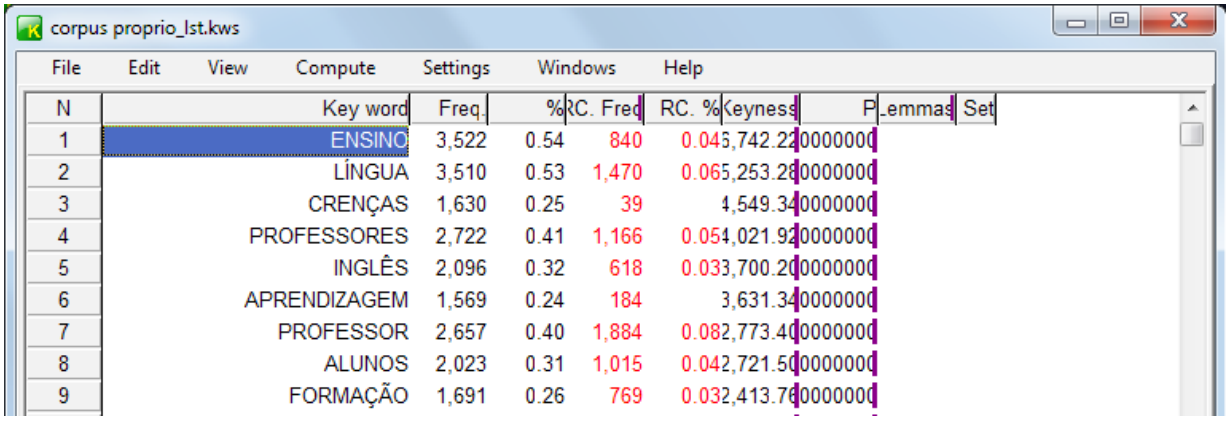
Já os utilitários do programa são:

File Manager: abre uma janela para o gerenciamento de arquivos.
Splitter: permite dividir um arquivo em vários arquivos menores.
Text Converter: oferece várias funções para o pré-

processamento de textos, como a substituição de palavras, partes de palavras ou partes de textos, simultaneamente num conjunto de arquivos, a renomeação em massa de arquivos e a mudança de pasta (diretório) de arquivos que apresentem certas características.

Viewer & Aligner: fornece meios para a visualização de textos e para o alinhamento (combinação) de dois textos num só.

Abaixo, trazemos um exemplo de *KeyWords* de um *corpus* próprio, composto por artigos acadêmicos de profissionais da área de Letras.



N	Key word	Freq.	%	RC. Freq	RC. %	Keyness	P	Lemmas	Set
1	ENSINO	3,522	0.54	840	0.04	5,742.22	0000000		
2	LÍNGUA	3,510	0.53	1,470	0.06	5,253.28	0000000		
3	CRENÇAS	1,630	0.25	39		4,549.34	0000000		
4	PROFESSORES	2,722	0.41	1,166	0.05	4,021.92	0000000		
5	INGLÊS	2,096	0.32	618	0.03	3,700.20	0000000		
6	APRENDIZAGEM	1,569	0.24	184		3,631.34	0000000		
7	PROFESSOR	2,657	0.40	1,884	0.08	2,773.40	0000000		
8	ALUNOS	2,023	0.31	1,015	0.04	2,721.50	0000000		
9	FORMAÇÃO	1,691	0.26	769	0.03	2,413.76	0000000		

Figura 3 - Quadro das palavras-chave.

O WordSmith Tools 5.0 pode ser utilizado em cinco áreas de pesquisa diferentes: ensino de língua estrangeira, análise de gênero, metáfora, tradução e linguística forense.

Segundo Berber Sardinha (2009), essas cinco áreas reúnem focos de pesquisa onde o WordSmith Tools pode atuar como instrumento central da análise de *corpus* e foram assim determinadas por serem campos onde a pesquisa de *corpus*, aliada ao WordSmith Tools, pode trazer muitos benefícios.

3.2 O programa AntConc 3.2.1

O programa AntConc 3.2.1 é um conjunto de ferramentas que permite buscas e faz o cálculo estatístico das ocorrências das palavras em um *corpus* escrito, desenvolvido por Laurence Anthony, da *Faculty of Science and Engineering* - Waseda *University*.

Este programa está disponível no site do LabLEX (http://cel08.fclar.unesp.br.), no site http://www.antlab.sci.waseda.ac.jp/antconc_index.html e no site de Laurence Anthony gratuitamente e permite fazer vários tipos de pesquisa em um determinado corpus.

O AntConc é um programa que analisa automaticamente textos, caracterizando-se como uma ferramenta que facilita a coleta e a análise de dados. Não é necessário fazer nenhuma instalação, pois o programa consiste em apenas um arquivo (de 3,67MB) que é executado com um duplo clique de *mouse* e que possibilita a extração de informação textual (lista de frequência de palavras, concordâncias).

Este programa funciona basicamente nos moldes do WordSmith Tools, ou seja, permite a extração de palavras (*WordList*), listas de concordâncias (*concordance*) e palavras-chave (*KeyWords*).

Os programas WordSmith Tools e AntConc apresentam uma diferença significativa, a saber: o WordSmith Tools, em sua versão gratuita/demo, permite

que sejam extraídas listas de palavras com um número restrito de itens. No caso do AntConc, o número de itens é ilimitado.

As instruções sobre como instalar e utilizar este programa, produzidas por Adauri Berzolin, podem ser encontradas no *site* do Projeto COMET – *Corpus* Multilíngue para Ensino e Tradução (HTTP://www.fflch.usp.br/dem/comet/).

Laurence Anthony (2012) afirma que o programa WordSmith Tools é um programa comercial que funciona somente no sistema Windows. Já o AntConc é um *software* livre que roda em Windows/Mac OS X/Linux. O autor destaca que a funcionalidade dos dois programas é muito similar, mas que o WordSmith Tools tende a ser mais rápido e pode processar grande quantidade de *corpora*, após a sua indexação.

Considerações Finais

Este trabalho buscou fazer uma breve ilustração do desenvolvimento da Linguística de *Corpus*, mostrando como o WordSmith Tools 5.0 e o AntConc 3.2.1 podem ser usados para investigar questões de áreas centrais da Linguística e da Linguística Aplicada.

Estes programas colocam à disposição do pesquisador ferramentas que podem ser úteis nas etapas fundamentais de um projeto de pesquisa, especialmente nas fases de levantamento dos dados, triagem e análise de casos relevantes.

Embora os dois programas apresentem ferramentas similares, um deles é gratuito (AntConc 3.2.1) e, segundo seu criador, não tão rápido com grandes quantidades de dados e o outro, WordSmith Toolls 5.0, se for usado em sua versão demo, apresenta limitações nas linhas de concordância.

Corpus Linguistics: possibilities and advances

Abstract

This study aims to show the historic and theoretical aspects of the Corpus Linguistics, as a research methodology, besides the WordSmith Tools 5.0, AntConc 3.2.1 programs considered as possibilities of analysis tools for the linguistics studies. It is characterized as a bibliographic research that shows as some software can be used to investigate central questions of Linguistics and Applied Linguistics. The final results from this theoretical cut show that the computer programs offer the researcher tools that can be useful in the fundamental steps of the research project, especially in the phases of getting data, selection and analysis of the relevant cases.

Keywords: Corpus Linguistics. WordSmith Tools 5.0. AntConc 3.2.1.

Referências

ANTHONY, L. Laurence Anthony's Website. Disponível em: <www.antlab.sci.waseda.ac.jp> . Acesso em: 15 de fev. 2012.

GERBER, M. e VASILÉVSKI, V. (Org.). *Um percurso para pesquisas com base em corpus***.** Florianópolis: Ed. da UFSC, 2007.

GRANGER, S. *Learner English on the computer*. London: Longman, 1998.

LEECH, G. *Corpora* and theories of linguistics performance. In: SVARTVIK, J. (Org.) *Directions in corpus linguistics.* Berlim: Mouton de Gruyter, 1992.

McENERY, T.; WILSON, A. *A corpus linguistics***.** Edinburg: Edinburg University Press, 1997.

McENERY, T.; XIAO, R.; TONO, Y. *Corpus-based language studies*. New York: Routledge, 2006.

SANCHEZ, A. Definicion e historia de los corpus. In: SANCHEZ, A et al (Org.) CUMBRE – *corpus linguistico de espanol contemporaneo*. Madrid: SGEL, 1995.

SARDINHA, A. P. B. *Linguística de corpus:* histórico e problemática. Revista D.E.L.T.A., São Paulo, v. 16, n. 2 , p. 323-367, 2000.

_____. *Pesquisa em Linguística de corpus com wordsmith tools***.** São Paulo: Mercado de Letras, 2009.

TAGNIN, S. *CoMAprend:* a experiência de construção de um corpus de aprendizes para estudos. Disponível em: WWW. dominiosdelinguagem.org.br/pdf/09-07-09/Texto%206. pdf. Acesso em: 25 de agosto de 2009.

THORNDIKE, E. *The teacher's book*. New York: Teachers College, 1921.

TRIBLE, C. (1997). Improvising corpora for ELT: quick and dirty ways of developing corpora for language teaching. In: MELIA, J.; LEWANDOWSKA- TOMASZEYK, B. (Org.). *PALC '97 Proceeding.* Lodz, Polônia: Lodz University Press. Disponível em: http://web.archive.org/web/20040203111227/http://web.bham.ac.uk/johnstf/timconc.htm. Acesso em: 27 de agosto de 2009.

Enviado em 01 de junho de 2011.
Aprovado em 14 de dezembro de 2011.