

Usando WordSmith Tools na investigação da linguagem

Tony Berber Sardinha

LAEL, PUCSP

Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem

Pontifícia Universidade Católica de São Paulo

Rua Monte Alegre 984

05014-001 São Paulo, SP, Brazil

tony4@uol.com.br

<http://sites.uol.com.br/tony4/homepage.html>

DIRECT Papers 40

ISSN 1413-442x

1999

1. Introdução

Tem havido uma crescente popularização dos computadores na vida cotidiana, e por isso cada vez o computador se faz presente nos mais diversos tipos de ambiente. Hoje em dia há computadores nas fábricas, bancos, consultórios médicos, estacionamentos, e restaurantes, para citar apenas alguns lugares. O computador também está presente em vários objetos do dia-a-dia, como automóveis, telefones, e relógios. Desta forma, o computador está influenciando na vida pessoal e profissional de um número maior de indivíduos.

Esta influência, no entanto, não é nova. O computador tem entrado nos ambientes profissionais há um bom tempo, pelo menos há 48 anos, quando foi lançado o primeiro computador digital comercial, o UNIVAC 1, pela IBM. Num nível mais pessoal, a entrada do computador se deu mais tarde, por volta do final dos anos 70, com a comercialização dos primeiros computadores pessoais, e já tem, portanto, cerca de 20 anos.

Em uma esfera da atividade humana, contudo, a adoção do computador como ferramenta de trabalho tem acontecido de modo mais tardio: a análise da linguagem. A bem da verdade, o emprego de computadores no armazenamento de dados para análise linguística já tem por volta de 40 anos, contados a partir do início da coleta do corpus Brown, o primeiro corpus linguístico computadorizado. Desde então, e devido à popularização dos computadores nas universidades, tem havido um aumento vertiginoso de estudos que se valem do computador como instrumento de análise ou armazenamento de dados. Mesmo assim, a parcela da pesquisa linguística assistida por computador ainda é minoritária. A maioria dos estudos da linguagem ainda se faz em cima de pequenas quantidades de dados, coletados, mantidos, e analisados à mão. O comentário de Phillips (1989) de 20 anos atrás segundo o qual a linguística se limita à análise de dados que cabem no quadro-negro ainda se aplica aos dias de hoje, infelizmente.

As razões para esta escassez do computador na pesquisa linguística são várias. Uma delas é a falta de conhecimento dos instrumentos disponíveis. Muitos pesquisadores acatam com prazer o computador e software de análise quando mostrados como utilizar

ferramentas computacionais na sua pesquisa acadêmica. Uma outra razão é a rejeição do computador e dos modelos de análise da natureza mais empírica que ele favorece. Esta atitude em geral vem como consequência de um posicionamento em relação à pesquisa linguística que não vê a necessidade de análise de dados empíricos na pesquisa da linguagem.

Um maior emprego de computadores na investigação da linguagem seria benéfica. Em primeiro lugar, eles são consistentes. Os computadores não se cansam, e assim podem fazer tarefas tediosas (como contar palavras, identificar todas as ocorrências de um termo, classificar a ordem dos itens listados, etc.) de modo eficiente e confiável (Biber et al., 1998; Stubbs, 1996, p.232). Em segundo lugar, eles permitem maior abrangência no quantidade de dados que se pode lidar (Biber, 1988). O exame de um corpus de 1 milhão de palavras é uma tarefa quase impossível para o ser humano, mas para um computador, mesmo um do tipo pessoal de mesa, é algo que se faz em poucos segundos. As vantagens relativas à consistência e à abrangência não são as únicas, porém.

Uma outra vantagem diz respeito à possibilidade da descoberta de fatos novos, ou mesmo da contestação de opiniões e crenças estabelecidas (Stubbs, 1996, p.232). A incorporação do computador à análise da linguagem pode ser comparada à introdução do microscópio nas ciências, séculos atrás:

‘Dentro de um curto período de tempo, os linguistas adquiriram novas técnicas de observação. A situação é semelhante ao período imediatamente anterior à invenção do microscópio e do telescópio, os quais de repente permitiram que os cientistas observassem coisas que jamais tinham visto.’ (Stubbs, 1996, p.231, tradução minha)

Além de permitir enxergar fenômenos novos, o computador pode também modificar como se enxerga a linguagem:

‘o desenvolvimento do computador com memória poderosa seria para a linguística o que o desenvolvimento do microscópio com lentes poderosas foi para a biologia – uma oportunidade não somente de ampliar nosso conhecimento mas de transformá-lo’ (Hoey, 1993; tradução minha)

Um dos fatores que pode favorecer o emprego maior de ferramentas computacionais na análise linguística é a existência de programas flexíveis e fáceis de usar. Neste sentido, um dos programas que cumpre estas exigências é WordSmith Tools, escrito por Mike Scott, e publicado pela Oxford University Press. WordSmith Tools já tem pelo menos 5 anos de existência (desde o lançamento em pequena escala de seus protótipos), e está em vias de aparecer na sua terceira versão.

Apesar de fácil de flexível e fácil de usar, o programa coloca à disposição do analista uma série de recursos, os quais, se bem usados, são extremamente úteis e poderosos na análise de vários aspectos da linguagem. Entre estes aspectos, estão a composição lexical, a temática de textos selecionados, e a organização retórica e composicional de gêneros discursivos. O presente trabalho pretende oferecer uma visão geral dos recursos do programa, e como estes recursos podem ser empregados, direta ou indiretamente, na consecução de vários tipos de investigação acerca da linguagem.

2. Componentes

O WordSmith Tools é composto de (a) ferramentas, (b) utilitários, (c) instrumentos, e (d) funções.

Há três ferramentas e quatro utilitários¹, nomeadamente:

- Ferramentas:
 1. WordList
 2. KeyWords
 3. Concord
- Utilitários:
 1. Renamer
 2. Text Converter
 3. Splitter
 4. Viewer

Os instrumentos de análise disponíveis nas três ferramentas totalizam 17, e são os seguintes (com os nomes em inglês usados no programa em parênteses):

- WordList:
 1. Lista de palavras individuais ('wordlist').
 2. Lista de múlti-palavras ('wordlist, clusters activated').
 3. Lista de palavras de consistência individuais ('detailed consistency').
 4. Lista de múlti-palavras² de consistência ('detailed consistency, clusters activated').
 5. Lista de dimensões e densidade lexical ('statistics').
- Concord:
 1. Concordância ('concordance').
 2. Lista de colocados ('collocates').
 3. Lista de agrupamentos lexicais³ ('clusters').
 4. Lista de padrões de colocados ('patterns').
 5. Gráfico de distribuição da palavra de busca ('plot').
- KeyWords:
 1. Lista de palavras chave ('keywords').
 2. Banco de dados de listas de palavras chave ('database').
 3. Lista de palavras chave chave ('key keywords').
 4. Lista de palavras chave associadas ('associates').
 5. Lista de agrupamentos textuais ('clumps').
 6. Gráfico de distribuição de palavra chave ('keyword plot').
 7. Listagem de elos entre palavras chave ('keyword plot links').

¹ O menu 'Tools / Utilities' apresenta um quinto utilitário, que é 'File Manager', mas este na verdade nada mais é do que o Gerenciador de Arquivos do próprio Windows.

² Assim chamadas nesse trabalho porque esse é o termo corrente na literatura na área da fraseologia ('multi-word units', 'MWU's', 'polywords', etc.; xxx). Outro termo compatível é n-grama, comum na área de Processamento de Linguagem Natural (xxx). Mais especificamente, usa-se bi-grama para conjuntos de duas palavras, tri-grama para três, e assim por diante.

³ Também conhecidos por 'lexias complexas' em um outro programa de computador similar ('Hyperbase', xxx).

As funções principais distribuídas nas três ferramentas são:

1. Lematização: agrupamento de duas ou mais formas diferentes em um mesmo item. Por exemplo, as formas ‘correm’ e ‘correram’ podem ser agrupadas sob o lema⁴ ‘correr’.
2. Classificação: ordenação de listas e concordâncias por ordem alfabética, frequencial, ou por posição (na ‘lista de colocados’).
3. Delimitação: escolha de quais partes do corpus serão lidas pelo programa. É útil porque permite ignorar cabeçalhos de textos etiquetados.

O funcionamento das ferramentas é influenciado por alguns tipos de ajustes que o usuário pode efetuar:

1. Horizonte de concordância: Janela, ou quantidade de palavras, dentro da qual o programa calcula os colocados. Esse valor também é usado na listagem de elos entre palavras chave (‘keyword plot links’).
2. Tamanho da múlti-palavra: Determina o tamanho que terá cada múlti-palavra (em palavras, ou seja, tamanho 2 indica múlti-palavras de dois elementos, tamanho 3 de três elementos, e assim em diante).
3. Tamanho do agrupamento lexical: Semelhante ao anterior, só que influencia os agrupamentos encontrados na concordância, e na produção de listas de palavras.
4. Prova estatística para identificação das palavra chave: Teste estatístico que será executado durante a comparação das frequências do corpus de estudo com o de referência.
5. Frequência mínima: Número mínimo que um item deve possuir para se tornar elegível para um instrumento. Vários instrumentos possuem ajustes independentes de frequências mínimas.
6. Extensão máxima: Número máximo de ocorrências de uma concordância ou de uma lista de palavras chave, por exemplo. O analista deve ter consciência desse valor pois a listagem ou concordância pode ser incompleta, caso haja mais casos do que o número máximo estipulado.

A apresentação do programa será feita em três etapas. Na primeira, serão descritas as ferramentas. A seguir, passar-se-á à apresentação dos instrumentos, e por fim serão feitas recomendações de procedimentos possíveis de serem utilizados para se responder algumas questões de pesquisa típicas. Para uma descrição detalhada dos comandos necessários para se acessar as ferramentas e instrumentos, o leitor deve se referir ao manual do programa WordSmith Tools (no arquivo ‘manual.doc’ que acompanha a distribuição).

3. Princípios abstratos

O programa WordSmith Tools, assim como outros programas de computador para análise linguística, funciona com base em três princípios abstratos básicos:

- (1) Ocorrência. Os itens devem estar presentes; itens que não ocorreram não são incorporados porque não são observáveis; na presença de regras pré-definidas é possível prever quais itens deveriam ocorrer, mas na ausência destes construtos prévios não é possível fazer tal previsão.

⁴ Também conhecido por ‘lexema’ na linguística.

- (2) **Recorrência.** Os itens devem estar presentes pelo menos duas vezes; isto não significa que itens de frequência 1 não tenham relevância. Pelo contrário, enquanto um nível de frequência ('ranking') eles são importantes, tanto que são conhecidos por um rótulo específico, qual seja o de 'hapax legomena'. Sabe-se que os 'hapax' formam a maioria dos itens da linguagem, por isso um corpus é representativo na medida em que representa estes itens. Itens de frequência 1 são em geral raros, e é a existência de itens raros que pressupõe a necessidade de corpora grandes na pesquisa, pois corpora maiores dão mais chance de itens raros aparecerem.
- (3) **Co-ocorrência.** Os itens devem estar na presença de outros. Um item isolado é muito pouco informativo. Ele obtém significância na medida que é interpretado como parte de um conjunto formado por outros itens. A co-ocorrência não implica em aparição sequencial, ou seja, os itens podem ou não ter co-ocorrido em sequência no discurso. O horizonte de co-ocorrência é uma janela que pode ir de algumas palavras ao redor de um item, às fronteiras do texto, ou até mesmo compreender um corpus multi-textual inteiro. Em outras palavras, é a orientação da pesquisa que vai determinar a amplitude desta janela de co-ocorrência. Por exemplo, o fato do item 'carta' aparecer no mesmo texto de 'registrada' (portanto numa janela do tamanho do texto) pode ser tão relevante quanto estes dois itens aparecerem imediatamente adjacentes (portanto numa janela de duas palavras de largura) formando a expressão 'carta registrada'.

4. As ferramentas

A seguir é feita uma descrição das ferramentas e de seus principais instrumentos.

4.1. WordList

Esta ferramenta propicia a feitura de listas de palavras. 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 destas listas é apresentada em uma janela diferente, e juntamente com as duas janelas correspondentes à lista alfabética ('A') e por 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.

4.1.1. Comandos principais⁵

Para se obter uma lista de palavras, os comandos básicos são os seguintes:

- (1) No Controller (a primeira janela que aparece ao se inicializar o programa), clique em Tools e depois em WordList;
- (2) Na janela do WordList, clique em File e depois em Start;

⁵ Esses comandos e os sugeridos na apresentação das demais ferramentas se referem à versão 2.0 do WordSmith Tools. Outras versões podem apresentar comandos ou telas diferentes.

- (3) Na janela 'Getting Started', clique em 'Choose Texts Now' se estiver fazendo a primeira lista desde que iniciou o programa. Se já tiver escolhido os textos e quiser mantê-los, clique em 'Make a WordList Now', mas se quiser mudar de textos, clique em 'Change Selection'. Na janela 'Choose Texts', clique em 'Clear Previous' e faça a seleção segundo as instruções a seguir.
- (4) Na janela 'Choose Texts', clique na pasta onde se encontram os textos, depois nos textos que deseja, e por fim em OK. Se os textos estiverem em mais de uma pasta, clique em 'Store' ao terminar de selecionar os textos da primeira pasta, mude de pasta, selecione os outros textos, e só quando tiver escolhidos os textos de todas as pastas clique em OK.
- (5) Na janela 'Getting Started', clique em 'Make a WordList Now'.

4.1.2. Listas alfabética e freqüencial

A lista de palavras ordenada alfabeticamente possui os seguintes elementos:

- Coluna 'Word': os itens (em geral palavras) contidos no(s) texto(s);
- Coluna 'Freq.': quantas vezes cada item ocorreu;
- Coluna '%': a porcentagem do total de itens do texto a que corresponde cada item;
- Coluna 'Lemmas': outros itens cujas frequências foram adicionadas ao item corrente. 'Lemas' ('lemmas' ou 'lemmata', em inglês e latim) são itens lexicais que incorporam formas derivadas. Por exemplo, o lema 'correr' pode compreender as formas 'corro', 'corre', 'correndo', 'correr', 'corrido', etc. É análogo ao conceito de 'lexema'. Para fazer a lematização, siga os passos abaixo:
 - Clique no item que será o lema ('head');
 - Pressione F5 ou clique no botão 'mark/unmark';
 - Clique nas demais formas que serão incluídas no lema;
 - Pressione F5 ou clique no botão 'mark/unmark';
 - Clique em 'join'. As formas selecionadas mudam para a cor cinza e passam a constar na coluna 'Lemmas' ao lado do lema ('head'). A frequência do item também muda, mostrando a somatória das frequências das várias formas incluídas no lema;
 - Para que as formas na cor cinza desapareçam, clique em 'Zap'. Ao fazer isto, a janela de estatísticas fica vazia.

4.1.3. Lista de estatísticas

Os principais elementos⁶ constantes dessa janela são:

- Coluna 1, 2, 3, ...: número de cada arquivo;
- Text File: nome do arquivo;
- Tokens: número de itens (ou ocorrências); por exemplo, a frase 'o João viu o Pedro' possui cinco itens: 'o' (1), 'João' (2), 'viu' (3), 'o' (4), 'Pedro' (5);
- Types: número de formas (ou vocábulos); a frase acima possui quatro formas: 'o' (2), 'João' (2), 'viu' (3), 'Pedro' (4);
- Type-Token Ratio: a razão forma/item (ou vocábulo/ocorrência) expressa em porcentagem; a razão forma/item (cuja abreviação é TT em inglês, e VO ou FI em

⁶ Há várias outras estatísticas menos importantes para os tipos de análise considerados neste trabalho incluídas na janela, e para estas o usuário pode recorrer ao 'help' do programa para maiores explicações.

português), na sua forma tradicional, é obtida dividindo-se o total de formas pelo total de itens. Na frase acima, a razão FI é 0,8 ($4 \div 5$). No WordList, entretanto, transforma-se este valor em porcentagem; assim, divide-se o total de formas pelo total de itens dividido por cem. Na frase ‘o João viu o Pedro’, portanto, o valor da razão fornecido pelo programa seria 8,0, ou seja, $4 \div (5 \div 100)$. Na prática, a razão forma/item indica a riqueza lexical do texto. 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;

- Standardised Type-Token: a razão forma/item padronizada. No cálculo da razão FI demonstrado acima, leva-se em conta todas as palavras do(s) texto(s) selecionado(s). A forma FI padronizada, pelo contrário, calcula FI em intervalos regulares, ou seja, faz este mesmo cálculo por partes do texto, e depois tira a média dos valores FI entre os vários trechos. Por exemplo, o texto ‘o João viu o Pedro. O Paulo, entretanto, viu o João mas não o Pedro’ possui uma razão FI igual a 60,0 (9 formas, 15 itens, ou seja, $9 \div (15 \div 100)$). Mas se dividirmos o texto na metade, teremos uma parte que consiste de ‘o João viu o Pedro. O Paulo, entretanto’, com 6 formas e 8 ocorrências (portanto FI = 75,0), e outra que corresponde a ‘viu o João mas não o Pedro’, com 6 formas e 7 itens (FI = 85,7). Tirando-se a média aritmética, chega-se a 80,3 ($75 + 85,7 \div 2$), que seria o valor da razão FI padronizada. A forma padronizada é empregada para neutralizar a influência do tamanho do texto na computação da razão FI, já que textos maiores por natureza apresentam mais repetições e por isso tendem a possuir valores de FI mais baixos do que textos curtos. A razão FI é portanto sensível à extensão do material textual, não sendo assim confiável para uso em comparações entre textos de tamanhos diferentes (que são a norma, aliás; textos autênticos de extensão igual são extremamente raros). A diferença entre os valores de FI simples (60,0) e padronizado (85,7) obtidos acima ilustra este efeito. O texto inteiro, por ser maior, dá mais espaço para repetições, as quais de fato ocorreram, e daí seu valor FI é mais baixo. O cálculo padronizado, por sua vez, impediu que se levasse em conta a repetição de palavras ocorridas no outro trecho, resultando assim em um valor médio mais alto.

Note que as estatísticas relativas ao número de sentenças ou períodos (‘sentences’) e parágrafos (‘paragraph’) dependem das convenções usadas para definir tais unidades, e por isso é necessário ter certeza de que os textos usados respeitam estas convenções. Segundo os ajustes pré-definidos, o programa identifica como sentença o espaço de caracteres entre marcas de pontuação (‘.!?’), e como parágrafo o espaço de texto que termina com uma linha em branco (conseguido ao se pressionar duas vezes consecutivas a tecla ‘enter’). Assim, a menos que os textos escolhidos tenham seguido rigorosamente estas convenções de delimitação de sentença e parágrafo, as contagens apresentadas pelo programa serão incorretas.

4.2. KeyWords

Esta ferramenta permite a seleção de itens de uma lista de palavras (ou mais) por meio da comparação de suas frequências com uma lista de referências. KeyWords contrasta uma lista de palavras (ou mais de uma) de um corpus de estudo com uma lista de palavras de um corpus de referência. O resultado do contraste é uma lista de palavras chave, ou palavras cujas frequências são estatisticamente diferentes no corpus de estudo e no corpus de referência.

Os componentes principais de uma análise de palavras chave são, portanto, dois:

- (a) um corpus de estudo, representado por uma lista de frequência de palavras. O corpus de estudo é aquele que se pretende descrever. A ferramenta KeyWords aceita a análise simultânea de mais de um corpus de estudo.
- (b) um corpus de referência, também formatado como uma lista de frequência de palavras. Também é conhecido como ‘corpus de controle’, e funciona como termo de comparação para a análise. A sua função é a de fornecer uma norma com a qual se fará a comparação das frequências do corpus de estudo. A comparação é feita através de uma prova estatística selecionada pelo usuário (qui-quadrado ou log-likelihood). *As palavras cujas frequências no corpus de estudo forem significativamente maiores segundo o resultado da prova estatística são consideradas chave*, e passam a compor uma listagem específica de palavras chave.

O procedimento seguido pelo programa para extração de palavras chave pode ser resumido segundo o algoritmo abaixo:

- (1) Selecione o primeiro item na lista de palavras do corpus de estudo;
- (2) Procure por este item na lista de palavras do corpus de referência;
- (3) Se o item constar do corpus de referência, vá para o passo a seguir, senão passe para o passo 7;
- (4) Compare as frequências através de uma prova estatística escolhida pelo usuário (log-likelihood é ‘default’, mas qui-quadrado também está disponível);
- (5) Se o resultado da comparação for estatisticamente significante (segundo o nível de significância definido pelo usuário), copie esta palavra para uma nova lista, e chame-a de lista de palavras chave;
- (6) Repita este procedimento até o último item da lista de palavras do corpus de estudo;
- (7) Se um item constante da lista de palavras do corpus de estudo não aparecer na lista de palavras do corpus de referência, assumo frequência 0 para o item no corpus de referência;
- (8) Execute os passos 4, 5, e 6.

4.2.1. Critérios de escolha dos corpora

São apresentados abaixo alguns critérios para seleção dos corpora de estudo e referência, quanto à sua extensão e composição.

- (1) O tamanho mínimo necessário para um corpus de estudo depende (a) do que se pretende estudar, e (b) da especialização do corpus. Em se tratando de categorias morfo-sintáticas, Berber Sardinha (2000) aplicou a metodologia de estimação de amostras linguísticas significativas sugerida por Biber (Biber, 1990, 1993) e encontrou os seguintes valores, em quantidade de palavras, como amostras mínimas para o inglês escrito:

Categoria	Corpus geral	Corpus especializado
Verbo	67.187	13.848
Substantivo	74.551	8.555
Adjetivo	149.694	21.234

Advérbio	205.206	68.953
Pronome	913.256	40.945
Numeral	1.180.815	91.161

Em termos do tamanho total do corpus em número de palavras, os achados de Berber Sardinha (2000) indicam os seguintes valores mínimos para categorias morfo-sintáticas:

Corpus	Todas as categorias	Excluindo-se a mais rara	Palavras de conteúdo
Geral	5.495.048	1.180.815	205.206
Especializado	91.161	65.432	68.953

Em relação a itens lexicais específicos, os tamanhos mínimos representativos de amostras também variam de acordo com a língua e a especialização do corpus. Segundo Berber Sardinha (2000), no caso do inglês escrito, as amostras mínimas para alguns itens seriam:

Item	Corpus	
	Geral	Especializado
And	142.346	36.703
Be	747.168	504.269
Can	2.191.802	978.534
Have	775.047	203.149
I	4.109.098	55.397
May	3.569.167	1.651.178
My	6.470.793	110.162
The	102.787	38.637
This	659.339	306.823
Will	2.763.524	860.985
You	4.800.634	267.837
Your	10.708.661	342.020

- (2) 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 frequências mais salientes no corpus de estudo tendem a se 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 1, a diferença será grande (10) e possivelmente significativa, ou seja, a palavra 'casa' tem chances de ser chave. Mas se os 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, um 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 percentagens, e é portanto desnecessário unir-se os corpora (caso o corpus de referência seja 5 vezes maior no mínimo; vide discussão abaixo acerca do valor crítico de 5). Tomando-se o exemplo acima, 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 1 seria equivalente a 0.2% (1/500). Juntando-se os

corpora, a frequência no corpus de referência passa a ser de 11, ou 1.8% (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 chavidade. Por exemplo, se em vez de 10, ‘casa’ tiver frequência de 1 no corpus de estudo, as percentagens 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 chavidade da palavra ‘casa’.

- (3) A composição do corpus de referência influencia os tipos de palavra que podem se tornar chave. Um corpus de características genéricas semelhantes ao corpus de estudo tende a ‘filtrar’, ou seja, eliminar, os elementos genéricos (i.e. relativos a um mesmo gênero) em comum, resultando em uma lista de palavras chave que não inclui estes elementos. Alguns traços linguísticos que podem vir a ser filtrados são, entre outros, marcadores discursivos privilegiados, escolhas lexicais típicas, e formas verbais flexionadas em comum. Por exemplo, se se comparar um corpus de estudo de artigos de pesquisa acadêmicos de medicina com um corpus de referência do mesmo tipo, pode se esperar que palavras como ‘resultados’, ‘análise’, ‘sugerem’ não se tornem chave. Já um corpus de referência de um gênero distinto do de estudo tende a não excluir tais palavras ‘genéricas’. Por isso, um corpus de referência geral, que inclua vários gêneros, é tida como a escolha não-marcada para estudos de palavras chave.

Para se ter uma idéia do tipo de filtragem que pode vir a ocorrer nas palavras chave, pode-se utilizar um quadro semelhante ao mostrado abaixo. Na coluna 2, deve-se colocar as características referentes ao corpus de estudo, um por linha. Faz-se o mesmo com o corpus de referência, na coluna 3. As características que *coincidirem*, entre as duas colunas, são passíveis promoverem *filtragem* entre as palavras chave. Por outro lado, as características que *diferirem*, entre os dois corpora, tendem a se *manter* entre as palavras chave *na forma que aparecem no corpus de estudo*. Ou seja, a característica referente ao corpus de estudo se mantém. O quadro a seguir ilustra uma situação hipotética de comparação de dois corpora:

1	2	3	4
Característica	Corpus de estudo	Corpus de referência	Resultado
Modo	Falado	Escrito	Traço retido: Linguagem falada
Gênero	Aulas de inglês	Jornais	Traço retido: Gênero aula de inglês
Assunto	Vários	Vários	Traço filtrado: Assuntos variados em comum
Período	Contemporâneo	Contemporâneo	Traço filtrado: Tópicos em comum relativos ao cotidiano

- (4) 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 estas dimensões retornam

significativamente mais palavras chave do que corpora de tamanhos menores. Um corpus de referência que é o dobro do tamanho do corpus de estudo retornou cerca de 7% das palavras (do corpus de estudo) como chave; com um corpus de referência que é o triplo, 9%; e com um corpus de referência que é o quintuplo, 14% das palavras do corpus de estudo eram chave.

A quantidade de palavras chave num corpus de estudo pode ser prevista matematicamente. A fórmula abaixo permite estimar-se de antemão, de modo estatisticamente significativo, a quantidade de palavras chave obtidas quando se sabe o tamanho dos corpora de estudo e referência empregados na análise.

$$\text{Palavras chave} = 95,890582 + 0,011432 \times (\text{itens do corpus estudo}) + 0,009217 \times (\text{formas do corpus de referência}) - 0,000105 \times (\text{itens do corpus de referência})$$

Equação 1: Fórmula para previsão do total de palavras chave

- (5) O tamanho recomendado de um corpus de referência é cinco vezes o tamanho do corpus de estudo. Um pesquisador não necessita, necessariamente, coletar ou procurar um corpus de referência maior do que este valor, pois a quantidade de palavras chave a serem obtidas seria igualável a quantidades obtidas com corpora maiores.

4.2.2. Comandos principais

Os comandos necessários para se conseguir uma lista de palavras chave são:

- (1) No Controller (a primeira janela que aparece ao se inicializar o programa), clique em Tools e depois em KeyWords;
- (2) Na janela do KeyWords, clique em 'Settings' e depois em 'Min/Max Frequencies'.
- (3) Na janela 'WordSmith Tools Settings', faça os ajustes que julgar necessários e clique em OK;
- (4) Volte à janela 'KeyWords', clique em File e depois em Start;
- (5) Na janela 'Choose WordLists', selecione a pasta e depois o arquivo que contém o corpus de estudo, na janela à esquerda ('WordList'). Na janela à direita, selecione a lista que contém o corpus de referência. Clique em OK.

4.2.3. Lista de palavras chave

A execução dos comandos acima traz à tela uma lista de palavras chave em uma janela separada. As informações constantes na lista, da esquerda para a direita, são:

- Coluna 'Word': os itens do(s) texto(s);
- Coluna 'Freq': a frequência do item no corpus de estudo;
- Coluna <nome do arquivo> %: a percentagem do item em relação ao total do corpus de estudo;
- Coluna 'Freq': a frequência do item no corpus de referência;
- Coluna <nome do arquivo> %: a percentagem do item em relação ao total do corpus de referência;

- Keynes: o resultado da estatística de comparação (log-likelihood ou qui-quadrado);
- P: o valor da significância estatística atingido pelo resultado da estatística.

4.2.4. Palavras chave chave

O programa KeyWords também proporciona a possibilidade da contagem de quantas vezes algumas palavras foram chave em várias listas. Palavras que foram chave em um número determinado de listas são chamadas de palavras chave chave (key key words).

Os passos seguidos pelo programa KeyWords para a obtenção de palavras chave chave são, resumidamente, os seguintes:

- (1) Abrir um conjunto de listas de palavras chave;
- (2) Identificar as palavras chave em comum entre a primeira lista e a segunda;
- (3) Se as palavras em comum tiverem uma frequência igual ou superior ao valor mínimo estabelecido pelo usuário, copie-as para uma lista de palavras chave chave;
- (4) Repetir os passos 2 e 3 acima até ter comparado cada listas com todas as outras.

4.2.4.1. Comandos

Para se obter uma lista de palavras chave chave, deve-se seguir os seguintes comandos:

- (1) No Controller (a primeira janela que aparece ao se inicializar o programa), clique em Tools e depois em KeyWords;
- (2) Na janela do KeyWords, clique em 'Settings' e depois em 'Min/Max Frequencies'.
- (3) Na janela 'WordSmith Tools Settings', faça os ajustes que julgar necessários e clique em OK;
- (4) Volte à janela 'KeyWords', clique em 'File' e depois em 'New Database';
- (6) Na janela 'Choose WordLists', no quadro à esquerda ('WordList'), selecione a pasta e depois os arquivos que contém os corpora de estudo. No quadro à direita, selecione a lista que contém o corpus de referência. Clique em OK.
- (5) Na janela 'Batch of Files', digite o nome da pasta onde será salvo o arquivo 'database' no quadro 'directory', e selecione a opção 'store in a database'. Clique em OK.
- (6) Quando a janela 'Files Created' se tornar ativa, veja o nome do arquivo criado e clique em OK.
- (7) Volte à janela 'KeyWords', clique em 'File' e depois em 'Open Database';
- (8) Na janela 'Open 1 KeyWords Database File' clique na pasta escolhida no passo (5) acima, e depois em no nome do arquivo surgido durante a execução do passo (6) acima. As palavras listadas são as palavras chave chave.

4.3. Concord

Essa ferramenta produz concordâncias, ou listagens das ocorrências de um item específico (chamado palavra de busca ou nóculo, que pode ser formado por uma ou mais palavras) acompanhado do texto ao seu redor (o co-texto). 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 chave⁷ no contexto', na

⁷ O termo 'key word' é usado na literatura juntamente com 'search word' (ou ainda 'search term', 'node', 'node word') para se referir à palavra central da concordância. Prefere-se 'search word' e sua tradução

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 é uma peça chave na investigação de corpora. No WordSmith Tools, pode-se usar o Concord em separado, fazendo-se concordâncias avulsas, ou pode-se usá-lo em conjunto com as ferramentas WordList e KeyWords, chamando-o a partir desses programas. Para tanto, basta selecionar um item de uma lista de palavras ou de palavras chave e clicar no botão ‘C’ na barra de tarefas do WordList ou KeyWords que 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. É importante notar que se 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.

4.3.1. Comandos principais

Os comandos necessários para se produzir uma concordância são:

- (1) No Controller (a primeira janela que aparece ao se inicializar o programa), clique em Tools e depois em Concord.
- (2) Na janela do Concord, clique em ‘Settings’, ‘Choose Texts’, e selecione os arquivos a partir dos quais deseja fazer a concordância. Clique em OK.
- (3) Clique em ‘Settings’, ‘Horizons’, faça os ajustes desejados, e clique em OK.
- (4) Volte ao Concord, clique em ‘Settings’, ‘Search Word’, e digite a palavra de busca, na caixa ‘Search Word or Phrase’. Para buscar palavras com variações, use um asterisco; por exemplo, para capturar ‘casa’, ‘casar’, ‘casou’, etc., digite ‘casa*’. Se quiser fazer a concordância já, clique em ‘Go Now’; se quiser fazer mais alguma definição, clique em OK.
- (5) Opcionalmente, se quiser encontrar a palavra de busca somente em companhia de outra, digite essa ‘palavra de contexto’ na caixa ‘Context Word’. Nas caixas ‘Context Search Horizon’ escolha a distância máxima, em relação à palavra de busca, que a palavra de contexto pode estar. Por exemplo, se quiser encontrar as formas ‘casa da sogra’ e suas variantes ‘casa da minha sogra’ e ‘casa lá da minha sogra’, mas não ‘sogra em casa’, digite ‘sogra’ na caixa ‘Context Word’ e escolha 0L e 4R como ‘Context Search Horizon’. 0L significa que o programa deve ignorar as ocorrências de ‘sogra’ à esquerda de ‘casa’, mas deve aceitar todas as ocorrências de ‘sogra’ até quatro palavras à direita de ‘casa’ (para permitir ‘casa lá (1R) da (2R) minha (3R) sogra (4R)’). Clique em ‘Go Now’ para fazer a concordância, ou em OK para efetuar mais algum ajuste.

4.3.2. Concordância

A tela da concordância possui os seguintes elementos:

- (1) Coluna N: O número sequencial da linha da concordância.
- (2) Coluna Concordance: A concordância em si.
- (3) Coluna Set: Espaço reservado para o analista inserir códigos de classificação das linhas da concordância. Os códigos podem ser as letras do alfabeto (maiúsculas ou

‘palavra de busca’ porque ‘key word’ possui um sentido especializado no WordSmith Tools, e portanto é preferível reservar ‘palavras chave’ para se falar das palavras identificadas pelo programa KeyWords.

minúsculas).

- (4) Coluna Tag: A etiqueta mais próxima do termo de busca. É útil para ajudar a ver, por exemplo, em qual parte do texto a palavra ocorreu, ou por qual falante foi proferida. Se o texto for de uma conversa, na transcrição pode-se digitar a identificação dos falantes com códigos especiais como [João], [Maria], etc. Essa opção do Concord só funciona se a função de reconhecimento de etiquetas tiver ativa. Para ativá-la, crie um arquivo texto com as etiquetas (e.g. [João], [Maria], etc.), salve-o, clique em 'Settings', 'Tags', e na caixa 'Tags to be included' digite o nome do arquivo. Depois selecione a opção 'Activated'. Se a opção 'Tags to ignore' estiver com 'Activated' selecionado, certifique-se de que o formato das etiquetas que aparece na caixa nesse espaço não é o mesmo dos códigos de falante. Se em 'Tags to Ignore' aparecer '[*]', o Concord irá ignorar as etiquetas de falantes especificadas, pois elas correspondem ao formato '[*]'. Por fim, clique em OK.
- (5) Coluna 'Word No.': o número correspondente à posição sequencial da palavra no corpus de estudo selecionado.
- (6) Coluna 'File': o arquivo no qual a palavra aparece.
- (7) Coluna '%': a posição sequencial da palavra em termos de porcentagem do total do corpus de estudo selecionado.

4.3.3. Lista de colocados

A lista de colocados é mostrada clicando-se no botão 'Show Collocates' na barra de tarefas. A janela dos colocados possui os seguintes elementos principais:

- (1) Coluna Word: O colocado em questão.
- (2) Coluna Total: O total de ocorrências do colocado ao redor da palavra de busca. É a soma das colunas 'Left', 'Right' e '*' (vide abaixo).
- (3) Coluna Left: O total de ocorrências do colocado à esquerda da palavra de busca (colunas L5⁸ a L1).
- (4) Coluna Right: O total de ocorrências do colocado à direita da palavra de busca (soma das colunas R1 a R5⁹).
- (5) Coluna L5: O total de ocorrências do colocado na posição referente a cinco palavras à esquerda da palavra de busca.
- (6) Colunas L4 a L1: Semelhante a L5, mas refere-se às posições referentes a quatro, três, duas, e uma palavra à esquerda da palavra de busca.
- (7) Coluna '*': O total de ocorrências do colocado enquanto termo de busca. Só é útil se houver mais de uma palavra de busca na mesma concordância.
- (8) Colunas R1 a R5: Semelhantes a L1 a L5, descritas acima, só que se referem às posições relativas à direita da palavra de busca.

5. O que é cada instrumento

Nessa seção é feita uma apresentação breve dos instrumentos disponíveis em cada ferramenta:

- WordList
 1. Lista de palavras individuais ('wordlist'): Lista contendo todas as palavras do arquivo ou arquivos selecionados, elencadas em conjunto com suas frequências

⁸ Caso não se tenha alterado as posições relativas pré-definidas no menu 'Settings', 'Horizons'.

⁹ Idem à nota 8.

absolutas e percentuais. É apresentada em duas versões: ordenada por frequência e por ordem alfabética.

2. Lista de múlti-palavras ('wordlist, clusters activated'): Semelhante à lista descrita acima, só que possui itens compostos, isto é, conjuntos de palavras, em vez de palavras individuais.
 3. Lista de palavras de consistência individuais ('detailed consistency'): Lista que combina duas ou mais listas de palavras, com palavras idênticas colocadas lado a lado. A lista mostra ainda o total de arquivos em que cada item aparece, e as frequências de cada item nas listas originais.
 4. Lista de múlti-palavras de consistência ('detailed consistency, clusters activated'): Idêntica a anterior, só que cada item corresponde a palavras compostas.
 5. Lista de dimensões do corpus e densidade lexical ('statistics'): Lista de estatísticas descritivas dos arquivos selecionados. Apresenta várias contagens relativas aos textos, tais como o tamanho em itens ('tokens') e formas ('types'), a densidade lexical simples e em intervalos (isto é, a razão entre itens e formas, para o texto todo e reiniciada em intervalos regulares; vide discussão acima), e a quantidade de parágrafos e sentenças.
- **Concord**
 1. Concordância ('concordance'): Lista contendo uma palavra específica (chamada de palavra de busca ou nóculo) juntamente com parte do texto ao seu redor (o co-texto).
 2. Lista de colocados ('collocates'): Lista de palavras que ocorrem ao redor da palavra de busca, em posições determinadas. A posição da primeira palavra à direita da palavra de busca é representada no programa por R1 ('Right 1', ou uma à direita), a segunda por R2, a terceira por R3, etc., até R5 (quinta posição à direita). O mesmo esquema é aplicado à esquerda: L1 ('Left 1') para primeira palavra à esquerda, L2 para a segunda, etc. até L5 (quinta palavra à esquerda). A coluna central da listagem refere-se à(s) palavra(s) de busca.
 3. Lista de agrupamentos lexicais ('clusters'): Lista de sequências fixas de palavras recorrentes na concordância. Ou seja, são múlti-palavras extraídas da concordância. Em geral são múlti-palavras que incluem a palavra de busca, mas podem não ser, pois o programa busca os itens recorrentes na concordância, sem se limitar a trechos nos quais aparece a palavra de busca.
 4. Lista de padrões de colocados ('patterns'): Lista de resumo dos colocados. Eles são agrupados nas posições em que são mais frequentes. Deve-se ter cuidado para não se ler a listagem como se os itens formassem múlti-palavras, pois não há garantia de que houve os encadeamentos sequenciais aparentes na lista.
 5. Gráfico de distribuição da palavra de busca ('plot'): Gráfico no qual as ocorrências da palavra de busca são marcadas por um pequeno traço ('|'), desenhadas em um retângulo que representa o texto (ou corpus). Um traço pode representar mais do que uma ocorrência, pois a pouca definição da tela do computador não permite que se reserve um espaço exclusivo para cada palavra.
 - **KeyWords**
 1. Lista de palavras chave ('keywords'): Lista extraída da lista original de palavras do corpus de estudo, produzida a partir da comparação do corpus de estudo com um de referência. As palavras chave são de dois tipos: positivas e negativas. As palavras chave positivas são aquelas cujas frequências são estatisticamente superiores no corpus de estudo em relação ao de referência. As negativas são aquelas cujas frequências são estatisticamente menores no corpus de estudo. As

- positivas aparecem no topo da lista, e as negativas, no fim, em vermelho.
2. Banco de dados de listas de palavras chave ('database'): Um arquivo especial que reúne várias listas de palavras chave. É um pré-requisito para a obtenção de palavras chave.
 3. Lista de palavras chave chave ('key keywords'): Lista de palavras chave que aparecem em um número de arquivos determinado. Indica, portanto, as palavras chave compartilhadas por dois ou mais arquivos.
 4. Lista de palavras chave associadas ('associates'): Lista de palavras chave que ocorrem nos textos nos quais uma certa palavra chave aparece.
 5. Lista de agrupamentos textuais ('clumps'): Lista textos que compartilham palavras chave, e nos quais ocorre uma ou mais palavra chave associada.
 6. Gráfico de distribuição de palavra chave ('keyword plot'): Semelhante ao gráfico de distribuição da palavra de busca da ferramenta 'Concord'. As ocorrências das palavras chave são indicadas por uma pequena marca ('|'), ao longo de um retângulo representando o texto (ou corpus). Às vezes o traço representa mais do que uma ocorrência de uma mesma palavra chave.
 7. Listagem de elos entre palavras chave ('keyword plot links'): Listagem do número de vezes em que duas ou mais palavras chave ocorrem dentro do horizonte colocacional definido pelo usuário.

6. Aplicações

Há várias aplicações no estudo da linguagem para o programa WordSmith Tools. Abaixo são apresentadas algumas questões comuns no estudo da linguagem que podem ser operacionalizadas por meio do uso dos recursos disponíveis no programa.

6.1. Como se caracteriza um gênero específico?

Pode-se tratar dessa questão de dois modos. No primeiro, o que se pergunta é o que há de mais distintivo, caracterizador ou típico no gênero em questão. Basicamente, o que se deve fazer é buscar saber como esse gênero difere de outros. Para tanto, exige-se um contraste dos dados do corpus de estudo com o de um corpus de referência. Os procedimentos principais seriam:

- (a) Feitura de uma lista de palavras para o corpus de estudo.
- (b) Feitura de uma lista de palavras para o corpus de referência, caso não haja.
- (c) Feitura de um recorte na lista de palavras do corpus de estudo, por meio da comparação do corpus de estudo com o de referência, extraindo-se assim as palavras chave.

Os instrumentos necessários para cada etapa acima seriam:

- (a), (b) Lista de palavras
- (c) Lista de palavras chave

A segunda maneira pela qual se pode tentar responder a essa questão é perguntando-se quais seriam os itens lexicais recorrentes, independente de serem chave. O instrumento principal, nesse caso, não seriam as palavras chave mas as palavras de consistência. Os procedimentos seriam os seguintes:

- (a) Feitura de uma lista de palavras para cada texto do corpus de estudo.

(b) Identificação dos itens recorrentes em um número determinado de textos.

Os instrumentos aplicados seriam:

- (a) Lista de palavras, com possível opção de feitura por lote ('batch').
- (b) Lista detalhada de palavras de consistência.

6.2. Qual o estilo de um autor específico ou período histórico?

Para se estudar essa questão são necessários os mesmos procedimentos da questão acima, já que se busca descobrir quais os elementos mais típicos do autor ou do período histórico por meio de um contraste.

6.3. Quais as inovações vocabulares de um período de tempo determinado?

Em termos simples, nessa questão busca-se saber quais os acréscimos que ocorreram de um período de tempo anterior para outro posterior. Para tanto, há duas opções, que variam de acordo com o conceito de 'acrécimo' que se emprega.

Se qualquer diferença entre frequências for tida como acréscimo, então pode-se usar os procedimentos a seguir:

- (a) Feitura de uma lista de palavras para o corpus de estudo referente ao período anterior;
- (b) Feitura de uma lista de palavras para o corpus de estudo referente ao período posterior;
- (c) Descoberta dos itens que possuem frequência zero no corpus anterior e mais que zero no corpus posterior.

E os instrumentos empregados seriam:

- (a), (b) Lista de palavras
- (c) Lista detalhada de palavras de consistência, com aplicação da função de classificação por frequência.

Por outro lado, se o conceito de 'acrécimo' implicar em uma diferença estatística significativa, então deve-se recorrer a palavras chave, e portanto os procedimentos seriam:

- (a) Feitura de uma lista de palavras para o corpus de estudo referente ao período anterior;
- (b) Feitura de uma lista de palavras para o corpus de estudo referente ao período posterior;
- (c) Descoberta dos itens acrescidos de modo estatisticamente significativo, ou seja, que são chave no corpus posterior. Para tanto, deve-se escolher o corpus posterior como sendo o de estudo, e o anterior como o corpus de referência.

Seriam empregados os seguintes instrumentos em cada etapa:

- (a), (b) Lista de palavras
- (c) Lista de palavras chave

6.4. Que agrupamentos de textos existem?

Nessa questão está embutida a identificação de textos diferentes que possuem itens lexicais compartilhados. Os procedimentos seriam:

- (a) Feitura de uma lista de palavras para cada texto do corpus de estudo.
- (b) Feitura de uma lista de palavras para o corpus de referência, caso não haja.
- (c) Feitura de um recorte na lista de palavras de cada texto do corpus de estudo, por meio da comparação do corpus de estudo com o de referência, extraindo-se assim as palavras chave de cada texto.
- (d) Identificação de itens compartilhados entre os textos, por meio da localização de palavras chave, ou seja, palavras chave que ocorrem em um número crítico de textos.
- (e) Identificação dos textos que compartilhem palavras chave.

Os instrumentos necessários para perfazer essas etapas são:

- (a), (b) Lista de palavras
- (c) Lista de palavras chave
- (d) Banco de dados de palavras chave
- (e) Lista de palavras chave associadas e lista de agrupamentos textuais ('clumps')

6.5. Quais os usos típicos de uma palavra determinada?

Aqui busca-se saber se há e quais são os padrões de uso mais típicos de um item lexical. Assume-se que o item lexical seja conhecido de antemão. Os procedimentos seriam os seguintes:

- (a) Feitura de uma concordância tendo-se o item lexical em questão como palavra de busca, para se observar o co-texto típico no qual está inserido o item.
- (b) Feitura de uma listagem de agrupamentos lexicais, a fim de se examinar os padrões fixos ou fraseologias do qual o item faz parte (note que em geral é necessário descartar os agrupamentos que não incluem a palavra de busca).
- (c) Feitura de uma listagem de colocados, para se observar os padrões colocacionais.

Os instrumentos empregados seriam os seguintes:

- (a) Concordância.
- (b) Listagem de agrupamentos lexicais ('clusters').
- (c) Lista de colocados e, opcionalmente, lista de padrões de colocados ('patterns').

6.6. Em quantas partes se divide o texto?

O objetivo dessa questão é descobrir quais os segmentos do texto e se há uma correspondência entre os segmentos e a ocorrência de itens lexicais. Para se saber se há correspondência deve-se examinar uma representação gráfica da localização das palavras chave no texto. É importante frisar que o arquivo de onde se fez a lista de palavras chave deve corresponder a um texto, e não a um conjunto de textos, senão não é possível descobrir-se as divisões internas do texto em si, mas sim do corpus (o que é mais difícil de justificar).

Para tanto deve-se executar os seguintes procedimentos:

- (a) Feitura de uma lista de palavras para o texto, ou para cada um dos textos.
- (b) Feitura de uma lista de palavras para o corpus de referência, caso não haja.
- (c) Feitura de um recorte na lista de palavras de cada texto do corpus de estudo, por meio da comparação do corpus de estudo com o de referência, extraindo-se assim as palavras chave de cada texto.
- (d) Feitura de um gráfico de distribuição das palavras chave no texto.

Os instrumentos exigidos são os seguintes:

- (a), (b) Lista de palavras
- (c) Lista de palavras chave
- (d) Gráfico de distribuição de palavra chave, com aplicação da função de classificação por ordem de primeira ocorrência

6.7. Qual a temática recorrente?

Essa questão enfoca mais diretamente o conteúdo do corpus de estudo, mais especificamente quais seriam os itens indicativos de um conteúdo recorrente por vários textos, corpora, ou períodos históricos. Há duas alternativas de procedimentos. Na primeira, trata-se essa questão como um problema de identificação de grupos de textos com elementos *chave* compartilhados, nesse caso a temática. Por isso, as etapas a serem percorridas nesse caso não diferem daquelas discutidas acima em 6.4. Na segunda alternativa, vê-se essa questão como um problema de localização de itens *recorrentes*. Essa opção seria realizada seguindo-se os procedimentos abaixo:

- (c) Feitura de uma lista de palavras para cada texto do corpus de estudo.
- (d) Identificação dos itens recorrentes em um número determinado de textos.

Os instrumentos necessários seriam:

- (c) Lista de palavras, com possível opção de feitura por lote ('batch').
- (d) Lista detalhada de palavras de consistência.

7. Comentários finais

O objetivo do trabalho apresentado aqui foi o de apresentar um panorama geral do programa WordSmith Tools. A apresentação visou a familiarização do leitor com os comandos relativos às principais ferramentas disponibilizadas pelo *software*. Espera-se, desse modo, ter-se contribuído para tornar o programa mais acessível a uma gama maior de lingüistas de várias orientações.

Referências

BERBER SARDINHA, A. P. (2000) Representatividade de corpus. *DIRECT Papers*, **45**.

BIBER, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

---. (1990) Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, **5**: 257-269.

---. (1993) Representativeness in corpus design. *Literary and Linguistic Computing*, **8**: 243-257.

BIBER, D. ET AL (1998). *Corpus linguistics - Investigating language structure and use*. Cambridge: Cambridge University Press.

HOEY, M. (1993). Introduction. IN: M. HOEY (org.). *Data, Description, Discourse -- Papers on the English Language in Honour of John McH Sinclair on his Sixtieth Birthday*. London: HarperCollins.

PHILLIPS, M. (1989). *Lexical Structure of Text* (Discourse analysis monographs: 12). Birmingham: ELR, University of Birmingham.

STUBBS, M. (1996). *Text and Corpus Analysis -- Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.