Literatura y Lingüística
ISSN: 0716-5811
literaturalinguistica@ucsh.cl
Universidad Católica Silva Henríquez
Chile

Recski, Leonardo Juliano
Concordâncias, listas de palavras e palavras-chave: o que elas podem nos dizer sobre a linguagem?
Literatura y Lingüística, núm. 16, 2005, p. 0
Universidad Católica Silva Henríquez
Santiago, Chile

Concordâncias, listas de palavras e palavras-chave: o que elas podem nos dizer sobre a linguagem?

Leonardo Juliano Recski
brasileiro,
Universidade Federal de Santa Catarina

Resumo
Corpora armazenados eletronicamente são excelentes recursos para uma série de atividades. Este artigo descreve três métodos para a investigação da linguagem: concordâncias, listas de palavras e palavras-chave. Sugere-se que tais métodos estão ao alcance de aprendizes de línguas e literatura, de professores e de pesquisadores empregando computadores pessoais comuns. Concentrado-se especificamente nestes três métodos, espera-se que o artigo possa aguçar o apetite do crescente corpo de professores, aprendizes e pesquisadores com acesso a corpora para que possam mais autonomamente desvelar fatos sobre o funcionamento da linguagem em todas as suas variedades.

Palavras-chave: - lingüística de corpus - concordâncias- listas de palavras - palavras-chave

Abstract
Computerized corpora have proved to be excellent resources for a wide range of tasks. This article describes three methods for carrying out research into small corpora, namely, the use of concordances, wordlists and keywords. It is suggested that such methods are within the reach of language and literature learners, teachers or researchers working with an ordinary computer. It is hoped that, by concentrating specifically in three methods, the article may be able to whet the appetites of the growing body of teachers, learners and researchers with access to corpora to discover more for themselves about how languages work in all their variety.

Keywords: - corpus linguistics - concordance- wordlist - keywords

1. Introdução

Para que possam trabalhar de maneira eficiente com corpora pequenos ou grandes, aprendizes, professores de línguas ou pesquisadores precisam, acima de tudo, de fácil acesso a eles e de software adequados. Sem muita dificuldade, qualquer pessoa interessada em investigar diversos aspectos lingüísticos pode ligar o seu computador, sem sair de casa, e descobrir fatos interessantes e, até porque não dizer, surpreendentes, sobre a linguagem.

Este artigo descreve alguns dos possíveis métodos para a pesquisa de pequenos corpora. Sugere-se que tais métodos sejam compatíveis com

as necessidades de aprendizes, professores de línguas e/ou pesquisadores interessados em realizar pesquisa empregando computadores pessoais comuns.

O artigo está estruturado em três partes. A seção 2 discute o uso de concordâncias para localizar ocorrências de uma dada palavra ou frase em um corpus, bem como para examinar quais palavras tipicamente co-ocorrem com estas palavras ou frases.

A seção 3 descreve o emprego de listas de palavras. Listas de palavras podem ser geradas a partir de um texto ou de uma coleção de textos. Informação sobre a freqüência de certas palavras pode ser de grande importância para que possamos identificar as características de um dado texto ou gênero.

Finalmente, a seção 4 discute o uso de palavras-chave. O princípio básico deste tipo de análise é que se uma palavra for muito mais freqüente em um dado texto do que sua freqüência em um conjunto de textos empregados como referência, ela provavelmente constitui uma palavra-chave.

2. Empregando concordâncias na investigação de aspectos lingüísticos

Concordâncias são de grande utilidade haja vista que não existe outra forma de obtermos uma grande quantidade de exemplos de morfemas, palavras ou frases em seus contextos de uso. Dicionários oferecem boas informações sobre a pronuncia, etimologia, aspectos gramaticais, significado e no máximo dois ou três exemplos de cada significado de uma palavra. Gramáticas propõem-se a exemplificar e explicar, mas uma boa parte das palavras ou frases não são exemplificadas. Logo, tanto aprendizes e professores, quanto lexicógrafos, podem utilizar concordâncias para obterem inúmeros exemplos. Ao examinarmos estes exemplos, é possível descobrir não apenas que palavras tipicamente co-ocorrem com a palavra que estamos examinando, mas obter também uma noção de sua freqüência.

Channell (2000) emprega o termo "significado pragmático" para aqueles aspectos do significado que estão relacionados a como uma palavra ou frase é tipicamente utilizada, ao invés daqueles que são inerentes a própria palavra ou frase. A pesquisa de Channell está em consonância com pesquisas anteriores, como as de Stubbs (1996) e Sincalir (1991), que empregam o termo "prosódia semântica" para descrever o fato de uma dada palavra ou frase poder ocorrer com maior freqüência no contexto de outras palavras ou frases que são predominantemente

positivas ou negativas em sua orientação semântica. Uma concordância do adjetivo fat utilizando o British National Corpus Sampler demonstra que das 61 ocorrências encontradas em cerca de 2 milhões de palavras, 46 ocorrências (75%) apresentam uma conotação negativa (Figura 1).

Figura 1. Concordância do adjetivo fat no British National Corpus Sampler

1	bear in fact knocked Lambert down and he was so	FAT	he could not get up! Lambert came to Stamford ...
2	... are you? Doctor! Doctor! I tend to get	FAT	in certain places. What should I do?
3	... should be like I, I always thought I was	FAT	regardless of what weight. Yep. I was and it ...
4	... I have to live with. I think I'm	FAT	! The reality of anorexia for you was what, ...
5	... at one time. I see boys calling the girls	FAT	and it makes my hair stand on end! Erm ...
6	.. on continually and these poor girls are oh you're	FAT	! You've got a great big bottom! And ...
7	... don't want her you can have her she's too	FAT	for me." The humiliation made me shudder. ...
8	y proletariat of paint in allegorical restraint. Where	FAT	silk-hatted bosses strut and cower. Around the walls in
9	She was about forty years old and a little	FAT	. She looked afraid. &bquo;You wanted me, ...
10	... out of his seat. The passenger was a short	FAT	in a grey suit. He shouted angrily in ...
11	sked by her/his commander why he had become so	FAT	, she replied: 'TIS strong beer and tobacco ...
12	... it's not nice to say that a girl is	FAT	these days, but she was all ample proportioned we ...
13	... man. Right you'd better see the cashier big	FAT	chap with the glasses. No I think on that ...
14	... so quick. Yeah you wolfed yours down, you	FAT	pig. I was starving though. Mm. I ...
15	... like and she mentioned Cardiff but I've gone so	FAT	I don't wanna go! Oh don't ...
16	... , that's all fat. I haven't got	FAT	legs really, I mean I put stockings on and ...
17	... ever so nice, I said oh I'm too	FAT	for that she said ooh have you seen, she ...
18	robably seen his mother butchered ah see that big	FAT	body just then a pair of feet I thought it ...
19	... 's why I'm thin and slim and you're	FAT	and la,'la, la, la. You're ...
20	... of her and'they say oh she's a bit	FAT	What abo't Gemma you'donIthink she's ...

As linhas de concordância da Figura 1 evidenciam que o fato de uma pessoa ser obesa pode ser considerado ruim ou não atrativo (pelo menos na cultura britânica). Logo, fat é utilizado aqui para demonstrar como informações advindas de um corpus corroboram a intuição de que geralmente este adjetivo é empregado pejorativamente.

Suponhamos agora, que um aprendiz de inglês como língua estrangeira, ao escrever um artigo científico em inglês, esteja interessado em iniciar um parágrafo com "This paper", "This article" ou "This study". Este mesmo aprendiz se pergunta: que tipos de verbos são normalmente empregados após "This paper / article / study"? Não existe uma maneira fácil de sabermos que tipos de verbos tipicamente co-ocorrem com article, paper e study em uma certa língua, a não ser através da consulta de linhas de concordância ou de um dicionário compilado com base em concordâncias.

Assim, para responder a pergunta acima, utilizo parte do corpus de artigos científicos compilado por Ken Hyland . As palavras pesquisadas foram article / paper / study com o pronome this situado num contexto de até duas palavras à esquerda. Após obter as linhas de concordância, a lista foi editada para conter apenas as ocorrências onde paper, article e study constituíam o sujeito da oração. O resultado pode ser observado na Figura 2.

Figura 2. Concordância de article, paper e study derivada do corpus de Ken Hyland

1	stematically understanding this phenomenon.	THIS ARTICLE	addresses this dilemma, encouraging the adopti
2	plications of game theory to retailing strategy	THIS ARTICLE	attempts to let the games speak for themselves.
3	have been applied with the retailer in mind.	THIS ARTICLE	examines three retailing issues-changes in tradi
4	(sr) for solid angle. One of the thrusts of	THIS ARTICLE	is to urge the promotion of radian to the status
5	subjects perform on the tasks given to them.	THIS ARTICLE	subjects perform on the tasks given to them.
6	ate bending problems is given by Adachi et al.	THIS PAPER	concerns the Laplace-transformed BEM formul
7	re and semiconductor properties of the device.	THIS PAPER	describes a physics-based multicell electrother
8	material handling equipment; (5) Plant layout.	THIS PAPER	focuses on stages 1 and 2 in the design of a ne
9	k out ways and means of utilizing waste heat.	THIS PAPER	investigates waste heat utilization in a ceramic
10	at are derived from polymer precursors [7, 8].	THIS PAPER	reports on the fabrication of a uniform and den
11	e material context of the Persian Gulf conflict.	THIS STUDY	contributes to work concerned with understand

12	, 1995; Shanna, Durand, and Gur-Arie, 1981).	THIS STUDY	demonstrates how failing to capture relevant hi
13	f a police shooting of Blacks in South Africa.	THIS STUDY	illustrates the processes through which the pap
14	ard, American fast food in particular. In sum,	THIS STUDY	investigates the perceptions and attitudes of Ho
15	portant research issues in automated welding.	THIS STUDY	will focus on this issue. In our initial study [11

Percebe-se que os substantivos article / paper / study podem ser seguidos de verbos como examine, address, report, concern, demonstrate, describe, focus, constribute, investigate, etc. Dentre todos estes verbos, os mais comumente encontrados foram investigate (12), examine (8) e address (7). Com o propósito de verificar se escritores brasileiros empregam os mesmos tipos de verbos com os substantivos artigo, estudo e trabalho, um corpus de artigos científicos escritos em português foi compilado2. As linhas de concordância na Figura 3 demonstram que escritores brasileiros empregam muitos dos verbos empregados no corpus de inglês acadêmico.

Figura 3. Concordância de artigo, trabalho e estudo derivada do corpus de artigos científicos escritos em português

1	LA E NA ESCRITA1 Edair Gorski* Resumo	EST ARTIGO	trata da topicalidade como uma propriedade
2	SINO DE LÍNGUA PORTUGUESA Resumo	EST ARTIGO	apresenta algumas considersemões acerca do
3	co-construction; negotiation; dialogy. Resumo	EST ARTIGO	discute a interação em sala de aula situada so
4	r; teacher; reflection; transformation. Resumo	EST ARTIGO	tem como objetivo discutir o papel do(a) co
5	d spiral approaches; learning process. Resumo	EST ARTIGO	aponta para as vantagens da aplicação do mo
9	lugar da atuação individual de cada um deles.	ESTE TRABALHO	objetiva, assim, discutir (a) a compreensão d
11	z parte de suas vidas. 2. Pressupostos teóricos	ESTE TRABALHO	fundamenta-se no referencial de pesquisa da
12	urse in face to face interaction. 0. Introdução	ESTE TRABALHO	focaliza a interação em sala de aula entre 3 p
13	ve triggered linguistic change. 1. Introdução	ESTE TRABALHO	examina o problema da insemerminação do s
16	cente nos últimos anos, pelo menos, no Brasil.	ESTE TRABALHO	assenta-se em dois pressupostos teóricos. Pri
19	lizado principalmente no romance Mar Morto,	ESTE ESTUDO	considera o mar como um elemento que prov
21	interpretação abstraída do episódio abordado.	ESTE ESTUDO	situa-se no âmbito de um projeto mais abran
22	em quando esta não é entendida. 1. Objetivos	ESTE ESTUDO	tem como objetivo investigar o efeito de dife

23	pela humanidade. A partir desta perspectiva,	ESTE ESTUDO	busca perceber que práticas de leitura circun
24	"Práticas de Leitura e Formação do Leitor". 1	ESTE ESTUDO	é parte das atividades da Bolsa de Iniciação

3. Análises baseadas em listas de palavras

Listas de palavras fornecem um tipo de informação bastante diferente das concordâncias. Elas auxiliam o pesquisador a identificar palavras comuns em um corpus, informação esta, que pode ser útil, por exemplo, quando queremos determinar quais itens lexicais devemos ensinar e quais devemos ignorar.

A distribuição de palavras em um corpus pode assumir formas um tanto estranhas. Tipicamente, encontraremos um pequeno número de palavras com uma freqüência muito grande. O caso mais extremo é o do artigo the, que normalmente constitui cerca de 5% das palavras de qualquer corpus (Sinclair, 1991). Do outro lado da escala de freqüência existe um grande numero de palavras que ocorrem apenas uma vez (normalmente chamadas de hapax legomena).

Consideremos uma lista de palavras baseadas no Michigan Corpus of Academic Spoken English (MICASE)3 com cerca de 1,7 milhões de palavras (Tabela 1).

Tabela 1. As 30 palavras mais freqüentes no MICASE

N	Word	Freq.	%	N	Word	Freq.	%
1	THE	80.495	4,38	16	WE	17.040	0,93
2	AND	48.075	2,61	17	UH	16.712	0,91
3	THAT	47.534	2,59	18	T	15.416	0,84
4	YOU	45.879	2,50	19	WHAT	14.447	0,79
5	I	42.365	2,30	20	LIKE	14.123	0,77
6	OF	41.910	2,28	21	THEY	13.674	0,74
7	IT	39.020	2,12	22	HAVE	13.299	0,72
8	TO	37.466	2,04	23	BUT	11.950	0,65
9	A	35.390	1,92	24	KNOW	11.548	0,63
10	S	34.974	1,90	25	FOR	11.278	0,61
11	IN	27.169	1,48	26	BE	10.955	0,60
12	IS	27.166	1,48	27	THERE	10.894	0,59
13	THIS	20.205	1,10	28	OKAY	10.830	0,59
14	SO	20.179	1,10	29	ON	10.191	0,55
15	UM	17.819	0,97	30	YEAH	10.098	0,55

As 30 palavras mais freqüentes nesta lista representam aproximadamente 25% (mais de 410 mil palavras) do total de palavras do corpus. Note que a maioria dessas palavras são funcionais, i.e., com pouco conteúdo lexical, ou são itens tipicamente encontrados no discurso oral (e.g. um, uh, okay, yeah).

Os primeiros itens lexicais normalmente encontrados em uma lista de palavras são verbos como know (ranking 24 na lista), think (46), make (114) e o primeiro substantivo encontrado é people (81). Hapax legomena começam no ranking 21.730 (com a palavra aback) e terminam no ranking 35.458 (com a palavra zwitterions). Isto significa que 39% de toda a lista é composta por palavras que ocorrem apenas uma vez. Um resultado semelhante é obtido quando investigamos o Freiburg-Brown Corpus of American English (FROWN)4 com cerca de 1 milhão de palavras. O primeiro hapax legomena ocorre no ranking 28.751 e o último no ranking 51.274 (44% do corpus é constituído por hapax legomena). É interessante ressaltar que grande parte destes hapax legomena são substantivos próprios. Alguns exemplos extraídos do FROWN são: Marisa, Marlette, Maranos, Mardsen, Marta, Martyn, Masanori, Mathew, Mathilde, etc.

Isto significa que aproximadamente um terço de uma lista de palavras é constituída por palavras que ocorrem apenas uma vez e que aproximadamente 30 palavras muito freqüentes representam um quarto da freqüência total de um corpus.

As implicações dos fatos revelados acima são bastante sérias para o ensino: mesmo que aprendizes leiam milhões de palavras, muitas delas serão vistas apenas uma vez. Os substantivos próprios podem ser considerados irrelevantes para o ensino, mas existem muitos hapax legomena que não são substantivos próprios (e.g. deflate, defrost, defunct, defuse, degenerative, deletion, delineate, etc).

Outro exemplo do possível uso de listas de freqüência é a listagem de classes gramaticais mais comuns em um corpus. Para que isto seja possível é necessário que o corpus seja anotado gramaticalmente. Hoje em dia, este tipo de serviço já pode ser obtido via email sem nenhum custo para o pesquisador e/ou aprendiz (vide, por exemplo, o AMALGAM Tagger by email - http://www.comp.leeds.ac.uk/amalgam/amalgam/amalgtag3.html, ou o Birminghan's email tagging service - http://www.clg.bham.ac.uk/tagger/index.html).

Diferentemente de listas de freqüências originadas a partir de textos não anotados, que fornecem apenas as freqüências das palavras, listas de

freqüência de elementos gramaticais constituem um exercício que requer uma maior capacidade interpretativa, tanto do pesquisador, quanto dos aprendizes. A partir de tais listas é possível, por exemplo, responder a perguntas como: a) quais são os itens gramaticais mais freqüentes em um dado texto; e b) qual a diferença entre as listas de freqüência de itens gramaticais em dois (ou mais) tipos de gêneros textuais.

Para responder às perguntas acima utilizo dois corpora. O primeiro (cerca de 200 mil palavras) foi compilado a partir de textos extraídos da renomada revista de ciências New Scientist. O segundo, representativo do discurso oral (cerca de 200 mil palavras) foi compilado a partir de transcrições de fala de programas exibidos pela rede de televisão norte americana CNN. O software utilizado para anotar os corpora foi o TOSCA Tagger5. Os resultados obtidos são apresentados na Tabela 2.

Tabela 2. Lista de freqüência de itens gramaticais no corpora New Scientist (science) CNN (spoken)

SCIENCE		SPOKEN	
Noun	63.328	Noun	43.469
Verb	36.086	Verb	42.726
Preposition	28.202	Punctuation	29.012
Punctuation	25.649	Pronoun	27.197
Article	21.393	Preposition	19.251
Adjective	20.902	Adverb	15.821
Pronoun	14.848	Article	15.173
Adverb	13.213	Adjective	13.465
Conjunction	10.127	Conjunction	10.292
Numeral	5.242	Particle	3.022
Particle	2.909	Numeral	2.551
Tag	1.790	Existential there	763
Genitive marker	1.096	Genitive marker	692
Existential there	368	Miscelaneous	532

Existem diferenças entre as listas para os dois tipos de textos. Por exemplo, no corpus de inglês escrito científico encontramos um número bem maior de substantivos se comparado ao corpus de discurso oral. Uma possível razão para esta diferença pode estar relacionada ao fato de textos escritos serem, normalmente, lexicalmente mais elaborados do que textos provenientes do discurso oral. Fica intuitivamente claro que muitos textos escritos, tais como artigos científicos, são densamente permeados com informações, ao passo que textos representativos do discurso oral são mais complexos gramaticalmente (Halliday & Matthiessen, 2004). Existe uma explicação funcional para os dados revelados na Tabela 2. Geralmente, um texto escrito é mais longo e

apresenta menos repetições do que uma transcrição de fala. O texto escrito é permanente, cuidadosamente editado e escrutinado antes de ser publicado, ao invés de ser espontâneo e não planejado como a maioria das interações orais.

Possíveis aplicações pedagógicas para o tipo de listagem fornecida na Tabela 2 incluem: a) aprendizes de inglês, como língua estrangeira, podem empregar este tipo de informação para descobrirem características estilísticas pertinentes a diferentes tipos de textos; e b) estes aprendizes podem observar como a quantidade de itens gramaticais varia de acordo com diferentes tipos de textos.

4. Descobrindo palavras-chave em um texto

Palavras-chave são extremamente úteis para a identificação de um texto ou gênero. Um das ferramentas encontrados no software WordSmith Tools (Scott, 1996) é a KeyWords, que compara listas de palavras pré-existentes. Uma lista (geralmente obtida a partir de um corpus relativamente grande) serve como referência; a outra, é baseada no texto que queremos investigar. O propósito de tal análise é descobrir que palavras caracterizam o texto que estamos interessados em investigar.

Para ilustrar o procedimento descrito acima, emprego um corpus compilado a partir de transcrições de discursos e entrevistas concedidas pelo presidente norte americano George W. Bush relativos a guerra do Iraque (disponibilizadas no site oficial da Casa Branca – www.whitehouse.gov). O corpus é composto por 57 textos (124.758 palavras) coletados entre 14/01/2001 e 18/02/2004.

A Tabela 3 contém parte da lista de palavras-chave gerada pela ferramenta KeyWords a partir do corpus descrito acima. O procedimento empregado pela ferramenta KeyWords está baseado em corpus de referência. O corpus de referência utilizado para esta ilustração foi compilado a partir de três corpora contemporâneos (MICASE, Switchboard, BNC Sampler)6 representativos do discurso oral (totalizando aproximadamente 8 milhões de palavras).

Tabela 3. Lista de palavras-chave nos discursos/entrevistas de George W. Bush relativos a guerra do Iraque.

N	Word	Freq.	Bush %	Freq.	Reference %	Keyness	P
1	IRAQ	836	0,67	102		6.164,0	0,000000
2	IRAQI	412	0,33	26		3.156,2	0,000000
3	OUR	1.196	0,96	9.450	0,13	2.581,5	0,000000
4	BUSH	405	0,32	300		2.344,3	0,000000
5	SADDAM	314	0,25	78		2.166,3	0,000000
6	NATIONS	301	0,24	138		1.907,2	0,000000
7	WILL	1 009	0,81	9.754	0,14	1.852,6	0,000000
8	AMERICA	374	0,30	562		1.803,0	0,000000
9	FREEDOM	311	0,25	239		1.785,7	0,000000
10	WEAPONS	291	0,23	225		1.668,6	0,000000
11	TERROR	216	0,17	19		1.626,0	0,000000
12	TERRORISTS	209	0,17	9		1.625,7	0,000000
13	HUSSEIN	235	0,19	59		1.619,2	0,000000
14	UNITED	348	0,28	762	0,01	1.477,3	0,000000
15	PRESIDENT	308	0,25	690		1.296,3	0,000000
16	REGIME	202	0,16	95		1.274,4	0,000000
17	SECURITY	269	0,22	468		1.237,5	0,000000
18	WAR	336	0,27	1.088	0,02	1.215,4	0,000000
19	COALITION	161	0,13	54		1.069,3	0,000000
20	FREE	295	0,24	989	0,01	1.050,2	0,000000
21	COUNTRY	368	0,29	2.075	0,03	995,2	0,000000
22	WORLD	385	0,31	2.568	0,04	935,0	0,000000
23	WE	2.296	1,84	64.305	0,90	933,0	0,000000
24	IRAQIS	124	0,10	13		923,2	0,000000
25	THE	7.257	5,82	287.331	4,01	910,8	0,000000
26	PEACE	199	0,16	392		877,4	0,000000
27	MILITARY	219	0,18	603		849,6	0,000000
28	NATION	168	0,13	223		840,2	0,000000
29	TERRORIST	109	0,09	16		791,6	0,000000
30	STATES	251	0,20	1.120	0,02	775,5	0,000000
31	DESTRUCTIO N	118	0,09	48		761,9	0,000000
32	PEOPLE	844	0,68	15.563	0,22	752,8	0,000000

Basta vislumbrarmos a lista parcial de 32 palavras-chave disposta na Tabela 3 para termos uma boa idéia de alguns dos tópicos enfatizados pelo presidente Bush em seus discursos relacionados ao Iraque. Algumas das palavras que emergem na lista são empregadas pelo líder político norte americano para justificar a operação militar no Iraque. Por exemplo, a expressão weapons of mass destruction aparece 103 vezes como uma justificativa para a retaliação norte americana; outras justificativas são subsidiadas por palavras como free / freedom (para o mundo, mas principalmente para o povo iraquiano); terror / terrorist(s) e security.

É interessante ressaltar que uma das palavras-chave reveladas na Tabela 3 é o verbo auxiliar will (1009 ocorrências). Mas qual será a estratégia retórica do líder norte americano por traz deste emprego tão freqüente

deste modal? Pode-se argumentar que uma característica compartilhada por muitos políticos é a de fazer promessas e previsões acerca do que vai acontecer no futuro como resultado de suas ações. O presidente Bush parece ser um mais um adepto desde clube; ele geralmente enfatiza o que será feito, porém sem explicitar como será feito. Assim, especulativamente, sugiro que o discurso altamente modulado do presidente Bush, através do emprego recorrente do auxiliar will, funciona para posicionar seus ouvintes -como parte de uma dialética onde ele, retoricamente, os manipula para naturalizar seus próprios pontos de vista.

Consideremos agora, como um aprendiz de literatura brasileira pode se beneficiar do uso de palavras-chave. Para ilustrar isto, utilizarei a obra literária Dom Casmurro de Machado de Assis. O corpus de referência empregado para a comparação das listas de palavras foi o corpus do Núcleo Interinstitucional de Lingüística Computacional (NILC)7 com cerca de 32 milhões de palavras.

A lista das primeiras 35 palavras-chave de Dom Casmurro, disposta na Tabela 4, revela alguns dos personagens do romance: os protagonistas Capitu e seu amigo de infância e posteriormente marido Bentinho; o amigo seminarista de Bentinho Escobar, a egoísta, ciumenta e intrigante Prima Justina; Tio Cosme -irmão de dona Glória (mãe de Bentinho); Ezequiel -o filho de Capitu e Bentinho e Sancha -a companheira de escola de Capitu.

Tabela 4. Lista de palavras-chave em Dom Casmurro

N	Word	Freq	Casmurro %	Freq	NILC %	Keyness	P
1	CAPITU	337	0,52	39		3.842,5	0,000000
2	ME	493	0,75	13.848	0,05	1.760,2	0,000000
3	EU	533	0,82	23.006	0,08	1.488,8	0,000000
4	ERA	552	0,84	26.175	0,09	1.450,1	0,000000
5	NÃO	1.523	2,33	208.549	0,74	1.428,9	0,000000
6	MINHA	342	0,52	7.923	0,03	1.341,9	0,000000
7	COUSA	96	0,15	23		1.048,8	0,000000
8	LHE	241	0,37	4.499	0,02	1.041,9	0,000000
9	MÃE	228	0,35	3.936	0,01	1.018,9	0,000000
10	QUE	2.686	4,11	595.425	2,11	990,5	0,000000
11	ESCOBAR	110	0,17	389		810,9	0,000000
12	OLHOS	164	0,25	2.197		810,4	0,000000
13	MIM	162	0,25	2.694		735,1	0,000000
14	DOUS	61	0,09	4		710,5	0,000000
15	MAS	603	0,92	73.553	0,26	659,2	0,000000
16	PADRE	111	0,17	1.099		611,0	0,000000
17	JUSTINA	55	0,08	13		601,4	0,000000
18	NEM	243	0,37	13.414	0,05	573,0	0,000000
19	BENTINHO	56	0,09	55		526,3	0,000000
20	COSME	53	0,08	85		460,0	0,000000
21	PRIMA	63	0,10	234		459,0	0,000000
22	MEU	169	0,26	7.660	0,03	456,6	0,000000
23	MAMÃE	57	0,09	171		436,4	0,000000
24	SANCHA	37	0,06	2		433,4	0,000000
25	SEMINÁRIO	85	0,13	1.128		421,4	0,000000
26	IA	88	0,13	1.544		390,4	0,000000
27	EZEQUIEL	51	0,08	194		369,4	0,000000
28	DISSE-ME	34	0,05	19		343,7	0,000000
29	TUDO	188	0,29	14.647	0,05	334,3	0,000000
30	COMIGO	64	0,10	789		326,1	0,000000
31	E	2.189	3,35	631.579	2,24	319,3	0,000000
32	DELA	83	0,13	2.020		317,8	0,000000
33	DEPRESSA	42	0,06	148		309,9	0,000000
34	PÁDUA	42	0,06	149		309,4	0,000000
35	PODIA	73	0,11	1.479		304,3	0,000000

Através das descrições feitas dos personagens, percebe-se uma palavra-chave interessante na lista acima: olhos (ranking 12 na Tabela 4). Os olhos são muito bem explorados por Machado de Assis, como em "Olhos de cigana oblíqua e dissimulada", "olhos de ressaca", "olhos dorminhocos", "olhos redondos, que me acompanham para todos os lados". Na verdade, esses elementos físicos, muitas vezes, parecem destacar o estado interior; tem-se um retrato íntimo das personagens. Em "olhos redondos" percebe-se uma característica física, mas, logo após, verifica-se um importante traço psicológico: "...que me acompanham para todos os lados; que me observam, me estudam". Outros exemplos do emprego de olhos por Machado de Assis estão expostos na Figura 4 abaixo.

Figura 4. Concordância do substantivo olhos em Dom Casmurro

1	da do tempo, desciam-lhe pelas costas. Morena,	OLHOS	claros e grandes, nariz reto e comprido, tinha a boca .
2	Pádua não é de todo má. Capitu, apesar daqueles	OLHOS	que o Diabo lhe deu... Você já reparou nos olhos ...
3	ranto da moça redobrou tanto que senti os meus	OLHOS	molhados e fugi. Vim para perto de uma janela. Pobre
4	etexto para mirá-los mais de perto, com os meus	OLHOS	longos, constantes, enfiados neles, e a isto atribuo que
5	com uma espécie de vertigem, sem fala, os	OLHOS	escuros. Quando eles me clarearam vi que Capitu tinh
6	cosida às saias de minha mãe, não atendia aos	OLHOS	ansiosos que eu lhe mandava; também não parecia esc
7	que a voz lhe tremia, e pareceu-me que tinha os	OLHOS	úmidos. Disse-lhe que também sentia a nossa separaça
8	evantei, nem sei se iria. Capitu fitou-me unsl	OLHOS	tão ternos, e a posição os fazia tão súplices, que ...
9	E, após alguma reflexão, fitando em mim uns	OLHOS	murchos e teimosos, perguntou-me: --Conservou o m
10	Le alguma cousa, um soneto. A insônia, musa de	OLHOS	arregalados, não me deixou dormir uma longa hora o
11	a assim perto, me envolvia, me encarava com os	OLHOS	furados e escuros. Quanto mais andava aquela Rua do
12	velhos e moços, sedas e chitas, e provavelmente	OLHOS	feios e belos, mas eu não vi uns nem outros. ...
13	scimo que, para desnortear a justiça, os mesmos	OLHOS	matadores seriam olhos piedosos, e correriam a chorar
14	migo Escobar era um tanto metediço e tinha uns	OLHOS	policiais a que não escapava nada. --São os olhos dele
15	e não vi que essa menina travessa e já de	OLHOS	pensativos era a flor caprichosa de um fruto sadio e
16	aos cinco anos, um rapagão bonito, com os seus	OLHOS	claros, já inquietos, como se quisessem namorar todas
17	afetos que me explicou daquela	OLHOS	teimosos. Outros olhos me

	maneira os seus		procuravam também, não
18	era muito amiga de si. Quando tornou, trazia os	OLHOS	vermelhos; disse-nos que, ao mirar o filho dormindo,
19	na xícara, e comecei a mexer o café, os	OLHOS	vagos, a memória em Desdêmona inocente; o espetácu
20	meu favor, mas a fé velava com os seus grandes	OLHOS	ingênuos. Minha mãe faria, se pudesse, uma troca de

Basicamente, pode-se dizer que Dom Casmurro trata, através das lentes do narrador (Bentinho), da dúvida da traição. A história se perpetua em torno da tríade Bentinho - Capitu - Escobar. A realização de um amor antigo através de um filho; o qual, porém, pode não ser um filho legitimo. Isto parece ser o grande ponto da história; o duelo de conclusões dos leitores, pois são eles quem acabam por decidir qual o final, a conclusão.

Em face disto, alguns itens lingüísticos de particular interesse para aprendizes de literatura seriam itens interpessoais como eu, minha, mim, meu, me, (empregados ao longo de toda a história) bem como os nomes dos personagens. O diagrama exposto na Figura 5 nos auxiliará a compreender isto melhor.

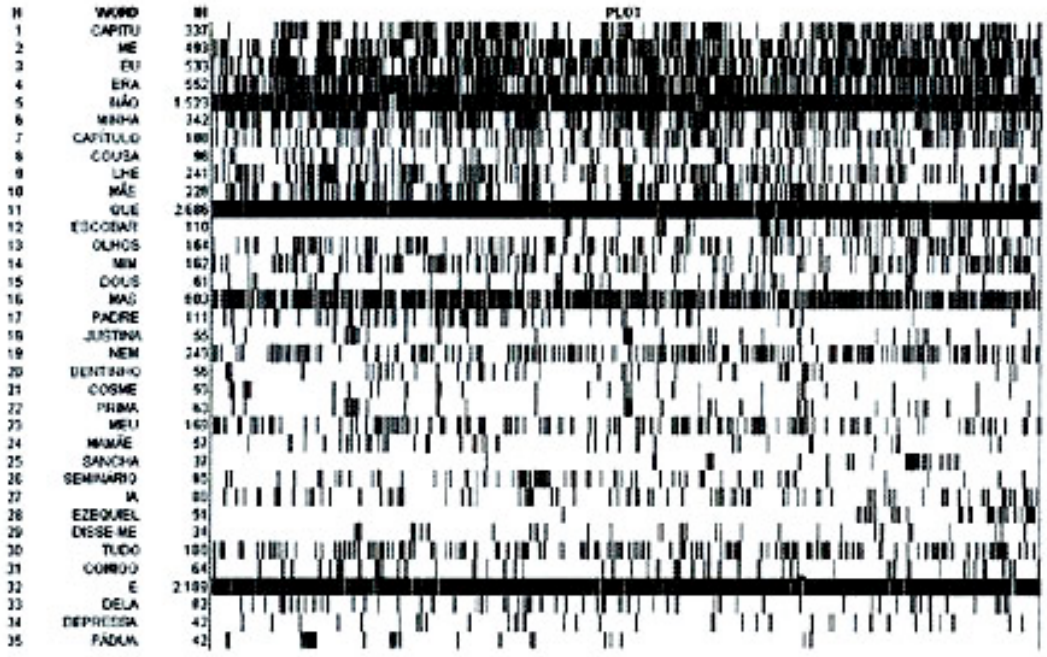

Figura 5. Diagrama das 35 primeiras palavras-chave em Dom Casmurro

A Figura 5 abaixo é um diagrama das palavras-chave, onde as margens esquerda e direita representam o início e o fim da história. Este diagrama foi ordenado para mostrar a primeira (margem esquerda) e última (margem direita) vez que cada palavra aparece no texto.

É possível notar que o personagem Escobar surge na história a partir de sua metade (quando Bentinho vai para o seminário) e que Ezequiel – o filho de Capitu e Bentinho – só aparece perto do final do livro (depois do casamento dos dois).

5. Conclusões

Este artigo procurou demonstrar como algumas ferramentas computacionais podem auxiliar aprendizes, professores e/ou pesquisadores a compreender melhor alguns aspectos pertinentes à linguagem humana.

Espero ter evidenciado que através de concordâncias seja possível vislumbrar e complementar alguns aspectos lingüísticos não disponíveis em dicionários e gramáticas convencionais. Listas de palavras nos permitem não apenas explorar uma série de itens lingüísticos que possam ser de grande utilidade para o ensino de vocabulário, mas também nos auxiliam a identificar itens peculiares a diferentes gêneros textuais. Palavras-chave podem nos auxiliar de diversas formas (recuperação de textos, investigações sobre questões relacionadas a plágio, estudos de estrutura textual, etc), muitas das quais estão bem aquém do que foi discutido aqui, onde os exemplos foram literários e estilísticos.

Finalmente, é necessário salientar que computadores e software lingüísticos são apenas ferramentas. Cabe a nós empregarmos nossa criatividade para decidir como utilizá-los. O fato de não ser possível predizer o que cada um de nós talvez encontre vasculhando um corpus, parece aguçar nossos poderes de observação, contribuindo assim, para uma forma de aprendizado mais autônoma, crítica e interessante.

Notas

1 O sub-corpus é formado por 15 textos representando oito disciplinas (biologia, engenharia, engenharia mecânica, marketing, lingüística, filosofia, sociologia e física) num total de 583.000 palavras.

2 Os periódicos utilizados foram: The ESPecialist, Intercâmbio, Revista de Letras, Revista do Gelne e Boletins da ALAB. O corpus possui cerca de 787 mil palavras.

3 Para maiores informações favor acessar http://www.lsa.umich.edu/eli/micase/micase.htm - MICASE

4 O FROWN corpus foi compilado para ser equivalente ao BROWN e LOB corpora com a diferença de conter amostras de inglês americano representativo do início da década de 90. Para maiores informações acesse o site http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM.

5 Para maiores informações favor consultar - http://lands.let.kun.nl/TSpublic/tosca/icle.html

6 Para obter maiores informações sobre estes corpora visite: http://www.isip.msstate.edu/projects/switchboard (Switchboard); http://www.comp.lancs.ac.uk/ucrel/bnc2sampler/sampler.htm - (BNC Sampler).

7 O corpus compilado pelo NILC, contém cerca de 35 milhões de palavras, e consiste de textos em prosa, divididos em textos corrigidos, textos não corrigidos e textos semi-corrigidos. Os textos classificados como corrigidos (corpus empregado como referência neste estudo), totalizando 32.590.000 palavras, são aqueles publicados para grande número de leitores (livros, jornais, revistas, etc). O corpus é composto por cerca de 4.300 textos de diversos gêneros: livros, revistas, a constituição brasileira e textos jurídicos e jornais.

Referências

Channel, Joanna, (2000). "Corpus-based analysis of evaluative lexis". In.: Hunston, S. e Thompson, G. (Eds.), Evaluation in Text: Authorial Stance and the Construction of Discourse, Oxford: Oxford University Press, p. 39-65.

Halliday, M. e Matthiessen, C., (2004). An Introduction to Functional Grammar (3rd Edition), London: Arnold.

Scott, Mike, (1996). WordSmith Tools, Oxford: Oxford University Press.

Sinclair, John, (1991). Corpus, Concordance, Collocation, Oxford, UK: Oxford University Press.

Stubbs, Michael, (1996). Text and Corpus Analysis, Cambridge, Massachusetts: Blackwell Publishers.