PROCEDIMENTOS METODOLÓGICOS EM ESTUDOS DA TRADUÇÃO: INTERFACE COM AS LINGUÍSTICAS SISTÊMICO-FUNCIONAL E DE CORPUS

Daniel Antonio de Sousa Alves
Eliza Mitiyo Morinaka

RESUMO : O arcabouço teórico e metodológico da Linguística de Corpus tem despertado a atenção de vários pesquisadores em Estudos da Tradução, abrindo-se, assim, uma vasta área de investigação. A rapidez com que as ferramentas tecnológicas processam uma grande quantidade de dados para análises de tradução impulsionou o avanço da disciplina desde a década de 1990. No Brasil, várias pesquisas foram feitas desde 2000, e este artigo visa apresentar um mapeamento de procedimentos metodológicos utilizados em pesquisas na interface entre Estudos da Tradução baseados em Corpora e Linguística Sistêmico-Funcional, dedicando uma grande atenção às instruções de funcionamento das ferramentas. Procura-se, dessa forma, oferecer um suporte a futuros pesquisadores nesta interface e abrir um diálogo para a construção de outros caminhos possíveis.
PALAVRAS-CHAVE: Estudos da Tradução, Linguística Sistêmico-Funcional, Linguística de Corpus.

RESUMEN. La estructura teórica y metodológica de la Lingüística de Corpus ha despertado la atención de muchos investigadores de los Estudios de Traducción, abriéndose así una vasta área de investigación. La rapidez con que las herramientas tecnológicas procesan una gran cantidad de datos para análisis de traducción impulsionó el avance de la disciplina desde la década de 1990. En Brasil, muchas investigaciones fueron hechas desde el año 2000, y este artículo tiene como finalidad presentar un mapeo de procedimientos metodológicos utilizados en investigaciones en la interfaz entre los Estudios de Traducción basados en Corpora y la Lingüística Sistémico Funcional, dedicando una gran atención a las instrucciones de funcionamiento de las herramientas. Se busca, con eso,

ofrecer soporte a futuros investigadores de esta interfaz y empezar un diálogo para la construcción de otros posibles caminos.
PALABRAS CLAVE: Estudios de Traducción, Lingüística Sistémico Funcional, Lingüística de Corpus.

1 Contextualização

As ferramentas tecnológicas garantem rapidez para processar uma grande quantidade de dados, atraindo a atenção de vários pesquisadores dos estudos da linguagem. Nas últimas décadas, a Linguística de Corpus (LC) vem se desenvolvendo e possibilitando a elaboração de trabalhos de naturezas diversas.[76] Na área de Estudos da Tradução (ET), as ferramentas de corpora vêm sendo aplicadas desde os primeiros trabalhos de Mona Baker[77] no começo da década de 1990, inicialmente orientadas para o processamento de corpora de grandes dimensões e, posteriormente, voltadas para a análise de corpora de pequenas dimensões.

Em 2001, Jeremy Munday propôs um modelo que integrava os Estudos da Tradução, a Linguística de Corpus e a Linguística Sistêmico-Funcional (LSF) pelo fato dessa teoria oferecer as ferramentas para as análises contextuais e culturais – um modelo que se mostrava bem sucedido, com o desenvolvimento de diversas pesquisas nessa interface nos anos que se seguiram. No Brasil, por exemplo, pesquisas desenvolvidas e orientadas pelas pesquisadoras Adriana Pagano (UFMG), Célia Magalhães (UFMG) e Maria Lúcia Vasconcellos (UFSC) adotaram essa orientação teórica e metodológica proposta por Sinclair.

Em 2004, após identificarem a necessidade de consolidar as metodologias utilizadas em suas pesquisas, Pagano, Magalhães e Vasconcellos propuseram um trabalho de cooperação entre as Universidades Federais de Minas Gerais e de Santa Catarina. Na ocasião, os autores deste artigo (então ligados à UFSC e à UFMG), apresentaram um manual contendo uma compilação dos procedimentos metodológicos adotados por

[76] Para um relato histórico da Linguística de Corpus ver Biber *et al* (1998) e Sardinha (2000, 2003).

[77] *Corpus Linguistics and Translation Studies: implications and applications.* O objetivo do artigo é procurar regularidades na língua da tradução em corpora de grandes dimensões, ou seja, a existência de uma linguagem única da tradução que se constituiria em universais da tradução. Ver também Baker (1995 e 1996).

pesquisadores(as) que desenvolviam trabalhos em Estudos da Tradução e interfaces com as Linguísticas Sistêmico-Funcional e de Corpus.

O resultado dessa cooperação interinstitucional foi um trabalho de 33 páginas, contemplando as metodologias de três monografias de graduação (a saber: Bueno, 2002; Franco, 2003; e Alves, 2004), quatro dissertações de mestrado (Campesatto, 2002; Cruz, 2003; Mauri, 2003; e Jesus, 2004) e uma tese de doutoramento (Fernandes, 2004). Na ocasião, também foram apontados quatro outros trabalhos em andamento (que, aqui, informamos com seus respectivos anos de conclusão: Feitosa, 2005; Morinaka, 2005; Paquilin, 2005; e Zuniga, 2006). Trabalhos desenvolvidos em outras instituições não foram incluídos na compilação devido à dificuldade em se obter cópias de dissertações e teses de outras instituições na época. O manual ficou disponível online de 2004 a 2009, tendo saído do ar apenas pela cessação das atividades do servidor no qual ele estava hospedado.

Apontada por Fábio Alves (2005) como uma contribuição pontual "para a consolidação de uma metodologia de pesquisa específica para estudos baseados em corpora de pequenas dimensões", a compilação foi utilizada por diferentes pesquisadores(as) como material de referência para o trabalho com corpora (citamos Morinaka, 2005; Jesus, 2008; e Fleury, 2011 como exemplos).

O objetivo deste artigo aqui apresentado é retomar a compilação de 2004 como um registro histórico, apresentando o mapeamento de procedimentos metodológicos compilados na época e abrindo espaço para discussões sobre avanços e possibilidades em termos metodológicos para trabalhos futuros na interface entre Estudos da Tradução baseados em Corpora e Linguística Sistêmico-Funcional.

2 Corpora

Corpora são compilações de textos eletrônicos, direcionados para análises linguísticas. Todos os corpora devem estar em formato eletrônico, de modo a possibilitar a análise por *software,* como, por exemplo, o *WordSmit Tools*. Os pesquisadores podem digitalizá-los manualmente, digitando-os ou capturando-os através de *scanners* e *softwares* de reconhecimento de caracteres (*OCRs*), ou obtê-los já digitalizados, capturando-os da internet

ou obtendo-os a partir da própria editora. O software reconhece textos nos seguintes formatos: .txt, .html, .sgml ou .xml .

No projeto Corpus Discursivo para Análises Linguísticas e Literárias (CORDIALL) do Núcleo de Estudos da Tradução (NET), mantido pela Faculdade de Letras da Universidade Federal de Minas Gerais, optou-se por adotar como padrão o formato .txt, por este ser o formato *default* do programa e por ser um formato mais simples e acessível que os demais. Os corpora devem ser catalogados em fichas de registro, contendo dados bibliográficos completos sobre os textos, e informações paratextuais se cabível. No projeto CORDIALL, optou-se também por marcar títulos, subtítulos, capítulos, seções e subgêneros dentro do gênero principal, por exemplo, um poema dentro de um romance, com anotações em etiquetas (detalhes na seção critérios de codificação). Categorias discursivas de análise, como, por exemplo, tema/rema, redes coesivas, transitividade e modalidade, também podem ser marcadas se necessário. Os passos seguidos para a manipulação dos corpora podem ser subdivididos em três etapas: preparação do corpus, processamento do corpus e análise do corpus.

2.1 PROCEDIMENTOS PARA A PREPARAÇÃO DO CORPUS

2.1.1 Textos não digitalizados

Digitalização do livro em arquivo eletrônico: utilizando-se *scanners* e *softwares* de reconhecimento de caracteres (*OCRs*), faz-se a digitalização do livro que, depois, deve ser salvo como documento do *Word*. Recomenda-se a digitalização página por página, mesmo que seja possível fazê-lo com o livro aberto.

Revisão e edição: é necessário revisar o texto a fim de corrigir erros de reconhecimento de caracteres. Por exemplo, a palavra "rio" pode aparecer como "no", pois o segmento "ri" pode ser reconhecido pelo *OCR* como a letra "n". Nessa etapa, recomenda-se trabalhar com o documento *Word*, pois o mesmo oferece um procedimento mecânico rápido de localização das palavras e substituição, através do comando 'Editar - Localizar' ou da combinação de teclas 'Ctrl+L'.

Conversão do Arquivo: consiste em converter o documento do *Word* para o formato de arquivo '.txt'.

Codificação: para maiores detalhes, consulte a seção 2.2.

2.1.2 Livros previamente digitalizados

Para os arquivos já digitalizados é necessário passar por todas as etapas descritas acima a partir da revisão e da edição.

2.2 CRITÉRIOS DE CODIFICAÇÃO

2.2.1 Palavras

Palavra é toda sequência de letras, que inclui hifens e apóstrofes em alguns casos, separada de outras palavras por um espaço ou qualquer sinal de pontuação. Pode ser também conceituada como 'a menor unidade de forma livre', ou seja, uma unidade do vocabulário e a menor unidade da sintaxe. (SINCLAIR, 1991, p. 176)

2.2.2 Detectando vocábulos (*types*) e ocorrências (*tokens*)

Vocábulo (*type*)[78] é cada palavra (diferente) usada em um texto (SINCLAIR, 1991). Por exemplo, o sintagma "características de uma mulher e uma menina" contém 6 (seis) vocábulos (*types*) pois a palavra "uma" ocorre 2 (duas) vezes. Ocorrência (*token*)[79] é o termo usado para se medir o tamanho do texto (SINCLAIR, 1991). Cada palavra é contada uma vez, mesmo que seja repetida. Por exemplo, o sintagma "características de uma mulher e uma menina" contém 7 (sete) ocorrências (*tokens*).

É importante ressaltar que o *software WordSmith Tools* não diferencia palavras com mais de um sentido. Nas frases abaixo, por exemplo, a palavra manga, mesmo sendo usada em diferentes contextos (com diferentes sentidos) será contada pelo software como duas ocorrências de um mesmo vocábulo.

[78] Sinclair utiliza a terminologia 'vocabulary'.
[79] Sinclair utiliza a terminologia 'running words'.

* Quando ele se chegou, ela puxou-lhe pela manga. * Pitangueira não dá manga.

2.2.3 Etiquetas (*Tags*)

São informações extras que podem ser adicionadas ao texto sendo trabalhado. Pode-se adicionar: etiquetas que indiquem os subgêneros dentro do gênero principal, como por exemplo, um poema dentro de um romance; etiquetas que indiquem classes gramaticais, por exemplo, verbo, substantivo ou adjetivo; e etiquetas que indiquem as categorias discursivas de análise, como, por exemplo, tema/rema, redes coesivas, transtividade e modalidade. Como um exemplo de etiquetamento de categorias discursivas de análise temos:

Entrou <Processo Material> de mansinho <Circunstância> e a <Participante> viu <Processo Mental> dormida numa cadeira, os cabelos longos espalhados nos ombros <Circunstância>.

Também é possível utilizar o etiquetamento para ignorar uma porção do texto, nesse caso, os *tags* devem estar juntos, ou seja, sem linhas em branco entre eles. Por exemplo:

Certo:	Errado:
"<Ignore esta informação> <Ignore esta também> TEXTO"	"<Ignore esta informação> <Ignore esta também> TEXTO"

Não há problema em existir espaços entre as etiquetas a serem ignoradas. Por exemplo:

"<Ignore toda essa informação>"

Não há como etiquetar o texto para que se diferenciem classes gramaticais na contagem de vocábulos e ocorrências. É possível somente ignorar algumas etiquetas que podem ser identificadas manualmente, por exemplo: "Ele mata<verbo> a mata<substantivo>".

2.2.4 Frases

Para que o programa detecte uma frase, deve haver um espaço ao fim desta (mesmo que ela esteja no fim do parágrafo).

Exemplo	Frases	Explicação
"WordSmith Tools is an integrated suite of programs for looking at how words behave in texts."	0	Como não há espaço depois do ponto final, o programa não detecta a frase.
"WordSmith Tools is an integrated suite of programs for looking at how words behave in texts. "	1	Existe um espaço depois do ponto final, o programa detecta e conta a frase.

2.2.5 Frases com reticências

Em todas as ocorrências de reticências, o programa obedece ao seguinte padrão: Se houver um espaço e uma letra maiúscula após as reticências, o programa considera como duas frases; e se houver uma letra minúscula após as reticências, o programa considera como uma só frase. Exemplos de formas de reticências:

* “A lemmatised head entry has... beside it.”
* “A lemmatised head entry has ... beside it.”
* “A lemmatised head entry has. . . beside it.”
* “A lemmatised head entry has . . . beside it.”
* “A lemmatised head entry has . . . beside it.”
* “A lemmatised head entry has . . . beside it.”

O programa consideraria todos os casos acima como uma frase cada, pois após as reticências e espaço existe a letra minúscula ‘b’.

* “A lemmatised head entry has... Beside it.”
* “A lemmatised head entry has ... Beside it.”
* “A lemmatised head entry has. . . Beside it.”
* “A lemmatised head entry has . . . Beside it.”
* “A lemmatised head entry has . . . Beside it.”
* “A lemmatised head entry has . . . Beside it.”

Já os exemplos anteriores, o programa contaria como duas frases cada, pois há um espaço e uma letra maiúscula após as reticências.

2.2.6 Frases em poemas

Com relação a poemas, o programa desconsidera as frases onde não há pontuação ou quando há uma letra minúscula depois da pontuação.

Exemplo	Frases	Explicação
“Tu mirada ES un atentado contra la razón tu sonrisa una bomba de tiempo para el corazón y tu piel es una trampa en la que vuelvo a caer	1	Não há espaço após o primeiro ponto final e a letra que o segue é minúscula.

una otra vez.(fim da primeira estrofe) son tus besos como un pasaporte rumbo a la pasión tus caricias son el sol ardiente que me derritió y tu cuerpo un terremoto, un volcan en el que sueño morir pero vuelvo a despertar.”		
“Tu mirada ES un atentado contra la razón tu sonrisa una bomba de tiempo para el corazón y tu piel es una trampa en la que vuelvo a caer una otra vez. (fim da primeira estrofe) son tus besos como un pasaporte rumbo a la pasión tus caricias son el sol ardiente que me derritió y tu cuerpo un terremoto, un volcan en el que sueño morir pero vuelvo a despertar.”	2	Há pontuação e letra maiúscula após os dois pontos finais.

2.2.7 Aspas

Em todos os casos de utilização de aspas abaixo, o programa conta duas frases. Deve-se estar atento para as outras regras e deixar um espaço após o fim da frase.

* “You can also see how frequent each one was in each of the texts”. The highest frequency will appear in red. * “You can also see how frequent each one was in each of the texts.” The highest frequency will appear in red.

* "You can also see how frequent each one was in each of the texts. " The highest frequency will appear in red.

2.2.8 Sinais de pontuação

Todos os sinais de pontuação são detectados normalmente pelo *software WordSmith Tools*.

2.2.9 Parágrafos:

Cada parágrafo é reconhecido pelo programa como uma sequência de dois toques na tecla <ENTER>. Se ela for pressionada várias vezes, cada sequência de dois toques será interpretada como um parágrafo. Ao fim do texto, a tecla deve ser pressionada somente duas vezes. Assim, o programa detectará o último parágrafo.

Exemplo	Parágrafos	Explicação
"WordSmith Tools is an integrated suite of programs for looking at how words behave in texts. You will be able to use tools to find out how words are used in your own texts, or those of others." "WordSmith Tools is an integrated suite of programs for looking at how words behave in texts. You will be able to use tools to find out how words are used in your own texts, or those of others.	0	Como a tecla <ENTER> não foi pressionada nenhuma vez após o final do texto, o software não reconhecerá nenhum parágrafo.
"The Wordlist tool lets you see a list of all the words or word-clusters in a text, set out in alphabetical or frequency order. The concordancer, Concord, gives you a chance to see any	0	A tecla <ENTER> foi pressionada apenas uma vez após o primeiro parágrafo e nenhuma vez após o segundo; logo,

word or phrase in context – so that you can see what sort of company it keeps. With KeyWords you can find the key words in a text." "WordSmith Tools is an integrated suite of programs for looking at how words behave in texts. You will be able to use tools to find out how words are used in your own texts, or those of others.		nenhum dos parágrafos será contado pelo software.
"The Wordlist tool lets you see a list of all the words or word-clusters in a text, set out in alphabetical or frequency order. The concordancer, Concord, gives you a chance to see any word or phrase in context – so that you can see what sort of company it keeps. With KeyWords you can find the key words in a text." "WordSmith Tools is an integrated suite of programs for looking at how words behave in texts. You will be able to use tools to find out how words are used in your own texts, or those of others.	1	A tecla <ENTER> foi pressionada duas vezes após o primeiro parágrafo e nenhuma vez após o segundo; assim, apenas o primeiro parágrafo será contado pelo software.
"The Wordlist tool lets you see a list of all the words or word-	2	A tecla <ENTER> foi pressionada duas

clusters in a text, set out in alphabetical or frequency order." "The concordancer, Concord, gives you a chance to see any word or phrase in context – so that you can see what sort of company it keeps. With Key-Words you can find the key words in a text."		vezes após cada parágrafo. O programa contar á os parágrafos corretamente.

Com relação a poemas, o programa considera a contagem de parágrafos quando há espaços entre as linhas.

Exemplo	Parágrafos	Explicação
"Amar o perdido deixa confundido este coração.(fim da primeira estrofe) Nada pode o olvido contra o sem sentido apelo do Não."(fim da segunda estrofe)	0	Não há linhas em branco nem entre os versos nem entre as estrofes.
"Amar o perdido deixa confundido este coração.(fim da primeira estrofe)	1	Há uma linha em branco entre as duas estrofes, mas não ao final do

Nada pode o olvido contra o sem sentido apelo do Não.”(fim da segunda estrofe)		texto.
“Amar o perdido deixa confundido este coração.(fim da primeira estrofe) Nada pode o olvido contra o sem sentido apelo do Não."(fim da segunda estrofe)	2	Há uma linha em branco entre as duas estrofes e uma ao final do texto.
“Amar o perdido deixa confundido este coração.(fim da primeira estrofe) Nada pode o olvido contra o sem sentido apelo do Não.”(fim da segunda estrofe)	3	Se forem deixadas linhas em branco também entre os versos, o programa os conta como parágrafos.

3 Ferramentas e procedimentos para análise

Na metodologia das dissertações e da tese analisadas destaca-se o uso de três ferramentas (*Wordlist, Keyword e Concord*) e um utilitário (*Viewer & Aligner*) do programa *WordSmith Tools*.

Wordlist: disponibiliza três listas de palavras do corpus - (i) por ordem alfabética, (ii) por frequência e (iii) dados estatísticos tais como o número de vocábulos, ocorrência, a razão vocábulo/ocorrência, número de frases, parágrafos, etc;

KeyWord: compara um corpus de estudo com um corpus de referência e identifica as palavras que apresentam, no corpus de estudo, uma frequência inesperada (tomando como parâmetro de comparação o corpus de referência), fornecendo uma lista de palavras chave;

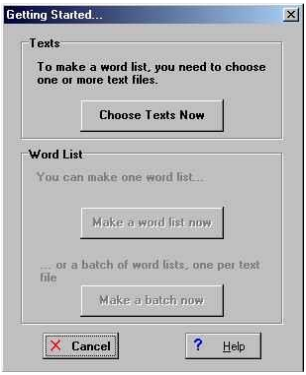
Concord: disponibiliza um índice da palavra que está sendo investigada e as respectivas linhas de concordância, ou seja, o contexto em que elas aparecem;

Viewer & Aligner: disponibiliza o texto fonte e o texto de chegada, alinhados por sentenças ou por parágrafos.

3.1 WORDLIST

Após acessar a opção *"START"* do menu , a caixa de diálogo *"Getting Started..."* será exibida. Clicando no botão *"CHOOSE TEXT NOW"*, podemos escolher o(s) arquivo(s) que contêm o corpus com o qual desejamos trabalhar. Será aberta outra caixa de diálogo em que é possível selecionar o(s) arquivo(s) do corpus. Antes de se prosseguir, é importante pressionar o botão *"CLEAR PREVIOUS"* para evitar que textos processados anteriormente sejam misturados com os atuais.

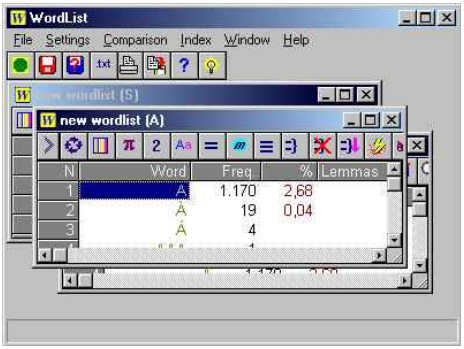
Figura 1
Caixa de diálogo *Getting Started* da ferramenta *WordList*



Seleciona-se então o(s) arquivo(s) a serem processados e pressiona-se o botão *OK*, para então clicar no botão *"MAKE A WORD LIST NOW"*. Após alguns momentos, as listas de palavras (*WordList*) serão exibidas na tela, organizadas da seguinte maneira:

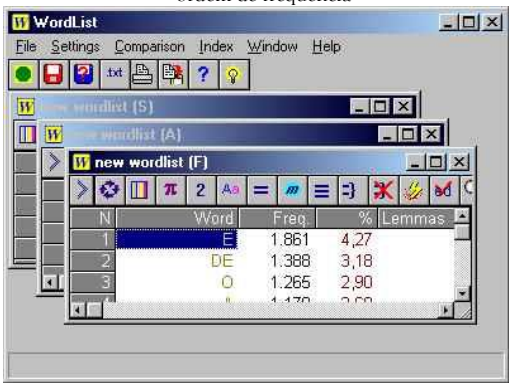
(i) Ordem alfabética - *New Wordlist (A)*: apresenta as palavras de um corpus de estudo em ordem alfabética e facilita o trabalho do pesquisador que trabalha com lemas. No estudo da transitividade, por exemplo, o pesquisador poderá ter um acesso rápido às formas de um verbo sendo investigado.

Figura 2
Lista de palavras - organizadas por ordem alfabética



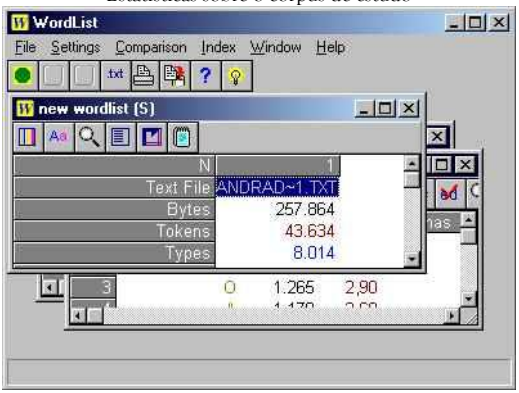
(ii) Frequência - *New Wordlist (F)*: apresenta as palavras de um corpus de estudo em ordem decrescente de ocorrência. Além de revelar dados importantes para o estudo comparativo de traduções, é uma lista útil para a investigação de *hapax legomena* como um indício de criatividade lexical;

Figura 3
Lista de palavras - organizadas por ordem de frequência



(iii) Dados estatísticos - *New Wordlist (S)*: apresenta estatísticas tais como o número de vocábulos, ocorrência, a razão vocábulo/ocorrência, número de frases, parágrafos, etc.

Figura 4
Estatísticas sobre o corpus de estudo



3.2 KEYWORDS

Essa ferramenta compara um corpus de estudo[80] com um corpus de referência e identifica as palavras que apresentam, no corpus de estudo, uma frequência inesperada (tomando como parâmetro de comparação o corpus de referência).

Em um exemplo imaginário, se trabalhamos com um corpus de referência de 500 palavras e um corpus de 100 palavras, suponhamos que a palavra 'superlativo' ocorra 25 vezes no corpus de referência e 10 vezes no corpus de estudo. A palavra 'superlativo' corresponderia, então, a 5% de todas as palavras do corpus de referência e a 10% de todas as palavras do corpus de estudo. Neste caso, tal palavra seria, provavelmente, dada como uma 'palavra chave', uma vez que sua frequência é duas vezes maior no corpus de estudo do que no corpus de referência.

Vale ressaltar que as palavras chave de um texto não são, necessariamente, as palavras que nele são mais frequentes. Como já mencionado anteriormente, para o *software*, palavras chave são aquelas que apresentam uma frequência inesperada.

Para facilitar a diferenciação entre as palavras chave (definidas pela ferramenta *KeyWords*) de um corpus e as suas palavras mais frequentes, faremos abaixo uma breve comparação entre as onze primeiras palavras chave do romance *Macunaíma - o herói sem nenhum caráter* e as onze palavras mais frequentes do mesmo romance.

Uma lista de todas as palavras no romance foi produzida, utilizando-se a ferramenta *WordList*, organizada por ordem de frequência:

[80] Aquele que se pretende descrever (Berber Sardinha, 1999, p.2)

Quadro 1
As onze palavras mais frequentes em
Macunaíma - o herói sem nenhum caráter.

N	Word	Freq.	%	Lemmas
1	E	1.861	4,27	
2	De	1.388	3,18	
3	O	1.265	2,90	
4	A	1.170	2,68	
5	Que	1.063	2,44	
6	Do	640	1,47	
7	Não	586	1,34	
8	Pra	522	1,20	
9	Com	512	1,17	
10	Macunaíma	496	1,14	
11	Da	487	1,12	

Foi também produzida uma lista com as palavras chave do romance utilizando-se a ferramenta *KeyWords*:

Quadro 2
Lista das onze primeiras palavras chave do romance
Macunaíma - o herói sem nenhum caráter

N	Word	Freq.	Andrad 1.Txt %	Freq.	Refptall.Lst %	Keyness	P
1	Pra	522	1,20	10		2.494,5	0,000000
2	Macunaíma	496	1,14	0		2.462,6	0,000000
3	Herói	274	0,63	13		1.255,5	0,000000
4	Jiguê	144	0,33	0		713,9	0,000000
5	Pro	129	0,30	3		611,4	0,000000
6	Maanape	113	0,26	0		560,1	0,000000
7	Então	245	0,56	408	0,09	422,2	0,000000
8	Manos	81	0,19	0		401,5	0,000000
9	Gigante	75	0,17	1		361,2	0,000000
10	Porém	147	0,34	162	0,03	329,5	0,000000
11	E	1.861	4,27	12.967	2,72	302,2	0,000000

Comparando-se a coluna *Word* do primeiro quadro com a coluna *Word* do segundo, é possível notar que as palavras chave de um corpus não necessariamente correspondem às suas palavras mais frequentes. Analisemos, por exemplo, a palavra 'e': com 1.861 ocorrências, tal palavra

corresponde a 4,27% do total de palavras do romance, sendo a palavra mais frequente deste corpus. Entretanto, a mesma palavra na lista de palavras chave é dada como a décima primeira palavra chave do mesmo corpus.

Além de apresentar as palavras que se mostram inesperadamente mais frequentes no corpus de estudo do que no corpus de referência, a ferramenta *KeyWords* também apresenta as palavras que se mostram inesperadamente menos frequentes no corpus de estudo do que no corpus de referência. Enquanto as primeiras são classificadas pela ferramenta como 'palavras chave positivas' (*positive key words*), as últimas são classificadas como 'palavras chave negativas' (*negative key words*).

De acordo com a definição dada no arquivo de ajuda do *software*, palavras chave positivas são aquelas que "ocorrem com mais frequência do que se esperaria em relação ao corpus de referência", e palavras chave negativas são aquelas que ocorrem com menos frequência do que se esperaria, considerando-se as mesmas condições. Retomando o exemplo imaginário anteriormente utilizado, suponhamos que a palavra 'comparativo' ocorra 25 vezes no nosso corpus de referência de 500 palavras e duas vezes no nosso corpus de 100 palavras. Nesse caso, a palavra 'comparativo' corresponderia a 5% das palavras do corpus de referência e a 2% das palavras do corpus de estudo. Como a frequência percentual de tal palavra no corpus de estudo corresponde a menos da metade do que se esperava (em relação ao corpus de referência) essa palavra, provavelmente, seria dada como uma palavra chave negativa.

3.2.1 Metodologia de trabalho com a ferramenta *KeyWords*

Após definir qual será o corpus de estudo, digitalizá-lo e corrigi-lo, será necessário compilar um corpus de referência. Berber Sardinha (1999, p. 3) recomenda que este corpus de referência tenha características gerais, composto por textos de diversos gêneros, e seja "cinco vezes maior que o de estudo" (Berber Sardinha, 1999, p. 11).

O primeiro passo a ser seguido é verificar o tamanho do corpus de estudo de modo a definir o tamanho do corpus de referência. Para tanto, utilizando a ferramenta *WordList*, é necessário produzir uma lista de palavras do corpus de estudo. Na lista de estatísticas aberta, verifica-se o

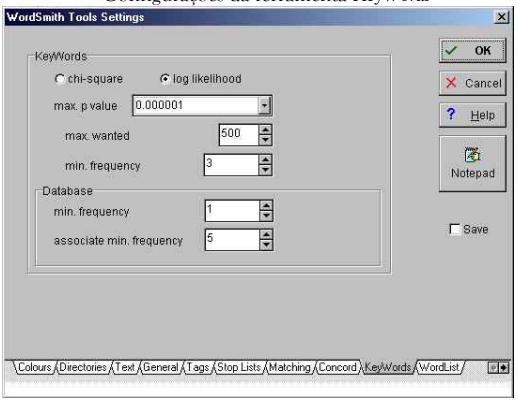
número de palavras (*tokens*) do corpus de estudo. Em seguida, multiplica-se esse número por cinco. O resultado obtido deverá ser o tamanho, aproximado, do corpus de referência.

A partir daí, será necessário compilar o corpus. Selecionam-se textos (também em formato .txt) de diversos gêneros (fictícios, jornalísticos, acadêmicos, etc.), tentando manter um razoável equilíbrio na proporção de cada gênero, e salva-se em uma pasta de trabalho. Produz-se, então, (utilizando a ferramenta *WordList*) uma lista de palavras de TODOS os arquivos que irão compor o corpus de referência.

Ambos os arrolamentos de palavras devem ser salvos, separadamente, na ferramenta *WordList* em formato .lst para que eles possam ser reconhecidos pelas outras ferramentas do *WordSmith Tools*. Tendo feito e salvo tais listas, será hora de usar a ferramenta *KeyWords*: primeiramente verificando suas configurações para depois usar as listas previamente salvas para produzir as relações de palavras chave.

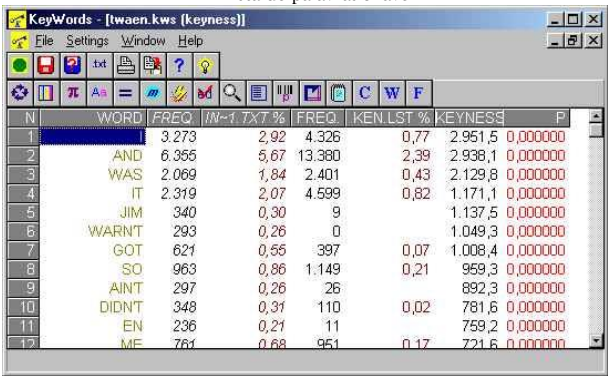
Tendo salvo as listas de palavras dos corpora necessários (de estudo e de referência), será necessário verificar as configurações da ferramenta *KeyWords* para produzir as listas de palavras chave. É preciso acessar no menu *"Settings"* da ferramenta *KeyWords*, a opção *"MIN. & MAX. FREQUENCIES"* e definir 16.000 como o valor máximo de palavras a serem exibidas (na caixa *"MAX.WANTED"*).

Figura 5
Configurações da ferramenta *KeyWords*



16.000 é o maior valor aceito pelo *software*, sendo, portanto, o que menos possivelmente imprimiria um corte artificial e não desejável que poderia causar uma eliminação de palavras significativas. Produz-se uma lista de palavras chave para depois acessar a opção "*START*" no menu da ferramenta *KeyWords*. Na caixa de diálogo aberta, pressiona-se o botão "*FIND THE KEYWORDS IN A TEXT*". Será aberta uma segunda caixa de diálogos em que os arquivos que contêm o corpus de estudo e o corpus de referência devem ser selecionados. Pressiona-se "*OK*" e após alguns momentos será exibida a lista de palavras chave.

Figura 6
Lista de palavras-chave



N	WORD	FREQ.	%N~1.TXT %	FREQ.	KEN.LST %	KEYNESS	P
1		3.273	2,92	4.326	0,77	2.951,5	0,000000
2	AND	6.355	5,67	13.380	2,39	2.938,1	0,000000
3	WAS	2.069	1,84	2.401	0,43	2.129,8	0,000000
4	IT	2.319	2,07	4.599	0,82	1.171,1	0,000000
5	JIM	340	0,30	9		1.137,5	0,000000
6	WARN'T	293	0,26	0		1.049,3	0,000000
7	GOT	621	0,55	397	0,07	1.008,4	0,000000
8	SO	963	0,86	1.149	0,21	959,3	0,000000
9	AIN'T	297	0,26	26		892,3	0,000000
10	DIDN'T	348	0,31	110	0,02	781,6	0,000000
11	EN	236	0,21	11		759,2	0,000000
12	ME	761	0,68	951	0,17	721,6	0,000000

3.3 CONCORD

Acessa-se a opção "*START*" do menu "*FILE*". Será exibida a caixa de diálogo "*Getting Started...*". Clicando o botão "*CHOOSE TEXT NOW*", podemos escolher o arquivo que contém o corpus com o qual desejamos trabalhar. Será aberta a caixa "*Choose Text(s)*".

É importante, antes de se prosseguir, pressionar o botão "*CLEAR PREVIOUS*" para evitar que textos processados anteriormente sejam misturados com os atuais. Seleciona-se, então, o(s) arquivo(s) a serem processados e pressiona-se o botão *OK*. A caixa "*Choose Text(s)*" será fechada e a caixa de diálogo "*Getting Started...*" voltará a ser exibida. Dessa vez, entretanto, o botão "*specify search-word*" estará habilitado.

Ao se pressionar o botão *"specify search-word"*, a caixa *"Concordance Settings"* será aberta. Nessa caixa, é possível definir:

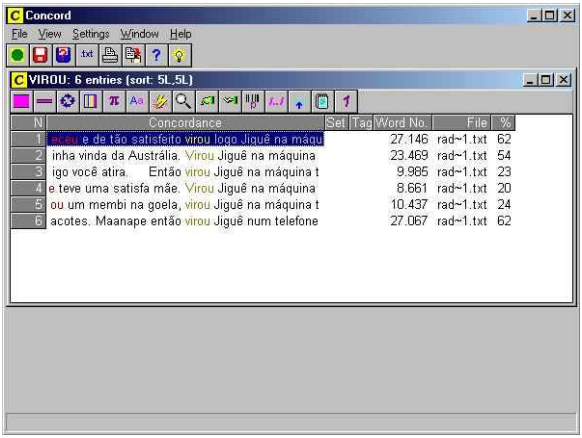
* A palavra nódulo (que é a palavra a ser pesquisada);

* Horizontes de contexto (ou um co-texto);

* A distância (tanto à esquerda quanto à direita) entre a palavra nódulo e o co-texto determinado;

Definidas as palavras a serem pesquisadas, basta pressionar o botão *"GO NOW"*. Após alguns momentos, uma tela com as linhas de concordância pedidas será exibida. A palavra definida como nódulo aparecerá em destaque ao centro da tela.

Figura 7
Linhas de concordância geradas pela ferramenta *Concord*



3.4 VIEWER & ALIGNER

Para alinhar o original e a tradução, basta clicar na opção *"Getting Started..."*. Aparecerá uma caixa com o comando *"Source Text"*. Ao clicar esse botão, aparecerá uma outra caixa *"Open a text file"*. Seleciona-se o texto original e pressiona-se o comando "OK". O texto aparecerá em uma tela à parte ao lado da caixa de comando. Para alinhá-lo com a tradução, basta clicar o comando "*Translation*" e selecionar a tradução.

4 Considerações finais

Ao organizar os procedimentos metodológicos utilizados em pesquisas na interface entre Estudos da Tradução baseados em Corpora e Linguística Sistêmico-Funcional, indicamos uma das possibilidades metodológicas para o estudo de textos traduzidos, que futuramente pode vir a estabelecer diálogo com as outras áreas de pesquisa para o contínuo crescimento dos Estudos da Tradução.

5 Referências

ALVES, Daniel Antonio de Sousa. *Focusing on KeyWords*: a study of a parallel bilingual corpus. 2004. Monografia (Graduação em Letras) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

ALVES, Fábio. *Relatório Avaliativo Programa Nacional de Cooperação Acadêmica PROCAD 097/10-2: Tradução e Linguística de Corpora: em busca de interfaces com vistas a processos de categorização cognitiva e contextualização pragmático/política*. Disponível em: http://goo.gl/66nlmQ - último acesso em 21 de novembro de 2013.

BAKER, Mona. Corpus Linguistics and Translation Studies: implications and applications. In: BAKER, Mona; FRANCIS, Gil and TOGNINI-BONELLI, Elena. (Ed.). *Text and technology*: in honour of John Sinclair. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1993. p 233-250.

BAKER, Mona. Corpora in Translation Studies: an overview and some suggestions for future research. In: *Target*, Amsterdam – Holland. v. 7 n. 2, 1995. p 223-243.

BAKER, Mona. Corpus-based translation studies: the challenges that lie ahead. In: SOMERS, Harold. (Ed.). *Terminology, LSP and translation:* studies in language engineering in honour of Juan C. Sager. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1996. p 177-186

BERBER-SARDINHA, Tony. Linguística de corpus: histórico e problemática. In: *D.E.L.T.A.*, São Paulo – SP. v. 16 n. 2, 2000. p 323-367.

_____. Comparing corpora with WordSmith keywords. In: *The ESPecialist*, São Paulo – SP, 2001. v. 22 n. 1. p 87-99.

_____. *Linguística de Corpus*. São Paulo: Manole, 2004. 410 p.

BIBER, Douglas; CONRAD, Susan e REPPEN Randi. *Corpus Linguistics:* investigating language structure and use. Cambridge: CUP, 1998. 300 p.

BUENO, Letícia Taitson. *Creative lexical items in Macunaíma*: a corpus-based research. 2002. 80 p. Monografia (Graduação em Letras) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

CAMPESATTO, Lucila Augusta. *Thematic structure in Brazilian Portuguese abstracts in English translation*: the issue of textual competence. 2002. 99 p. Dissertação (Mestrado em Letras/Inglês e Literatura Correspondente) – Centro de Comunicação e Expressão, Universidade Federal de Santa Catarina, Florianópolis.

CRUZ, Osilene Maria de Sá e Silva. *Harry Potter and the Chamber of Secrets e sua tradução para o português do Brasil*: uma análise dos verbos de elocução com base na Linguística Sistêmica e nos Estudos de Corpora. 2003. 207 p. Dissertação (Mestrado em Estudos Linguísticos), Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

FEITOSA, Marcos Pereira. *Uma proposta de anotação de corpora paralelos com base na Linguística sistêmico-funcional*. 2005. 137 p. Dissertação (Mestrado em Estudos Linguísticos), Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

FERNANDES, Lincoln Paulo. *Brazilian practices of translating names in children's fantasy literature*: a corpus-based study. 2004. 270 p. Tese, (Doutorado em Letras/Inglês e Literatura Correspondente) – Centro de Comunicação e Expressão, Universidade Federal de Santa Catarina, Florianópolis.

FLEURY, Lilian J. *Proposta de Sistematização Metodológica para Pesquisas em Análise Textual e Tradução: Uma Interface com a Linguística Sistêmico-Funcional e Linguística de Corpus*. Apresentação em outubro de 2011.

Disponível em: http://goo.gl/HSgBrT - último acesso em 21 de novembro de 2013.

JESUS, Silvana Maria de. *Representação do discurso e tradução*: padrões de textualização em corpora paralelo e comparável. 2004. 126 p. Dissertação (Mestrado em Estudos Linguísticos), Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

JESUS, Silvana Maria de. *Relações de tradução: SAY/DIZER em corpora de textos ficcionais*. 2008. 209 p. Tese (Doutorado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais.

MAURI, Cristina. *Um estudo da tradução italiana de Laços de família, de Clarice Lispector, a partir da abordagem em corpora*: a construção da introspecção feminina através dos verbos de elocução. 2003. 108 p. Dissertação (Mestrado em Estudos Linguísticos), Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

MORINAKA, Eliza Mitiyo. *Gabriela, cravo e canela and its (re)textualization in English*: representation through lexical relations. 2005. 108 p. Dissertação (Mestrado em Letras/Inglês e Literatura Correspondente) – Centro de Comunicação e Expressão, Universidade Federal de Santa Catarina, Florianópolis.

PAQUILIN, Viviane. *The various facets of a message*: an analysis of the thematic structure in Bridget Jones's Diary in the light of the systemic functional grammar, corpus linguistics and translation studies interface. 2005. 110 p. Dissertação (Mestrado em Letras/Inglês e Literatura Correspondente) – Centro de Comunicação e Expressão, Universidade Federal de Santa Catarina, Florianópolis.

SCOTT, Mike. *WordSmith Tools*. Oxford: Oxford University Press, 1996.

SILVA e FRANCO, Arabela Vieira dos Santos. *The semantic prosody of three keywords in a Parallel Corpus (Canadian English / Brazilian Portuguese) short stories*: a corpus-based translation study. 2003. 87 p. Monografia (Graduação em Letras) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

SINCLAIR, John. *Corpus, concordance, collocation*. Oxford: OUP, 1991. 179 p.

SINCLAIR, John. Preface. In: GHADESSY, Mohsen *et al* (Ed.). *Small corpus studies and ELT*: theory and practice Amsterdam: John Benjamins, 2001. p vii-xv.

ZUNIGA, Gleimara. Regina Ferreira. *Construing the translator*: a meta-reflection grounded in corpus-based translation studies and systemic functional linguistics. 2006. 105 p. Dissertação (Mestrado em Letras/Inglês e Literatura Correspondente) – Centro de Comunicação e Expressão, Universidade Federal de Santa Catarina, Florianópolis.