6
Instrumentos

In all fields of education, the computer has come to stay.[100]

(de Smedt, 2002, p. 89)

Se há pouco tempo, mais especificamente 23 anos, discutia-se a implementação de computadores na área de humanas (Andersen, 1984), esta fase está superada dado o impacto que os mesmos têm nesta área atualmente e, de forma mais geral, em várias esferas da atividade humana. A importância de ferramentas no cotidiano é ressaltada por Scott (1998, p. 12):

> Ferramentas são necessárias em quase todo esforço humano, desde a fabricação de cerâmica até a previsão do tempo. Ferramentas computacionais são úteis porque elas permitem que certas ações sejam realizadas facilmente, e esta facilidade significa que se torna possível realizar trabalhos mais complexos. Torna-se possível ter *insigths* porque quando você pode testar uma idéia mais rápida e facilmente, você pode experimentar e da experimentação surge o *insight*. Além disto, remodelar um conjunto de dados em uma nova forma permite que o ser humano identifique padrões.[101]

Assim sendo, parece que as ferramentas são inerentes a muitas das atividades humanas. No âmbito deste estudo, as ferramentas computacionais são especialmente importantes por realizarem tarefas para as quais os seres humanos não são talhados, especialmente tarefas laboriosas e repetitivas como de busca e contagem.

---

[100] Tradução livre para o português: "Em todos os campos da educação, o computador veio para ficar".

[101] Tradução livre do seguinte fragmento: "Tools are needed in almost every human endeavour, from making pottery to predicting the weather. Computer tools are useful because they enable certain actions to be performed easily, and this facility means that it becomes possible to do more complex jobs. It becomes possible to gain insights because when you can try an idea out quickly and easily, you can experiment, and from experimentation comes insight. Also, re-casting a set of data in a new form enables the human being to spot patterns".

A presente seção enfoca as três ferramentas computacionais que foram utilizadas na etapa de processamento de dados desta pesquisa, a saber, *WordSmith Tools* (Scott, 1999), SPSS e *Excel*. De forma resumida, a primeira é útil na exploração automática de corpora. O SPSS permite a realização de testes estatísticos de forma a verificar se as diferenças encontradas nas populações investigadas podem ser generalizadas para contextos maiores. Já o *Excel* é empregado como um facilitador para o cálculo da significância das colocações. O enfoque privilegiado nesta parte da dissertação é a descrição dos aspectos relevantes destas ferramentas para a pesquisa em tela.

6.1. *WordSmith Tools*

O *WordSmith Tools* é um programa desenvolvido por Mike Scott e comercializado pela *Oxford University Press*, estando em sua quinta versão atualmente. Ele reúne ferramentas para verificação de como as palavras ocorrem em textos ou corpora. O programa é empregado em pesquisas lexicográficas realizadas pela editora que o comercializa e em análises diversas desenvolvidas por lingüistas, professores e estudantes de línguas em escala mundial (Scott, 1998, p. 8).

Este programa é considerado uma ferramenta e não um produto. Enquanto um produto tem uma finalidade específica (por exemplo, um jogo), uma ferramenta transcende seus limites, possibilitando que resultados diversos sejam extraídos com sua ajuda. Às vezes, o alcance da ferramenta não pode ser previsto por aqueles que a desenvolveram (Scott, 1998, p. 13).

Para investigar textos com o auxílio do programa, é preciso que estes estejam em formato digital, preferencialmente TXT. Assim sendo, um primeiro momento de pré-processamento de textos, faz-se necessário. Nos arquivos TXT a serem investigados, os usuários podem optar por utilizar cabeçalhos ou outros tipos de marcação desde que os mesmos ocorram entre os sinais de menor e maior (< e >). Por definição padrão, o programa ignora automaticamente tudo que estiver entre estes sinais. Outra possibilidade é de marcar os textos com etiquetas específicas (como em páginas HTML) que indiquem o início e o fim da mesma, por exemplo, <cabeçalho> e </cabeçalho>. Neste caso, porém, é preciso

configurar o programa para não considerar tudo o que estiver no meio destas duas etiquetas.

Tentativas de investigações anteriores realizadas com o *WordSmith Tools* (Scott, 1999) indicam também que outros cuidados precisam ser tomados para que os textos sejam processados adequadamente. A contagem do número de frases em um texto é afetada caso não haja um espaço após as mesmas, especialmente no tocante àquelas que terminam um parágrafo. Por este motivo, antes de inserir uma quebra de linha, é necessário haver um espaço ao final da frase para que ela possa entrar nos dados estatísticos gerados pelo programa. O número de parágrafos também pode ser afetado por um descuido no pré-processamento de textos. Caso não haja uma linha em branco entre dois parágrafos, o resultado final fica comprometido. No caso do uso de barras ('/'), não é preciso inserir um espaço antes e depois da mesma porque o programa identifica automaticamente que '*his*/*her*', por exemplo, corresponde a duas palavras. No entanto, o uso das aspas é algo que gera problemas na contagem final de palavras e na forma como o programa reconhece as mesmas. É possível indicar que as aspas simples são caracteres válidos dentro de uma palavra. Desta forma, '*can't*' seria considerada uma palavra única. No entanto, o procedimento não é tão simples se os textos foram trabalhados em um processador de texto como o *Word*, que altera automaticamente uma aspa simples reta (') para uma aspa simples curvada (’). Esta substituição não ocorre no caso do Bloco de Notas, por exemplo. O *WordSmith Tools* (Scott, 1999) só reconhece a aspa reta, o que gera a necessidade de todas as aspas serem trocadas antes do início da etapa analítica dos dados. Este procedimento, no entanto, pode ser realizado com o auxílio do utilitário *Text Converter*.

Em suma, o que programa faz é processar textos e apresentar os mesmos dados neles contidos em uma forma distinta da linear que os caracteriza. Segundo Scott & Tribble (2006, p. 12), esta mudança de formato é transformacional visto que abre todo um novo leque de possibilidades para exploração textual. "Esta simples transformação serve para focar a atenção do leitor não na mensagem dos

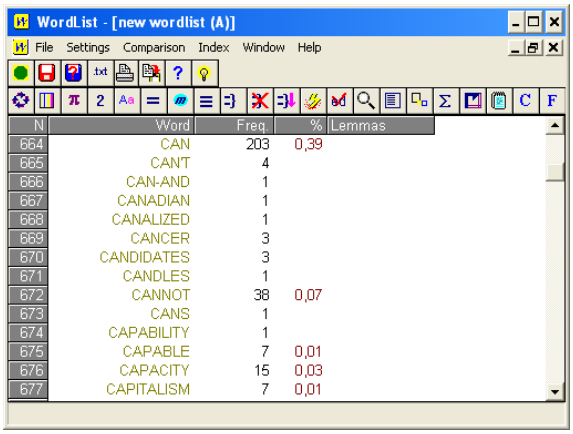
textos originais, mas na forma ou em outros aspectos das palavras individuais neles contidas"[102] (Scott & Tribble, 2006, p. 12).

O *WordSmith Tools* (Scott, 1999) já se encontra em sua versão 5 atualmente, mas, nesta pesquisa, optou-se por empregar a versão 3 devido à familiaridade do pesquisador com a mesma. Esta versão específica oferece três ferramentas e quatro utilitários. As ferramentas englobam *Concord*, *WordList* e *KeyWords*. Os utilitários compreendem *File Manager*, *Splitter*, *Text Converter* e *Viewer & Aligner*, que servem para gerenciar arquivos, seccionar arquivos, converter textos ou seus atributos e visualizar e/ou alinhar arquivos, respectivamente. Por não serem empregados nesta pesquisa, os utilitários não serão descritos nas seções seguintes, que, por outro lado, enfocarão as três ferramentas principais do programa, mostrando como as mesmas foram utilizadas na presente pesquisa.

6.1.1. *WordList*

A ferramenta *WordList* do *WordSmith Tools* permite a geração de listas de todas as formas – ou seja, todas as palavras diferentes – que ocorrem em um único texto ou em grandes corpora até o limite de 8 milhões de entradas. O princípio que rege esta ferramenta é o da ocorrência visto que somente os itens – isto é, as palavras – presentes são listados. Como explicam Scott & Tribble (2006, p. 12-23), o que tal ferramenta faz é vasculhar os dados e reduzir todos os itens a uma lista quantificada de formas. Estas listas podem ser ordenadas alfabeticamente, como ilustrado na Figura 26.

[102] Tradução livre do seguinte fragmento: "This simple transformation serves to focus the reader's attention not on the message of the original texts but on the form or other aspects of the individual words in them".



N	Word	Freq.	%	Lemmas
664	CAN	203	0,39	
665	CANT	4		
666	CAN-AND	1		
667	CANADIAN	1		
668	CANALIZED	1		
669	CANCER	3		
670	CANDIDATES	3		
671	CANDLES	1		
672	CANNOT	38	0,07	
673	CANS	1		
674	CAPABILITY	1		
675	CAPABLE	7	0,01	
676	CAPACITY	15	0,03	
677	CAPITALISM	7	0,01	

Figura 26: Lista de palavras por ordem alfabética

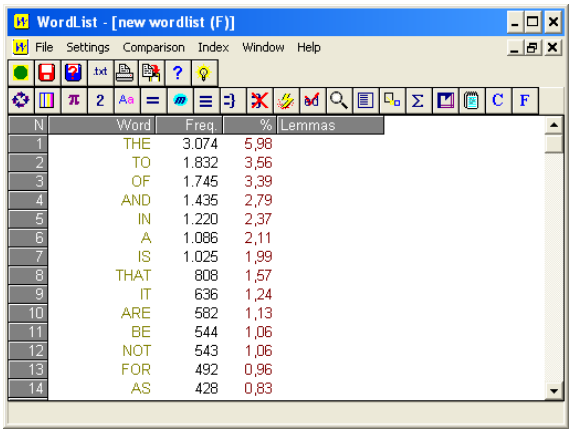
Este tipo de lista é especialmente interessante quando se deseja saber a freqüência de uma dada palavra no corpus. Desta forma, é mais fácil localizar a freqüência de, por exemplo, '*can*' com esta ordenação. A tela reproduzida na Figura 26 indica que '*can*' ocupa a posição 664 na lista alfabética, tendo totalizado 203 instâncias. Este número bruto representa 0,39% dos itens no Br-ICLE. No entanto, não se pode saber a partir desta informação a que classe gramatical a palavra em questão pertence (substantivo ou verbo).[103] Como Bowker e Pearson (2002, p. 113-114) afirmam, "listas de freqüência não discriminam entre palavras que têm a mesma forma, mas pertencem a categorias gramaticais diferentes, e pode ser difícil de identificar todas as instâncias de ambigüidade categorial, como este fenômeno é chamado".[104]

É a partir da lista de freqüência alfabética que se pode mais facilmente realizar a lematização das entradas de forma manual, como ocorreu neste estudo. O procedimento consiste em reunir todas as variantes de uma mesma forma canônica. Em termos práticos, é a lematização que possibilita a verificação da freqüência total de, por exemplo, '*will*', combinando a freqüência desta forma com a de '*'ll*' e '*won't*'.

Outro tipo de lista contempla as mesmas formas, mas em ordem de freqüência das mesmas no corpus investigado (cf. Figura 27).

[103] Isto poderia ser contornado caso o corpus tivesse sido etiquetado. Porém, como se trata de dados de alunos, contendo desvios ortográficos e gramaticais, a eficácia de etiquetadores automáticos fica comprometida.

[104] Tradução livre do seguinte fragmento: "Frequency lists do not discriminate between words that have the same form but belong to different grammatical categories, and it can be difficult to spot all instances of categorial ambiguity, as this phenomenon is called".

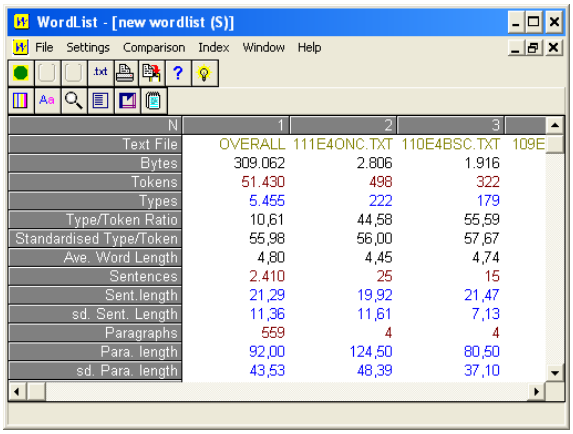


N	Word	Freq.	%	Lemmas
1	THE	3.074	5,98	
2	TO	1.832	3,56	
3	OF	1.745	3,39	
4	AND	1.435	2,79	
5	IN	1.220	2,37	
6	A	1.086	2,11	
7	IS	1.025	1,99	
8	THAT	808	1,57	
9	IT	636	1,24	
10	ARE	582	1,13	
11	BE	544	1,06	
12	NOT	543	1,06	
13	FOR	492	0,96	
14	AS	428	0,83	

Figura 27: Lista de palavras por ordem de freqüência

A Figura 27 é igual à Figura 26 no que concerne às informações oferecidas, somente a disposição das formas é alterada. A forma mais freqüente no Br-ICLE é o artigo definido '*the*'. O mesmo ocorre 3.074 vezes, o que representa 5,98% dos itens no corpus. Isto já era esperado já que as palavras mais freqüentes são geralmente as funcionais, as quais são referidas como "colas textuais" (Scott & Tribble, 2006, p. 15).

Por fim, a ferramenta *WordList* também fornece estatísticas a respeito do corpus investigado como indicado na Figura 28.



N	1	2	3	
Text File	OVERALL	111E4ONC.TXT	110E4BSC.TXT	109E
Bytes	309.062	2.806	1.916	
Tokens	51.430	498	322	
Types	5.455	222	179	
Type/Token Ratio	10,61	44,58	55,59	
Standardised Type/Token	55,98	56,00	57,67	
Ave. Word Length	4,80	4,45	4,74	
Sentences	2.410	25	15	
Sent.length	21,29	19,92	21,47	
sd. Sent. Length	11,36	11,61	7,13	
Paragraphs	559	4	4	
Para. length	92,00	124,50	80,50	
sd. Para. length	43,53	48,39	37,10	

Figura 28: Tela de estatísticas do *WordList*

Entre os dados fornecidos, podem ser encontrados o número de itens (*tokens*) e de formas (*types*), a razão forma/item regular e padronizada (*type/token ratio* e *standardised type/token ratio*), o tamanho médio de palavras, o número de frases, o tamanho regular e padronizado destas, o número de parágrafos, o

tamanho regular e padronizado destes, quantidade de cabeçalhos e quantidade de palavras com 1 até 37 letras. As estatísticas para o Br-ICLE e para o LOCNESS podem ser encontradas nos Anexos 10.4 e 10.9, respectivamente. Nos Anexos 10.5-10.8, 10.10 e 10.11, pode-se verificar as estatísticas por universidade (no caso do Br-ICLE) e por nacionalidade (no caso do LOCNESS).

Ressalta-se aqui o cálculo realizado para a razão forma/item, uma medida de riqueza lexical comumente empregada em estudos baseados em corpora. O resultado final é obtido a partir da divisão do número de formas pelo número de itens, sendo o resultado multiplicado por 100 para que possa ser expresso em termos percentuais. Como se sabe, esta razão é sensível ao tamanho do texto ou do corpus uma vez que quanto maior o mesmo, menor será o valor expresso por esta razão. Por isso, emprega-se a razão forma/item padronizada, que consiste basicamente no mesmo cálculo, mas este é realizado por intervalos no texto. O padrão é que ele seja realizado para grupos de 1.000 itens, mas o usuário pode diminuir este valor até 50 itens em intervalos pré-definidos pelo programa. Nas estatísticas apresentadas nos Anexos 10.4-10.11, optou-se pela utilização de grupos de 300 itens já que este era o limite mínimo para que uma redação não fosse descartada dos corpora. No entanto, a razão forma/item não foi empregada nas análises realizadas neste estudo por tratar-se de uma investigação de um corpus com dados de falantes de ILE no qual qualquer desvio foi preservado por ser característico da população em estudo. Todos os casos de erros ortográficos são considerados pelo computador como uma nova palavra, o que acaba gerando distorções neste resultado. A título de exemplificação, ilustra-se tal ponto com dois exemplos do Br-ICLE:

> In other words, the wealth of the world is concentrated on the small part of the globe and just fewer has the control of it whereas this fortune *shoul* be divided equaly through the world in an attempt to minimize the social inequalities and brings a decent quality of life for all peoples. (058F4jSn – enfâse minha)

> The world situations is about to enter in a colapse because of the gravity of the social problem, the masses can not support all this injustice anymore and it is necessary to review all this system ideologie in order to organize a equalitarian society where, at least, the basics rights *shoud* be conserved for all nations. (058F4jSn – ênfase minha)

Nestes dois casos, o mesmo participante opta por escrever a palavra '*should*' de duas formas distintas: '*shoul*' e '*shoud*', o que leva o programa a reconhecer as

mesmas como formas distintas. Portanto, um alto valor na razão forma/item (seja regular ou padronizada) em corpus de aprendiz não necessariamente indica um emprego lexical rico, mas pode indicar que seja comum a criação de palavras inexistentes (cf. Granger, 2002).

6.1.2. *KeyWords*

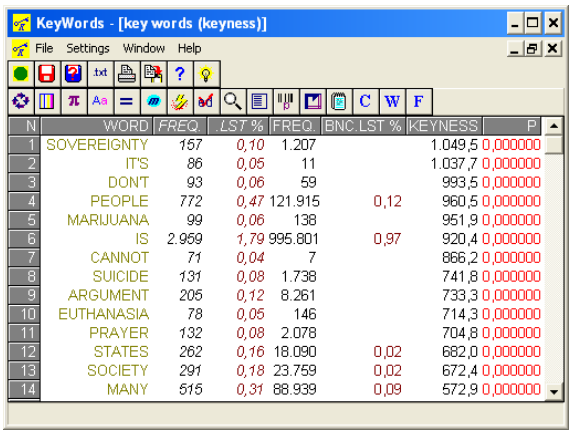
A ferramenta *KeyWords*, descrita por Berber Sardinha (1999b, p. 1) como uma das características mais proveitosas do programa, permite a comparação de duas listas de palavras (geradas pelo próprio programa), representativas de dois corpora distintos (um de estudo e outro de referência). Objetiva-se com esta ferramenta a identificação das formas que são características de cada um dos textos / corpora comparados. Necessariamente o corpus de referência precisa ser maior do que o de estudo para que as palavras-chave possam ser extraídas. Recomenda-se que os corpora de referência sejam duas, três ou cinco vezes maior do que os de estudo porque tais tamanhos geram um número maior de palavras-chave (Berber Sardinha, 2004, p. 102; para mais detalhes a respeito do cálculo realizado, ver Berber Sardinha, 1999a).

O conceito de chavicidade empregado pelo *KeyWords* implica "uma qualidade [que] as palavras podem ter em um dado texto ou conjunto de textos, sugerindo que elas são importantes, [que] elas refletem sobre o que o texto é, evitando ninharias e detalhes insignificantes"[105] (Scott & Tribble, 2006, p. 55-56). Não se trata, portanto, do conceito corriqueiro de palavra importante, sendo baseado em freqüência (Berber Sardinha, 1999b, p. 2). A ferramenta, então, fornece um meio de caracterizar um texto ou um gênero.

A partir da comparação dos corpora, o *WordSmith Tools* (Scott, 1999) faz um levantamento das formas que são utilizadas recorrentemente no corpus de estudo, mas não aparecem com semelhante freqüência no corpus de referência. Estas palavras são denominadas de palavras-chave positivas, sendo características do corpus menor. As palavras-chave negativas, por sua vez, são aquelas freqüentes no corpus de referência, mas que não aparecem como esperado no

corpus de estudo. Além do princípio da ocorrência (explicado na Seção 6.1.1), o princípio da recorrência é igualmente importante aqui visto que o padrão desta ferramenta é só considerar as palavras que ocorreram no mínimo três vezes em uma das listas comparadas. O usuário, no entanto, pode alterar esta configuração.

Ao contrastar o LOCNESS com um corpus geral como o BNC, como indicado no Capítulo 1, podem-se levantar as palavras que caracterizam o gênero de redação produzida por universitários, falantes de IL1. O resultado de tal comparação é indicado na Figura 29.



N	WORD	FREQ.	LST %	FREQ.	BNC LST %	KEYNESS	P
1	SOVEREIGNTY	157	0,10	1.207		1.049,5	0,000000
2	IT'S	86	0,05	11		1.037,7	0,000000
3	DON'T	93	0,06	59		993,5	0,000000
4	PEOPLE	772	0,47	121.915	0,12	960,5	0,000000
5	MARIJUANA	99	0,06	138		951,9	0,000000
6	IS	2.959	1,79	995.801	0,97	920,4	0,000000
7	CANNOT	71	0,04	7		866,2	0,000000
8	SUICIDE	131	0,08	1.738		741,8	0,000000
9	ARGUMENT	205	0,12	8.261		733,3	0,000000
10	EUTHANASIA	78	0,05	146		714,3	0,000000
11	PRAYER	132	0,08	2.078		704,8	0,000000
12	STATES	262	0,16	18.090	0,02	682,0	0,000000
13	SOCIETY	291	0,18	23.759	0,02	672,4	0,000000
14	MANY	515	0,31	88.939	0,09	572,9	0,000000

Figura 29: Palavras-chave positivas características do LOCNESS

A informação é apresentada pela ferramenta em sete colunas. A primeira identifica a palavra. A forma mais característica das redações do LOCNESS é '*sovereignty*', que ocorre 157 neste corpus, representando 0,10% de todos os itens. Esta mesma palavra ocorre 1.207 vezes no BNC, mas não tem porcentagem representativa dado o tamanho deste corpus geral. A chavicidade, valor que expressa a importância da relação observada nos dois corpora, é de 1.049,5 e o valor de *p*, que indica a significância estatística, é menor do que 0,0000001.

[105] Tradução livre do seguinte fragmento: "a quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is really about, avoiding trivia and insignificant detail".

6.1.3. *Concord*

Se, por um lado, tanto o *WordList* como o *KeyWords* fornecem informações de natureza quantitativa para o pesquisador, o *Concord* permite que o analista tenha à sua disposição informações qualitativas. Esta ferramenta, denominada de concordanciador, fornece linhas de concordância, ou seja, "listas [que] apresentam todas as ocorrências de uma palavra ou estrutura em uma base de dados, com uma pequena quantidade de contexto em cada lado"[106] (Conrad, 2002, p. 75).

Como esclarecem Tribble & Jones (1990, p. 7), a concordância referia-se originalmente (na Idade Média) a uma compilação manual de todas as palavras utilizadas em um determinado texto ou na coletânea de obras de um autor específico juntamente com os seus cotextos de ocorrência. A noção moderna de concordância é bastante semelhante a esta, mas a forma de realizar este levantamento sofreu uma grande modificação. Mesmo se forem comparados os diversos estágios da realização de concordâncias com o uso do computador, podem ser notadas diferenças sensíveis. A título de exemplificação, menciona-se a existência de uma seção a respeito do uso mouse no manual do *MicroConcord* (Murison-Bowie, 1993), publicado há 15 anos.

Para gerar linhas de concordância com o *Concord*, é preciso eleger uma palavra de busca (de até 80 caracteres) para que a ferramenta possa listar todas as instâncias na qual ela ocorre. Esta palavra de busca geralmente surge após a análise da lista de palavras ou da lista de palavras-chave já que as mesmas permitem identificar pontos que devem ser mais detidamente investigados. Nesta ferramenta, o princípio mais importante é o da co-ocorrência, ressaltam-se os itens que são utilizados na presença de outros.

No caso da Figura 30, definiu-se '*should*' como palavra de busca. Caso o usuário deseje buscar todas as ocorrências de '*should*' e '*shouldn't*', basta indicar '*should**' de forma que sejam listadas todas as ocorrências deste verbo modal tanto em sua forma afirmativa como negativa reduzida. O asterisco indica ao programa que se deve buscar toda e qualquer palavra que seja iniciada pela seqüência de seis letras definida. Este procedimento também tem suas

[106] Tradução livre do seguinte fragmento: "listings display all the occurrences of a word or structure in a database, with a small amount of context on each side".

desvantagens no sentido de incluir palavras irrelevantes para a pesquisa. Portanto, a utilização de um caractere coringa necessariamente implica a necessidade de verificar as linhas de concordância para que se possa certificar do que foi incluído nas mesmas. Outra possibilidade de busca, neste caso, é de escrever todas as formas possíveis separadas por barras: '*should/shouldn't*'.

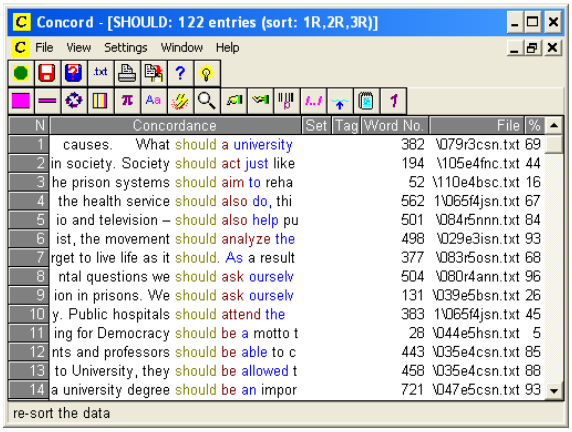


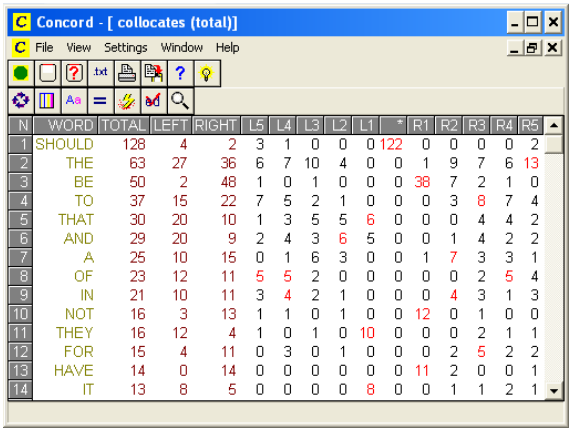
Figura 30: Linhas de concordância para '*should*' no Br-ICLE

No topo da tela do Concord, há a indicação de quantas linhas de concordância há para a palavra de busca. Ao todo, são 122 linhas sendo que somente as 14 primeiras encontram-se reproduzidas na Figura 30. Estas linhas podem ser classificadas com o uso de letras maiúsculas ou minúsculas conforme a conveniência do analista. Tais letras apareceriam na coluna '*set*' (cf. Viana, 2007 para mais detalhes a respeito deste procedimento em uma pesquisa lingüística de caráter semântico). A coluna '*tag*' reproduz a etiqueta mais próxima da palavra de busca. A próxima coluna informa o usuário a respeito do número desta palavra no texto ao qual ela pertence. A origem do fragmento é indicada na próxima coluna na qual se tem acesso ao nome do arquivo de onde a linha de concordância foi extraída. Na versão 3 do *WordSmith Tools* (Scott, 1999) é importante que este nome não ultrapasse oito letras. Caso contrário, ele não é adequadamente exibido nesta parte e a rápida identificação da fonte fica comprometida. Finalmente, há informação a respeito da parte do texto (em valores percentuais) no qual a palavra foi encontrada.

As linhas de concordância mostram a palavra de busca no meio da tela e em uma cor diferente de forma a possibilitar a identificação de padrões mais

facilmente. Este tipo de linha é conhecida como *key word in context* (KWIC). Adicionalmente as linhas da Figura 30 foram ordenadas em ordem alfabética das três palavras à direita da palavra de busca (1R, 2R e 3R). Tal procedimento facilita a identificação de padrões à direita de '*should*'. Outras ordenações também podem ser realizadas. O usuário tem à sua disposição vários critérios [1ª a 5ª palavra à esquerda / direita, palavra de busca, palavra no cotexto (caso a mesma tenha sido especificada quando da busca), classificação registrada na coluna '*set*', nome do arquivo, etiqueta (a que é exibida na coluna '*tag*') e a distância desta etiqueta], mas só pode escolher no máximo três destes. A reordenação das linhas de concordância através do uso de critérios variados permite que padrões léxico-gramaticais se tornem aparentes ao analista.

Como colocam Scott & Tribble (2006, p. 40), "para entender um grande número de linhas de concordância, precisa-se de algum tipo de guia [...]. A melhor estratégia neste caso é usar alguns dos recursos que o computador pode prover, envolvendo principalmente filtragem e ordenação de vários modos".[107] Com a utilização do *Concord*, o corpus pode ser investigado de outras formas. Pode-se, por exemplo, conseguir uma lista dos colocados da palavra de busca, isto é, as palavras que se encontram no entorno da mesma no horizonte da 5ª palavra à esquerda até a 5ª palavra à direita como ilustrado na Figura 31.



N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	SHOULD	128	4	2	3	1	0	0	0	122	0	0	0	0	2
2	THE	63	27	36	6	7	10	4	0	0	1	9	7	6	13
3	BE	50	2	48	1	0	1	0	0	0	38	7	2	1	0
4	TO	37	15	22	7	5	2	1	0	0	0	3	8	7	4
5	THAT	30	20	10	1	3	5	5	6	0	0	0	4	4	2
6	AND	29	20	9	2	4	3	6	5	0	0	1	4	2	2
7	A	25	10	15	0	1	6	3	0	0	1	7	3	3	1
8	OF	23	12	11	5	5	2	0	0	0	0	0	2	5	4
9	IN	21	10	11	3	4	2	1	0	0	0	4	3	1	3
10	NOT	16	3	13	1	1	0	1	0	0	12	0	1	0	0
11	THEY	16	12	4	1	0	1	0	10	0	0	0	2	1	1
12	FOR	15	4	11	0	3	0	1	0	0	0	2	5	2	2
13	HAVE	14	0	14	0	0	0	0	0	0	11	2	0	0	1
14	IT	13	8	5	0	0	0	0	8	0	0	1	1	2	1

Figura 31: Lista de colocados para '*should**' no Br-ICLE

[107] Tradução livre do seguinte fragmento: "To make sense of a huge number of concordance lines one needs some sort of guidance [...]. The best strategy in this case is to use some of the resources the computer can supply, chiefly involving filtering and sorting in various ways".

A primeira linha sempre indicará a palavra de busca; neste caso, '*should**'. A seguir são indicadas todas as palavras que ocorrem freqüentemente no ambiente lingüístico da mesma, ordenadas pela freqüência de co-ocorrência. Assim sendo, na 3ª posição, encontra-se o verbo '*be*', que é o verbo mais freqüentemente empregado juntamente com '*should*', totalizando 50 ocorrências. Destas, '*be*' aparece à esquerda de '*should*' duas vezes e à direita, 48 vezes. A posição mais recorrente de '*be*' (marcada em vermelho na Figura 31) é a R1, ou seja, uma palavra a direita de '*should*', gerando a seqüência '*should be*'.

Esta facilidade da ferramenta *Concord* não foi utilizada nesta dissertação quando do levantamento dos colocados dos nove verbos modais investigados. Optou-se pela verificação manual dos verbos que eram modalizados em cada uma das linhas de concordância. Este procedimento possibilitou observar que verbos lexicais eram efetivamente modalizados mesmo quando estes se localizavam em um horizonte maior do que cinco palavras à direita e à esquerda. Além disto, a análise manual tornou possível a desconsideração de eventuais verbos auxiliares como, por exemplo, '*have*' no aspecto perfeito, totalizando unicamente as ocorrências de '*have*' como verbo lexical.

6.2. SPSS

O programa computacional *Statistical Package for the Social Sciences* (SPSS) foi criado em 1969 na *Stanford University* por Niw, Hull e Bent. Originalmente, como indica o nome, o programa surgiu para dar conta do processamento de dados das ciências sociais. Em 1970, o programa tornou-se popular; porém, foi somente em 1980 que se lançou uma versão do SPSS para computadores pessoais (Griffith, 2007, p. 10).

O SPSS auxilia o analista na realização de estatísticas, que, por sua vez, são sempre baseadas em uma parcela da população investigada. O que os testes mostram é exatamente a possibilidade de generalizar os dados encontrados para uma maior fatia da população do que somente aquela selecionada para participar da coleta de dados. De outra forma, os testes estatísticos indicam ou não se as diferenças observadas devem-se ao fator acaso.

Um princípio básico da versão 15 do programa, utilizada nesta pesquisa, é que ela trabalha com números, não sendo possível solicitar que o mesmo processe palavras, por exemplo. Por esta razão, é preciso transformar as informações obtidas na análise de corpus em dados exclusivamente quantitativos para que testes estatísticos possam ser realizados.

O procedimento mais comum é que sejam informadas as freqüências de ocorrência para uma palavra, padrão ou categoria para o programa. Assim sendo, o SPSS se revela como um poderoso instrumento aliado à investigação de corpus já que permite que o analista investigue mais profundamente as diferenças encontradas entre corpora. Este procedimento é descrito por McEnery & Wilson (1996, p. 66-67):

> o uso de quantificação na lingüística de corpus tipicamente vai além da simples contagem: muitas técnicas estatísticas sofisticadas são usadas, que podem tanto prover uma análise matemática rigorosa de dados geralmente complexos – pode-se quase dizer, coloquialmente, trazer ordem ao caos – e ser usado para mostrar com alguns graus de certeza que diferenças entre textos, gêneros, linguagens, etc são reais e não simplesmente um acaso do procedimento de amostragem.[108]

A primeira tarefa na realização deste estudo estatístico foi a definição das variáveis do estudo. A variável independente – aquela "que é selecionada e sistematicamente manipulada pelo pesquisador para determinar se, ou o grau ao qual, ela tem algum efeito na variável dependente"[109] (Brown, 1988, p. 11) – foi representada pelo corpus (1 = Br-ICLE, 2 = LOCNESS). Já a variável dependente, isto é, aquela que é sensível a uma mudança da variável independente correspondeu a cada um dos verbos modais pesquisados.

Outra diferença entre as variáveis nesta pesquisa é que a independente comporta dados categoriais, que representam características (neste caso, a informação a respeito do corpus) enquanto a dependente relaciona-se a dados numéricos, que permitem contagem (aqui, a freqüência dos verbos modais).

---

[108] Tradução livre do seguinte fragmento: "the use of quantification in corpus linguistics typically goes well beyond simple counting: many sophisticated statistical techniques are used which can both provide a mathematically rigorous analysis of often complex data – one might almost say, colloquially, to bring order out of chaos – and be used to show with some degree of certainty that differences between texts, genres, languages, and so on, are real ones and not simply a fluke of the sampling procedure".

[109] Tradução livre do seguinte fragmento: "that is selected and systematically manipulated by the researcher to determine whether, or the degree to which, it has any effect on the dependent variable".

Segundo Rumsey (2003, p. 45), o primeiro tipo também é denominado de dados qualitativos ao passo que o segundo é nomeado de dados quantitativos ou mensuráveis.

Em relação às escalas empregadas, há dois tipos distintos nesta pesquisa. A variável independente lança mão de uma escala nominal, que permite dar conta de nomes e categorias. Já as variáveis dependentes, que se relacionam à contagem da freqüência dos verbos, empregam escalas intervalares nas quais o zero é definido e a diferença entre múltiplos é conhecida, sendo passível de mensuração.

Apesar de a contagem das freqüências ter sido realizada em termos brutos com o auxílio das ferramentas *WordList* e *Concord* do *WordSmith Tools*, os dados inseridos no SPSS foram relativos à freqüência normalizada do lema (cf. Seção 6.1.1) de cada verbo modal. Este cálculo permite que diferenças de tamanho entre textos ou corpora sejam levadas em consideração. Por exemplo, dois textos que contenham 5 ocorrências de uma palavra não necessariamente têm a mesma freqüência. Se um totaliza 50 itens e o outro, 500, torna-se evidente que é no primeiro que a palavra se faz presente de forma mais repetida. Assim sendo, o que a normalização faz é que os valores brutos observados sejam transformados para uma mesma base de forma que dois ou mais textos ou corpora de tamanhos distintos possam ser comparados.

Como os dois corpora investigados são de tamanhos distintos, decidiu-se normalizar os resultados obtidos por corpus para uma base comum de 100.000 itens. Este cálculo é feito multiplicando-se pela base comum (100.000) o resultado da divisão do número de ocorrências de um determinado verbo pelo tamanho do corpus (em itens).

No entanto, não foi possível trabalhar com o resultado geral obtido para cada corpus no SPSS. Houve a necessidade de verificar a ocorrência de cada verbo em cada uma das redações do corpus para que o programa pudesse rodar os testes estatísticos necessários ao presente estudo. Neste caso específico, optou-se por outra base para a realização da normalização. Uma vez que as normas de coleta do ICLE indicam que as redações devem variar entre 500 a 1.000 palavras, decidiu-se por normalizar as freqüências observadas em cada texto por 1.000.

Ressalta-se ainda que todo o procedimento de normalização foi realizado com o auxílio do programa *Excel* no qual os dados foram inicialmente

trabalhados. Em uma etapa seguinte, os resultados foram importados para o SPSS.

6.2.1. ANOVA

Há dois tipos principais de estatísticas: descritiva e inferencial. A primeira é útil na descrição de padrões já que resume todos os dados coletados. A segunda, diferentemente, permite que sejam verificadas diferenças significativas entre os grupos contrastados. Nesta pesquisa, foram empregadas estatísticas inferenciais para verificar de que forma os dados variavam a partir da influência da variável independente.

O teste ANOVA, um acrônimo para *Analysis of Variance*, foi empregado na presente dissertação para investigar as diferenças significativas no uso de verbos modais pelos dois grupos investigados – falantes brasileiros de ILE e falantes de IL1. Este teste analisa a variação existente tanto dentro dos grupos como entre os grupos analisados. O valor de *F*, reportado na tabela produzida pelo SPSS (cf. Anexos 10.14 e 10.15), é resultado da divisão do quadrado da média entre grupos pelo quadrado da média dentro dos grupos. Em outras palavras, este valor equivale à proporção da variação entre grupos em relação à observada dentro dos grupos (Brown, 1988, p. 171). Como explicam Van Peer, Hakemulder & Zyngier (2007), quanto maior o resultado de *F*, menor será a semelhança entre os grupos. Contudo, este valor é geralmente considerado como estatisticamente significativo se o valor de *p* for menor do que 0.05.

6.3. Excel

A planilha de dados *Excel*, parte integrante do conjunto de aplicativos conhecido como *Office* e comercializado pela Microsoft, foi também empregado nesta pesquisa para a realização dos cálculos estatísticos de associação lexical (cf. Berber Sardinha, 2004).

Os cálculos automatizados com o uso de *Excel* referem-se ao levantamento da informação mútua (MI) e do escore T (T). São os resultados destes testes que indicam na presente pesquisa se o uso combinado de um verbo modal com um

lexical constitui em somente um padrão de co-ocorrência ou em uma colocação (cf. Seção 2.2).

Em uma perspectiva prática, ao seguir o procedimento descrito por Berber Sardinha (2004), é possível configurar o *Excel* para que o mesmo realize os cálculos necessários. Desta forma, o pesquisador não precisa se preocupar com os detalhes e os refinamentos matemáticos e estatísticos, informando apenas alguns dados observados no corpus. Visto por outro prisma, tal facilidade permite que o lingüista concentre-se na análise dos resultados obtidos.

6.3.1. Informação mútua

A informação mútua, advinda da teoria da informação, mede a força colocacional entre duas palavras a partir dos valores freqüenciais observado e esperado no corpus para tal emprego conjunto (McEnery, Xiao & Tono, 2006). Quanto menor o resultado obtido, menos recorrente é a associação. Por outro lado, um alto valor pode indicar um padrão colocacional livre do fator acaso. Para que se considere este resultado como importante, ele precisa ser maior do que três (Hunston, 2002; Berber Sardinha, 2004).

Quando as minúcias da informação mútua são consideradas, tem-se que o seu cálculo é realizado a partir da seguinte fórmula: $\log_2 O/E$. Em outras palavras, a informação mútua é obtida a partir do logaritmo na base 2 do quociente resultante da divisão entre o valor observado e o esperado. O valor observado é obtido a partir da divisão da freqüência registrada para o nódulo e o colocado (em emprego conjunto) pelo número de itens no corpus. Por sua vez, para calcular o valor esperado, é preciso (a) dividir a freqüência do nódulo pelo tamanho do corpus, (b) dividir a freqüência do colocado pelo tamanho do corpus e (c) multiplicar os dois resultados anteriores (Berber Sardinha, 2004).

No entanto, uma crítica que é feita a informação mútua reside no fato de o mesmo desconsiderar a direção das palavras sob análise, podendo elas funcionar tanto como nódulo ou colocado.

6.3.2.
Escore T

O escore T é uma medida de associação de palavras que leva em consideração a direcionalidade das mesmas. Desta forma, este cálculo supre, de certo modo, uma limitação da informação mútua.

O cálculo é realizado através da seguinte fórmula: T = (O - E) / (raiz quadrada de f(n,c) / N). Neste caso, O corresponde ao valor observado da co-ocorrência, E relaciona-se ao valor esperado da co-ocorrência, f(n,c) é a freqüência do nódulo com o colocado e N indica o tamanho do corpus (Berber Sardinha, 2004, p. 205). Para que uma associação seja considerada, o valor do escore T deve ser, no mínimo, superior a dois. Em outras palavras, isto indicaria um emprego não-aleatório de itens.

De acordo com McEnery, Xiao & Tono (2006, p. 57), "as colocações com altos resultados de MI tendem a incluir palavras de baixa freqüência ao passo que aquelas com altos valores de escore T tendem a mostrar pares de alta freqüência".[110] Por este motivo, os pesquisadores – assim como Berber Sardinha (2004) – sugerem que estes dois cálculos sejam integrados na identificação de colocações.

6.4.
Resumo

Foi descrita nesta seção a utilização de três programas computacionais na etapa de processamento de dados da presente pesquisa. Procurou-se mostrar em detalhes como os programas e seus princípios foram empregados para a execução da investigação em tela. Como foi demonstrado, tanto o *WordSmith Tools* quanto o SPSS podem ser utilizados em investigações baseadas em corpora. Apesar de o segundo não estar necessariamente relacionado à análise de corpus, seu poder estatístico é bem-vindo a tal área já que permite o estabelecimento de até que ponto as diferenças observáveis em termos de freqüência podem ser consideradas como diferentes para a população investigada como um todo. Em outras palavras, o suporte estatístico fornece o embasamento necessário para conclusões mais

[110] Tradução livre do seguinte fragmento: "Collocations with high MI scores tend to include low-frequency words whereas those with high *t*-scores tend to show high-frequency pairs".

solidamente justificáveis na área de lingüística de corpus. Além disto, demonstrou-se como o *Excel* pode ser empregado no cálculo de associações colocacionais.

Ressalta-se também o caráter tecnológico da metodologia aqui empregada que prioriza a pesquisa baseada em ferramentas computacionais. O uso do computador no presente estudo auxilia no estabelecimento da objetividade, tornando a pesquisa replicável, que é um dos princípios básicos da investigação científica.

Por fim, destaca-se que a ferramenta computacional não substitui o analista humano, mas permite ajudar o mesmo na identificação de padrões, permitindo novos olhares sobre dados familiares (Viana, 2008). O trabalho de como interpretar estes achados fica a cargo do pesquisador, que certamente desempenha o papel mais importante no estudo.