DOMÍNIOS DE LINGU@GEM
Revista Eletrônica de Lingüística (www.dominiosdelinguagem.org.br)
Ano 2, nº 1 – 1º Semestre de 2008 – ISSN **1980-5799**

A CONSTRUÇÃO E ANÁLISE DE *CORPORA* PARA ALIMENTAÇÃO DE UM BANCO DE DADOS TERMINOGRÁFICO: UM EXEMPLO

Guilherme Fromm[*]

Resumo: o presente artigo pretende demonstrar a criação de corpora técnicos bilíngües (português e inglês), com a finalidade de alimentar um banco de dados de caráter terminográfico. Para tanto, foram construídos dois corpora, bilíngües nas áreas de Informática e Lingüística, retirados da Internet, com aproximadamente um milhão de palavras cada um. Os corpora resultantes foram usados para levantamento de candidatos a termos nas duas áreas citadas (através do uso do software WordSmith Tools) e forneceram dados para a construção da microestrutura de verbetes técnicos, através de exemplos reais de uso de língua e dados morfo-sintático- semânticos.

Abstract: the following article intends to show the creation of bilingual (English and Portuguese) technical corpora aiming the feeding of a terminographical data bank. Two corpora were built, in Information Technology and Linguistics areas and both were taken from the Internet, with around one million words each. The resulting corpora were used to find term candidates in both areas (using the software WordSmith Tools) and they offer data to build the microstructure of technical dictionaries entries, which show real examples of language usage and grammatical data.

O que é um *corpus*?

Um *corpus*, segundo Tagnin (2004), é "[...] uma coletânea de textos em formato eletrônico, compilada segundo critérios específicos, considerada representativa de uma língua (ou da parte que se pretende estudar), destinada à pesquisa". Bidermann (2001, p. 79) coloca como *corpus* um conjunto homogêneo de amostras de língua de qualquer tipo que deve possibilitar, mediante análise lingüística, a ampliação do conhecimento das estruturas lingüísticas da língua que ele representa. A área da Lingüística que trata dos estudos sobre *corpora* (assim como de suas compilações), é a Lingüística de *Corpus*. Para Berber Sardinha,

> A Lingüística de *Corpus* ocupa-se da coleta e da exploração de *corpora*, ou conjuntos de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador. (2004, p. 3).

[*] Doutor em Língua Inglesa pela FFLCH/USP. Professor Adjunto da UNIBAN.

Ainda segundo Berber Sardinha (p. 20/21), quanto à tipologia, os *corpora* podem ser de diferentes:

a. modos: falados (transcrições) ou escritos;

b. tempos: sincrônicos ou diacrônicos; contemporâneos ou históricos;

c. seleções: por amostragem (estático, amostra finita da linguagem como um todo), monitor (dinâmico, reciclável), balanceado (textos distribuídos em quantidades semelhantes);

d. conteúdos: especializados (gêneros ou registros definidos), regionais ou dialetais, multilíngües;

e. autorias: de aprendiz (falantes não-nativos) ou de língua nativa (falantes nativos);

f. disposições internas: paralelos (original e tradução) e alinhados[1];

g. finalidades: de estudo (*corpus* a ser descrito), de referência (para contrastar com o *corpus* de estudo) e de treinamento (para desenvolvimento de aplicações e ferramentas de análise).

Os *corpora* construídos

A construção do *corpus* para a alimentação do banco de dados passou por várias fases. Pensou-se o uso de *corpora* bilíngües comparáveis já prontos, adotando o princípio da reusabilidade. Esses *corpora* seriam buscados num dos vários projetos do COMET: o CORTEC. Segundo o site do COMET (www.fflch.usp.br/dlm/comet), o CORTEC "[...]é um *corpus* comparável de textos técnicos e/ou científicos originalmente escritos em português brasileiro e em inglês." As áreas iniciais abrangidas pelo projeto são: Direito Contratual, Informática, Hipertensão Arterial, Culinária e Ecoturismo. O projeto prevê a inserção contínua de *corpora* em novas áreas e a complementação, também contínua, dos *corpora* já existentes.

Após conseguir os *corpora* completos do CORTEC, verificou-se que, para o projeto de levantamento da macroestrutura e construção da microestrutura de um dicionário técnico, os mesmos não apresentavam alguns aspectos essenciais:

1. não havia árvores ou mapas conceituais para todos os campos envolvidos, o que é essencial para verificar se todas as áreas foram contempladas na construção;

[1] Utiliza-se, neste artigo e nos trabalhos propostos pelo projeto COMET, a oposição entre corpora paralelos (textos originais e suas traduções) e comparáveis (textos equivalentes em línguas diferentes), diferente, portanto, dessa apresentada por Berber Sardinha.

2. o balanceamento desses *corpora* estava bastante irregular;
3. o planejamento original dos mesmos não previa um fim lexicográfico/terminográfico, de modo que nem sempre incluía textos que permitissem a construção de definições para o banco de dados;
4. o tamanho de cada *corpus*, de aproximadamente duzentas mil palavras, também não se mostrou suficiente para selecionar uma quantidade de termos em todas as áreas e/ou a possibilidade de criar suas respectivas definições.

Verificada a necessidade de novos *corpora*, partiu-se, em primeiro lugar, para a reconstrução do *corpus* de Informática (ou Computação). Embora já houvesse um *corpus* semelhante, organizado durante o mestrado de Fromm (2002), o mesmo era monolíngüe (português). Decidiu-se, então, pela ampliação dos *corpora* desenvolvidos para o CORTEC, aproveitando o que já havia sido levantado. A estruturação final desses *corpora* ficou assim delineada: escritos, sincrônicos, de amostragem (embora exista a possibilidade de se transformarem em monitor), especializados, bilíngües, de língua nativa, comparáveis (segundo os critérios do COMET) e de estudo.

Ontologia/Taxonomia: a árvore de campo

Um dos pontos básicos para a elaboração de um banco de dados é a criação de uma estrutura para organizar a informação a ser coletada. Vários tipos de estruturas podem ser elaboradas de acordo com o objetivo final. Segundo Almeida e Bax (2003, p. 7):

> [e]struturas que se organizam a partir da utilização de termos são os *arquivos de autoridade*, *glossários* e *dicionários*. Estruturas que se organizam com a classificação e a criação de categorias são os *cabeçalhos de assunto* e os *esquemas de classificação* (ou *taxonomias*). As estruturas que se organizam a partir de conceitos e de seus relacionamentos são as *ontologias*, os *tesaurus* e as *redes semânticas*. (grifos dos autores)

Embora essa classificação dos autores pareça bastante clara, há diversos problemas em torná-la universal. Sowa (1999), por exemplo, defende uma idéia de categorização para ontologias ao colocar que

> O assunto da *ontologia* é o estudo das *categorias* de coisas que existem ou podem vir a existir em algum domínio. O produto de tal estudo, chamado *ontologia*, é um catálogo de tipos de coisas que se pressupõe existirem em um domínio de interesse ***D*** da perspectiva de uma pessoa que usa uma língua ***L*** para o propósito de falar sobre ***D***.[2]
> (grifos do autor; minha tradução)

Tendo em vista esses diferentes conceitos para denominar o que é uma ontologia e uma taxonomia (técnica de classificação, segundo Hoauiss), para o presente trabalho foi escolhido o termo *taxonomia* para indicar a construção da árvore do campo pesquisada para a construção do *corpus*. No site desenvolvido para a inserção dos dados do banco (http://jr.icmc.sc.usp.br/~comet/dic/; acesso restrito), no entanto, optou-se pelo uso do termo *ontologia* para designar essa mesma árvore. O termo *ontologia*, cada vez mais, está associado ao uso de ferramentas computacionais para diversos tipos de análise, o que se prova pertinente para o presente caso.

O modelo tomado como base para a construção de uma árvore de campo foi aquele apresentado por Marinotto (1995) para a área de Aeronáutica e a divisão hierárquica proposta para o saber humano: campo, área, domínio, subdomínio e outros. A árvore do campo da computação já havia sido previamente desenvolvida por Fromm (2002) para a informática[3] em geral; aquela, no entanto, não mais representa um estado da arte do campo em questão: o extremo dinamismo desse campo na criação de novas tecnologias e produtos requer uma atualização constante da mesma. A construção de uma taxonomia, no entanto, não é infalível: há sempre controvérsias por parte dos especialistas quanto à sua montagem.

[2] The subject of *ontology* is the study of the *categories* of things that exist or may exist in some domain. The product of such a study, called *an ontology*, is a catalog of the types of things that are assumed to exist in a domain of interest ***D*** from the perspective of a person who uses a language ***L*** for the purpose of talking about ***D***.

[3] Embora os termos *informática* e *computação* não se apresentem como sinônimos para Houaiss, eles pertencem ao mesmo campo. Tomo, aqui, esses termos como sinônimos.

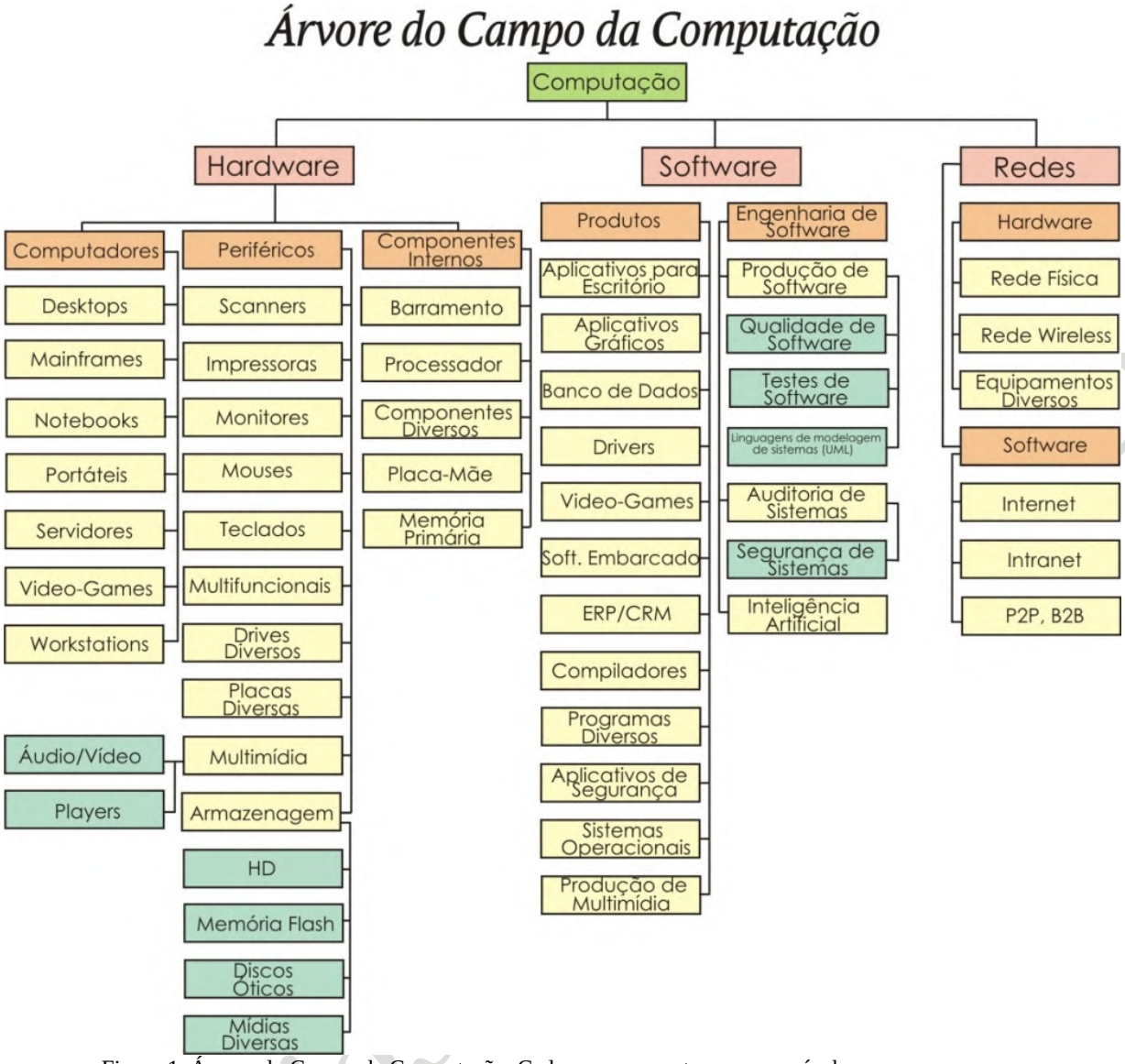


Figura 1. Árvore do Campo da Computação. Cada cor representa um novo nível.

A árvore acima (figura 1) representa o consenso entre a opinião de alguns professores especialistas na área, o que não quer dizer que seja unanimidade. Além disso, ainda que a Árvore de Campo (também designada Árvore de Domínio[4]) seja uma das possíveis formas de representar uma taxonomia, uma outra forma bastante comum é aquela apresentada na seqüência abaixo, quando da inserção das áreas feita pelo administrador no banco de dados.

[4] "Árvore de domínio: diagrama ou estrutura que organiza, de modo funcional, os conceitos de uma área temática. Tal árvore não representa uma classificação científica, mas uma maneira funcional de agrupar os conceitos de acordo com seu parentesco". DUBUC, R. Manual práctico de terminología. 3.ed. corr. atualiz.; trad. de Ileana Cabrera. Santiago de Chile: Unión Latina; Ril Ed, (1999, apud Lara, Tálamo, 2007).

- Grandes Áreas
  - Computação
    - Hardware
      - Componentes Internos
      - Computadores
      - Periféricos
        - Armazenagem
          - Cartão Flash
          - Discos Óticos
          - HD
          - Pen-Drive
        - Drives Diversos
        - Impressoras
        - Monitores
        - Mouses
        - Multifuncionais
        - Multimídia
        - Placas Diversas
        - Scanner
        - Teclados
    - Rede
    - Software

A coleta dos textos

Estabelecida a árvore, o passo seguinte foi a captura de uma quantidade de textos, em todas as áreas, suficiente para exibir contextos que pudessem criar definições para os termos. Já existem programas que fazem a coleta e extração de termos automaticamente, como o BootCaT, e ambientes de criação, armazenamento e análise de *corpora*, como o Corpógrafo (bem detalhados por ALMEIDA; OLIVEIRA; ALUÍSIO, 2006). Preferiu-se aqui, no entanto, não utilizar essas e outras ferramentas disponíveis, pois muitas ainda estão em fase de testes e não garantem o balanceamento de *corpus* exigido pelo trabalho; a coleta foi feita manualmente e depois os textos foram processados de acordo com as necessidades aqui propostas.

Segundo Aubert (1996), as fontes de busca para a definição de um termo podem apresentar três tipos de contextos possíveis:

> O *contexto associativo* apresenta o termo como pertinente ao tema objeto da pesquisa, mas não indica os traços conceptuais específicos destes termos, [...] Já os *contextos explicativos* apresentam alguns traços conceptuais pertinentes específicos do termo sob observação, freqüentemente relativos à materialidade, finalidade, funcionamento e

> similares. [...] Talvez mais desejáveis, mas certamente menos encontradiços, os *contextos definitórios* proporcionam um conjunto completo dos traços conceptuais distintivos do termo. Tal distintividade, no entanto, representa freqüentemente um certo nível de abstração, sem indícios claros da gama efetiva de usos em situação do termo. (p. 66-67)

A busca por contextos associativos, no caso da presente pesquisa, pode ser automatizada através dos programas de análise lexical (como o WordSmith Tools, que será explicado adiante). Os contextos explicativos e definitórios, por outro lado, exigem certo conhecimento do terminográfo sobre como localizá-los.

A necessidade de refazer os *corpora* e não apenas reutilizar os que já existiam deveu-se justamente à falta de contextos explicativos e definitórios. Verificou-se que, ao proceder à análise computadorizada dos textos previamente selecionados, havia lacunas em alguns subdomínios. Mesmo nos subdomínios com vários textos já coletados, o levantamento dos contextos foi insatisfatório.

Levando tudo isso em conta, ao começar uma nova coleta de textos, foi estabelecido um número mínimo de vinte mil palavras para cada subdomínio da árvore, quantidade que se acreditou razoável (e que se mostrou acertada, após alguns testes iniciais com um dos subdomínios disponíveis e a construção de alguns termos como teste) para o levantamento dos termos e um bom balanceamento entre esses subdomínios. Notou-se, porém, que já havia mais de um milhão de palavras, em cada língua, quando do término da primeira área da árvore (*hardware*). Decidiu-se, então, limitar os *corpora* a esse tamanho para o desenvolvimento da pesquisa. Como o objetivo da construção do banco não era fazer um levantamento completo de um campo técnico e sim coletar alguns exemplos de termos e seus contextos para posterior análise, o número obtido foi julgado suficiente, inclusive por abranger uma área completa.

A coleta dos *corpora*

Os textos coletados para os *corpora* de análise no campo da computação foram totalmente levantados pela Internet em sites especializados, muitos de caráter enciclopédico. A escolha se deveu à facilidade de encontrar textos do campo na rede (isso é uma característica marcante, já que nem todos os campos do saber estão bem representados em termo de quantidade e qualidade na Internet) e a velocidade com que os mesmos podem ser resgatados. Embora existam muito mais sites em inglês sobre o

campo da computação, não houve dificuldade para achar sites semelhantes (ou até mesmo traduzidos, como o *How Stuff Works*[5]) em português. Foi dada preferência aos sites de revistas especializadas, acadêmicos ou aqueles especializados em determinado assunto para o levantamento da pesquisa. Um site enciclopédico, no entanto, foi deixado de lado: a Wikipedia. A razão é que os termos apresentados pela mesma são disponibilizados, na íntegra, no site que dá acesso ao banco de dados (http://jr.icmc.sc.usp.br/~comet/dic/).

Para coletar os *corpora*, criou-se um diretório no computador que exibia pastas na mesma estrutura da árvore de campo (figura 1). O mesmo foi subdividido entre as línguas (inglês e português) e todas as áreas, domínios e subdomínios. Novos textos coletados e aqueles remanescentes do projeto original de Informática do CORTEC, na área de *hardware*, já foram distribuídos dentro de suas respectivas pastas. Os textos remanescentes das áreas *software* e rede foram alocados, também, nas respectivas pastas. Embora somente na área de *hardware* novos textos tenham sido coletados, todos aqueles já coletados para o CORTEC foram aproveitados para o estudo.

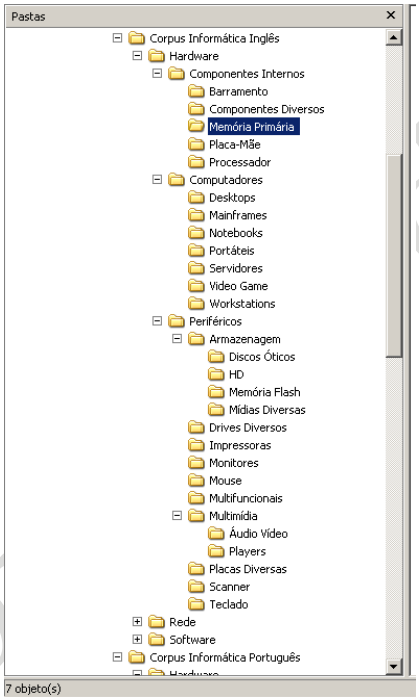


Figura 2. Diretório com pastas na forma da árvore de campo; área: *hardware*.

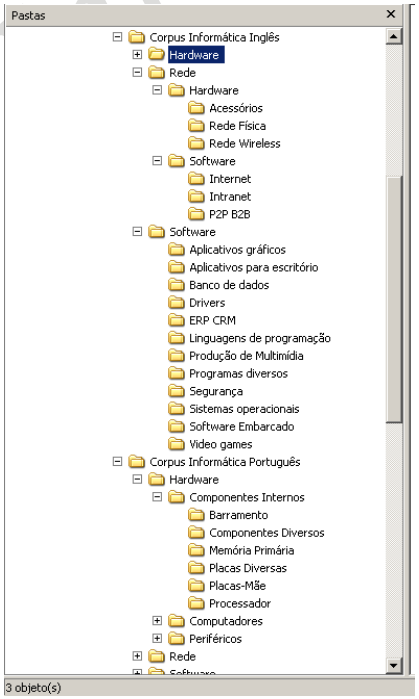


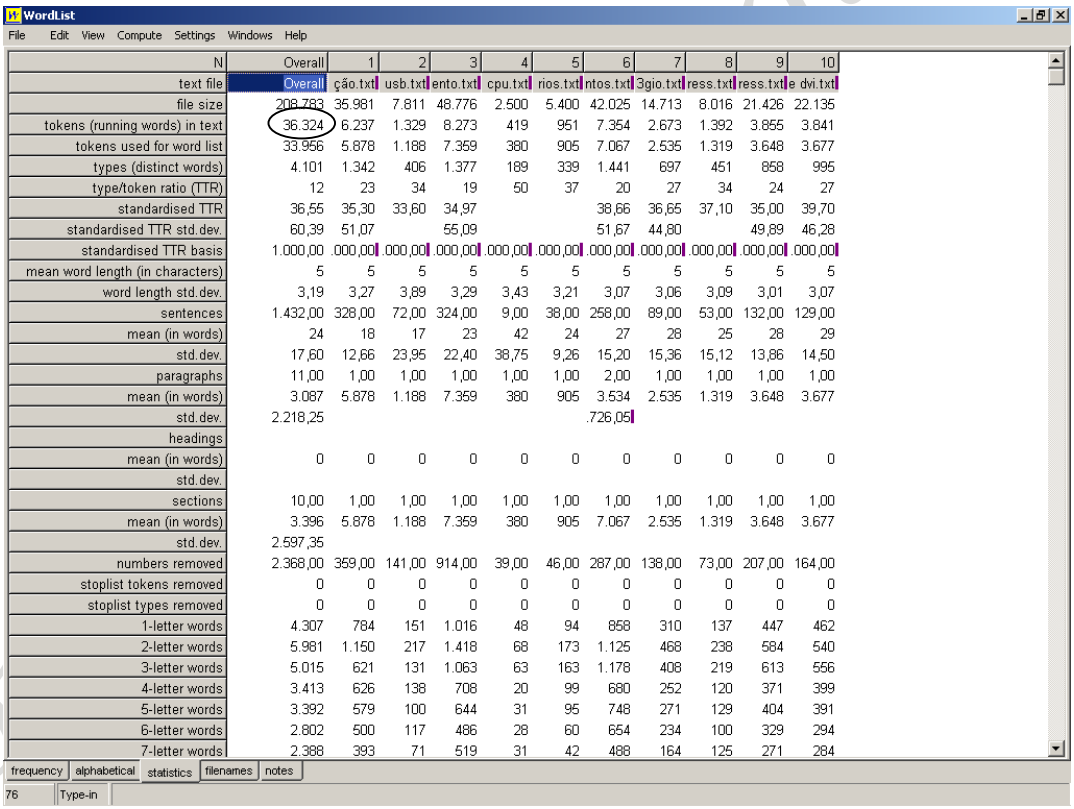
Figura 3. Idem; áreas: *software* e rede.

[5] < http://hsw.uol.com.br>. O endereço desta e de todas as outras páginas consultadas está disponível no banco de dados.

A análise dos *corpora*

Para verificar se a quantidade de textos alocados a cada pasta obedecia ao critério de vinte mil palavras por subdomínio[6], foi usada a ferramenta Wordlist (listagem de palavras) do programa de Análise Lexical WordSmith Tools, versão 4, de Scott (2007), para fazer a contagem (veja figura 4, no destaque). Embora haja vários programas de análise computadorizada, conforme estudos anteriores (FROMM, 2004), o WordSmith Tools é o mais indicado para grande quantidade de dados e para os tipos de análise que serão demonstrados a seguir.

O volume total de palavras para o *corpus* de computação foi de 1.029.187 palavras em inglês e 1.055.375 palavras em português. Segundo Berber Sardinha (2004, p.26), esses *corpora* seriam classificados, de acordo com a quantidade de palavras, como médios (de 250 mil a um milhão de palavras).


WordList

File Edit View Compute Settings Windows Help

N	Overall	1	2	3	4	5	6	7	8	9	10
text file	Overall	ção.txt	usb.txt	ento.txt	cpu.txt	rios.txt	ntos.txt	3gio.txt	ress.txt	ress.txt	e dvi.txt
file size	208.783	35.981	7.811	48.776	2.500	5.400	42.025	14.713	8.016	21.426	22.135
tokens (running words) in text	36.324	6.237	1.329	8.273	419	951	7.354	2.673	1.392	3.855	3.841
tokens used for word list	33.956	5.878	1.188	7.359	380	905	7.067	2.535	1.319	3.648	3.677
types (distinct words)	4.101	1.342	406	1.377	189	339	1.441	697	451	858	995
type/token ratio (TTR)	12	23	34	19	50	37	20	27	34	24	27
standardised TTR	36,55	35,30	33,60	34,97			38,66	36,65	37,10	35,00	39,70
standardised TTR std.dev.	60,39	51,07		55,09			51,67	44,80		49,89	46,28
standardised TTR basis	1.000,00	.000,00	.000,00	.000,00	.000,00	.000,00	.000,00	.000,00	.000,00	.000,00	.000,00
mean word length (in characters)	5	5	5	5	5	5	5	5	5	5	5
word length std.dev.	3,19	3,27	3,89	3,29	3,43	3,21	3,07	3,06	3,09	3,01	3,07
sentences	1.432,00	328,00	72,00	324,00	9,00	38,00	258,00	89,00	53,00	132,00	129,00
mean (in words)	24	18	17	23	42	24	27	28	25	28	29
std.dev.	17,60	12,66	23,95	22,40	38,75	9,26	15,20	15,36	15,12	13,86	14,50
paragraphs	11,00	1,00	1,00	1,00	1,00	1,00	2,00	1,00	1,00	1,00	1,00
mean (in words)	3.087	5.878	1.188	7.359	380	905	3.534	2.535	1.319	3.648	3.677
std.dev.	2.218,25						,726,05				
headings											
mean (in words)	0	0	0	0	0	0	0	0	0	0	0
std.dev.											
sections	10,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
mean (in words)	3.396	5.878	1.188	7.359	380	905	7.067	2.535	1.319	3.648	3.677
std.dev.	2.597,35										
numbers removed	2.368,00	359,00	141,00	914,00	39,00	46,00	287,00	138,00	73,00	207,00	164,00
stoplist tokens removed	0	0	0	0	0	0	0	0	0	0	0
stoplist types removed	0	0	0	0	0	0	0	0	0	0	0
1-letter words	4.307	784	151	1.016	48	94	858	310	137	447	462
2-letter words	5.981	1.150	217	1.418	68	173	1.125	468	238	584	540
3-letter words	5.015	621	131	1.063	63	163	1.178	408	219	613	556
4-letter words	3.413	626	138	708	20	99	680	252	120	371	399
5-letter words	3.392	579	100	644	31	95	748	271	129	404	391
6-letter words	2.802	500	117	486	28	60	654	234	100	329	294
7-letter words	2.388	393	71	519	31	42	488	164	125	271	284

frequency alphabetical statistics filenames notes

76 Type-in

Figura 4. A subárea *componentes diversos* (em português) apresenta uma quantidade de 36.324 palavras no total (em destaque).

[6] Verificada através da quantidade de *tokens* que a listagem apresenta. Os *tokens* representam a quantidade total de palavras nos textos, os *types* representam a quantidade de palavras não repetidas (distintas) nos textos.

A cada vinte mil palavras levantadas, partia-se para uma nova subárea. Algumas subáreas, no entanto, têm um valor bastante superior a esse. O limite de vinte mil palavras, portanto, foi o mínimo a ser levantado; não houve preocupação com o volume máximo. Berber Sardinha, ao citar Sinclair[7] (1997, p.27-39 *apud* BERBER SARDINHA, 2004, p.26), comenta uma entre as possíveis abordagens a respeito da extensão do *corpus* (no caso, a Impressionística):

> Sinclair [...] postula que o *corpus* deva ser tão grande quanto a tecnologia permitir para a época, deixando subentender que a extensão de um *corpus* deva variar de acordo com o padrão corrente nos grandes centros de pesquisa, que possuem equipamentos de última geração" (p. 26).

As variações de tamanho deram-se em virtude dos tipos de arquivos baixados: de algumas páginas o texto foi retirado no formato .html, copiado e colado para um arquivo formato .txt; outras páginas forneceram arquivos no formato .pdf que, sempre que possível, foram copiados para .txt também (alguns não puderam ser copiados e foram, portanto, descartados). Os arquivos em formato .pdf, normalmente estudos acadêmicos sobre a área, manuais de instrução ou propaganda dos fabricantes, têm uma quantidade maior de palavras. Essa preocupação em transformar todos os arquivos para o formato .txt dá-se por causa da velocidade de análise do programa WordSmith Tools 4. Embora ele também leia arquivos salvos em outros formatos, é no .txt que ele tem o máximo de desempenho. Os arquivos foram salvos com o título do texto (quando havia repetição dos títulos, foram acrescentados números seqüenciais) e, para fins de posterior análise, depois de copiados os textos, foram incluídos o endereço do site e a data de coleta (figura 5).

Ao término da coleta e primeira análise dos *corpora* em forma de Wordlist, partiu-se para o segundo passo, que é a criação das palavras-chave (*Keywords*). Antes de iniciar a ferramenta *Keywords* do WordSmith Tools, é necessária a criação dos chamados *corpora* de referência, que são grandes *corpora* de textos gerais da língua em análise e que servem como parâmetro de comparação para a ferramenta. Em português, foi usada a versão beta do Banco de Português (BERBER SARDINHA, 2007), totalizando 689.294.592 palavras; em inglês usou-se uma combinação das listas de palavras do BNC (British National *Corpus*) e de uma versão beta do ANC (American

[7] SINCLAIR, J. Corpus evidence in language description. In: WICHMANN, A. S. et al. **Teaching and language corpora**. Londres/Nova Iorque: Longman, 1997.

National *Corpus*)[8], totalizando 122.224.832 palavras. Em ambos os casos, os *corpora* de referência são bem maiores do que a proporção de cinco para um (o *corpus* de referência é cinco vezes maior que o *corpus* de análise) proposta por Berber-Sardinha (2004, p.102) como o tamanho recomendado.

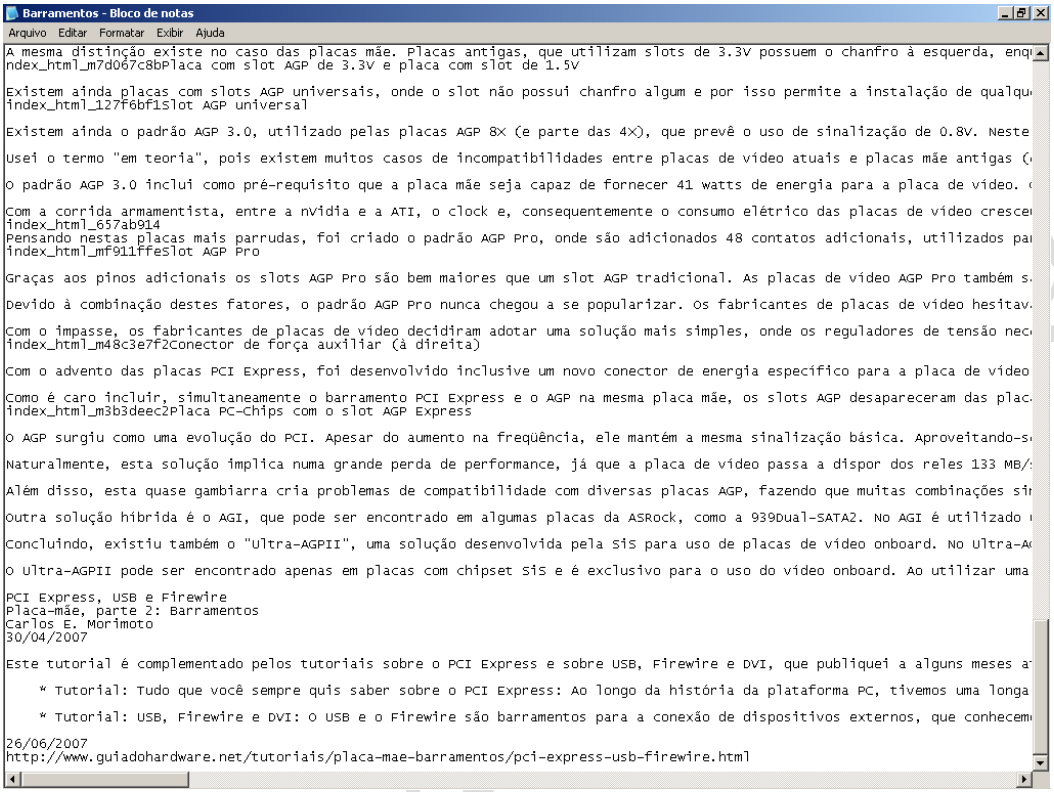


Figura 5. Arquivo .txt do *corpus*, com data de coleta e endereço na Internet (final da página).

Com a ferramenta Keywords do WordSmtih Tools foram criadas, então, as listagens de palavras-chave em cada língua. Essas palavras, escolhidas por meio de análises estatísticas (*log likelihood*) entre o *corpus* de estudo e o *corpus* de referência, correspondem aos contextos associativos já citados[9]. Esses contextos não foram usados para a construção das definições na presente pesquisa, apenas os explicativos/definitórios. Em outros tipos de estudo, no entanto, quando o terminógrafo não conhece a área, os contextos associativos podem se configurar como um ponto de partida para análises preliminares sobre os candidatos a termos. Na figura 6 temos a

[8] A listagem do BNC foi obtida no site do programa Wordsmith Tools. A listagem do ANC foi elaborada tendo a segunda versão do CD como *corpus* e o programa Wordsmith Tools como ferramenta de análise.
[9] O programa faz uma análise contrastiva entre os dois *corpora* e verifica as palavras que se destacam, pela freqüência de uso, no *corpus* de especialidade. As palavras apresentadas na listagem fazem parte, portanto, do campo que está sendo estudado.

tela do programa com as palavras-chave em inglês; na planilha 1, a tela com as palavras-chave em português, agora numa listagem em Excel.

Informática - Keywords em Inglês.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P	Lemmas	Set
1	DISK	2.322	0,23	2.887		15.117,46	0,0000000000		
2	MEMORY	2.709	0,26	8.744		13.550,35	0,0000000000		
3	DATA	3.246	0,32	31.771	0,03	9.986,71	0,0000000000		
4	CPU	1.092	0,11	316		8.958,52	0,0000000000		
5	DRIVE	2.154	0,21	11.440		8.928,05	0,0000000000		
6	ARN	882	0,09	12		8.315,24	0,0000000000		
7	DRIVES	1.233	0,12	1.735		7.802,31	0,0000000000		
8	PCI	833	0,08	84		7.413,06	0,0000000000		
9	PC	1.267	0,12	2.755		7.162,23	0,0000000000		
10	VIRTUAL	1.053	0,10	1.225		6.954,83	0,0000000000		
11	KEYBOARD	983	0,10	1.098		6.549,06	0,0000000000		
12	BIT	2.468	0,24	31.963	0,03	6.398,03	0,0000000000		
13	USB	639	0,06	37		5.829,93	0,0000000000		
14	ADDRESS	1.491	0,14	9.535		5.695,09	0,0000000000		
15	DEVICES	1.016	0,10	2.611		5.466,72	0,0000000000		
16	CONNECTOR	664	0,06	200		5.423,94	0,0000000000		
17	PARALLELS	744	0,07	559		5.350,71	0,0000000000		
18	INTERFACE	952	0,09	2.188		5.296,06	0,0000000000		
19	CD	910	0,09	1.834		5.254,41	0,0000000000		
20	DVD	707	0,07	555		5.045,30	0,0000000000		
21	MOV	527	0,05	6		4.978,49	0,0000000000		
22	FLOPPY	698	0,07	559		4.963,20	0,0000000000		
23	COMPUTER	1.644	0,16	17.184	0,01	4.868,08	0,0000000000		
24	SOFTWARE	1.425	0,14	11.505		4.861,09	0,0000000000		
25	DEVICE	987	0,10	3.395		4.829,11	0,0000000000		
26	SERVER	844	0,08	1.770		4.819,72	0,0000000000		
27	DISKS	715	0,07	829		4.725,61	0,0000000000		
28	WORKSTATION	697	0,07	759		4.668,34	0,0000000000		
29	HARDWARE	866	0,08	2.321		4.599,46	0,0000000000		
30	HARD	1.923	0,19	28.720	0,02	4.518,88	0,0000000000		
31	PROCESSOR	772	0,08	1.578		4.440,14	0,0000000000		

KWs plot links clusters filenames notes source text

3.946 Type-in PAGE

Figura 6. Palavras-chave na área de computação, em inglês.

As palavras na primeira coluna indicam as palavras mais freqüentes que, assim indicam os candidatos prováveis a termos naquela área; a ordem de palavras na primeira/segunda colunas leva em conta a sétima coluna, ou seja, sua chavicidade (*keyness*), que significa o quanto a palavra em destaque, na relação entre o *corpus* de análise e o *corpus* de referência, é representativa na freqüência relativa (o programa compara, estatisticamente, a freqüência desta palavra em ambos os *corpora*; se ela apresenta um uso mais [ou menos] destacado no *corpus* de análise do que no de referência, ela é incluída na lista).

				WordSmith Tools 4.0 -- 27/6/2007			
N	**Key word**	**Freq.**	**%**	**RC. Freq.**	**RC. %**	**Keyness**	**P**
1	COMPUTADOR	3380	0,3203	28792		22302,23	7E-23
2	IMPRESSORA	1812	0,1717	2526		17610,62	1E-22
3	CLIQUE	1553	0,1472	2491		14760,42	3E-22
4	PLACA	2104	0,1994	19180		13613,9	3E-22
5	WINDOWS	1603	0,1519	11497		11086,75	6E-22
6	PROCESSADOR	1212	0,1148	2901		10738,48	7E-22
7	BITS	1183	0,1121	2547		10688,28	7E-22
8	USB	913	0,0865	326		10412,25	7E-22
9	IMPRESSÃO	1617	0,1532	19473		9616,599	9E-22
10	BARRAMENTO	815	0,0772	344		9159,807	1E-21
11	PLACAS	1514	0,1435	17867		9061,677	1E-21
12	MEMÓRIA	1964	0,1861	49114		8969,695	1E-21
13	GEFORCE	672	0,0637	2		8686,718	1E-21
14	SELECIONE	755	0,0715	496		8111,6	2E-21
15	PCI	782	0,0741	1351		7341,269	2E-21
16	VÍDEO	1379	0,1307	25009		7134,15	2E-21
17	TELA	1241	0,1176	20775		6607,95	3E-21
18	CONTROLADOR	874	0,0828	4499		6575,389	3E-21
19	DVD	755	0,0715	2026		6544,251	3E-21
20	VOCÊ	2539	0,2406	193135	0,028	6410,745	3E-21
21	RADEON	491	0,0465	0		6366,831	3E-21
22	CONSULTE	630	0,0597	701		6330,112	3E-21
23	*HARDWARE*	684	0,0648	1707		6012,326	4E-21
24	PALM	700	0,0663	2140		5911,728	4E-21
25	TECLADO	692	0,0656	2275		5757,233	4E-21
26	DISCO	1345	0,1274	39388		5742,442	4E-21
27	MOUSE	750	0,0711	3769		5675,047	4E-21
28	MONITOR	796	0,0754	6107		5405,326	5E-21
29	XP	558	0,0529	862		5335,347	5E-21
30	MB	781	0,074	6179		5258,846	6E-21
31	BOTÃO	693	0,0657	3595		5203,829	6E-21
32	EAX	388	0,0368	9		4945,257	7E-21
33	MEMORIA	480	0,0455	466		4914,386	7E-21

Planilha 1. Palavras-chave em português.

Identificados os candidatos a termos nas duas línguas, é preciso verificar quais deles estão presentes em ambas as listas. Nesse momento é necessário um pouco da *expertise* (conhecimento sobre a área) do pesquisador para delimitar quais termos são equivalentes nas duas línguas. Alguns são empréstimos (nessa área, em especial, são bastante numerosos; como, por exemplo, *mainframe*) ou decalques (em que os verbos

são destaque: *deletar*, *chipar*, etc.), outros são acrônimos ou abreviações usados indistintamente nas duas línguas (como *AGP*), outros ainda requerem uma consulta a obras bilíngües já existentes para verificar, num primeiro momento, se são equivalentes (Platters – Discos, componentes do disco rígido; a primeira acepção de *platter*, segundo o American Heritage Dictionary, é o equivalente, em português, a *travessa* ou *prato grande*; o termo corrente em português, neste caso, é *disco*).

Na planilha 2, são mostradas as colunas das palavras-chave, numa planilha em Excel, indicando sua ordem pela chavicidade dos termos em cada língua.

Ordem	Português	Ordem	Inglês
1	COMPUTADOR	23	COMPUTER
2	IMPRESSORA	65	PRINTER
4	PLACA	782	BOARD
6	PROCESSADOR	31	PROCESSOR
8	USB	13	USB
10	BARRAMENTO	71	BUS
12	MEMÓRIA	2	MEMORY
15	PCI	8	PCI
18	CONTROLADOR	46	CONTROLLER
25	TECLADO	11	KEYBOARD
26	DISCO	132	PLATTERS
36	MHZ	203	MHZ
37	APLICATIVOS	88	APPLICATIONS
38	TECLA	199	KEY
45	AGP	237	AGP
47	DADOS	3	DATA
57	DISPOSITIVO	25	DEVICE
64	SERVIDOR	26	SERVER
67	INSTALAR	194	INSTALL
68	DRIVE	5	DRIVE
72	SCSI	45	SCSI
77	HTTP	82	HTTP
81	INTERFACE	18	INTERFACE
83	ROM	53	ROM
90	CHIP	142	CHIP
92	RÍGIDO	30	HARD
102	DRIVER	499	DRIVER
107	CONFIGURAÇÕES	181	SETTINGS
108	FIREWIRE	330	FIREWIRE
110	MAINFRAME	157	MAINFRAME
113	RAID	684	RAID

Planilha 2. Relação de termos equivalentes nas duas línguas.

A equivalência dos termos na listagem não garante, contudo, que todos eles apresentem contextos explicativos ou definitórios. Como o objetivo da construção do banco de dados e da página de consulta é fornecer um ambiente de pesquisa que indique também a definição do termo, é necessário identificar um desses dois contextos, explicativos ou definitórios, para termos equivalentes nas duas línguas. Muitos termos, nessa comparação, não foram aprovados pela dificuldade em se achar contextos claros (já prevendo essa, foram selecionados cem termos equivalentes em cada língua para haver uma margem de descarte). A planilha 3 apresenta a listagem parcial dos candidatos a termos equivalentes na área de computação. As escalas de cinza das legendas indicam os termos com contextos explicativos e/ou definitórios encontrados nas duas línguas, encontrados somente em uma língua ou não encontrados em nenhuma das duas[10]. Conforme os termos eram inseridos no banco de dados, uma marca com tons de cinza ou preto também era feita ao lado. Os números, antepostos ao termo, assim como na planilha 2, indicam sua ordem de chavicidade.

	Português		Inglês
1	COMPUTADOR	23	COMPUTER
2	IMPRESSORA	65	PRINTER
4	PLACA	782	BOARD
6	PROCESSADOR	31	PROCESSOR
8	USB	13	USB
10	BARRAMENTO	71	BUS
12	MEMÓRIA	2	MEMORY
15	PCI	8	PCI
18	CONTROLADOR	46	CONTROLLER
25	TECLADO	11	KEYBOARD
26	DISCO	132	PLATTERS
36	MHZ	203	MHZ
37	APLICATIVOS	88	APPLICATIONS
38	TECLA	199	KEY
45	AGP	237	AGP
47	DADOS	3	DATA
57	DISPOSITIVO	25	DEVICE
64	SERVIDOR	26	SERVER
67	INSTALAR	194	INSTALL
68	DRIVE	5	DRIVE

Legenda

- definição encontrada nas duas línguas
- definição não encontrada nas duas línguas
- definição encontrada em inglês, mas não em português
- definição encontrada em português, mas não em inglês

- adicionado ao banco de dados
- não adicionado ao banco de dados

[10] Uma possível ampliação do *corpus* de estudo, inclusive diacronicamente, deve fornecer todos os contextos necessários para o campo de definição dos termos.

72	SCSI	45	SCSI
77	HTTP	82	HTTP
81	INTERFACE	18	INTERFACE
83	ROM	53	ROM
90	CHIP	142	CHIP
92	RÍGIDO	30	HARD
102	DRIVER	499	DRIVER
107	CONFIGURAÇÕES	181	SETTINGS
108	FIREWIRE	330	FIREWIRE
110	MAINFRAME	157	MAINFRAME
113	RAID	684	RAID

Planilha 3. Área de computação, alguns candidatos a termos.

Para obter os contextos de cada termo, utilizamos uma terceira ferramenta do WordSmith Tools: o concordanciador (*Concordancer*). Ao selecionar o termo na listagem de palavras-chave e pedir suas concordâncias, o programa cria uma nova tela, com o termo em questão centralizado e na cor azul (tela KWIC, *key word in context*), mostrando suas ocorrências em todos os textos (figura 7). Basta clicar duas vezes na linha desejada, na coluna *File*, para que o texto seja mostrado por completo.

Para descobrir quais dessas linhas (cada uma representa a seleção de uma linha de um texto) podem nos fornecer os contextos desejados, foram usados, basicamente, dois artifícios:

1. uma busca por sinais de pontuação. Nos textos da figura 7, foi feita, inicialmente, uma busca usando os parâmetros de **:** (dois pontos), **(** (parênteses) ou **,** (vírgula). A idéia era achar esses contextos depois de pontuação (dois pontos ou parênteses) ou como aposto (entre vírgulas). Para realizar essa busca no programa, é necessário acrescentar o asterisco (*) depois da pontuação desejada. No caso do exemplo acima, a busca seria realizada como: **computador:***, **computador (*** ou **computador,***;

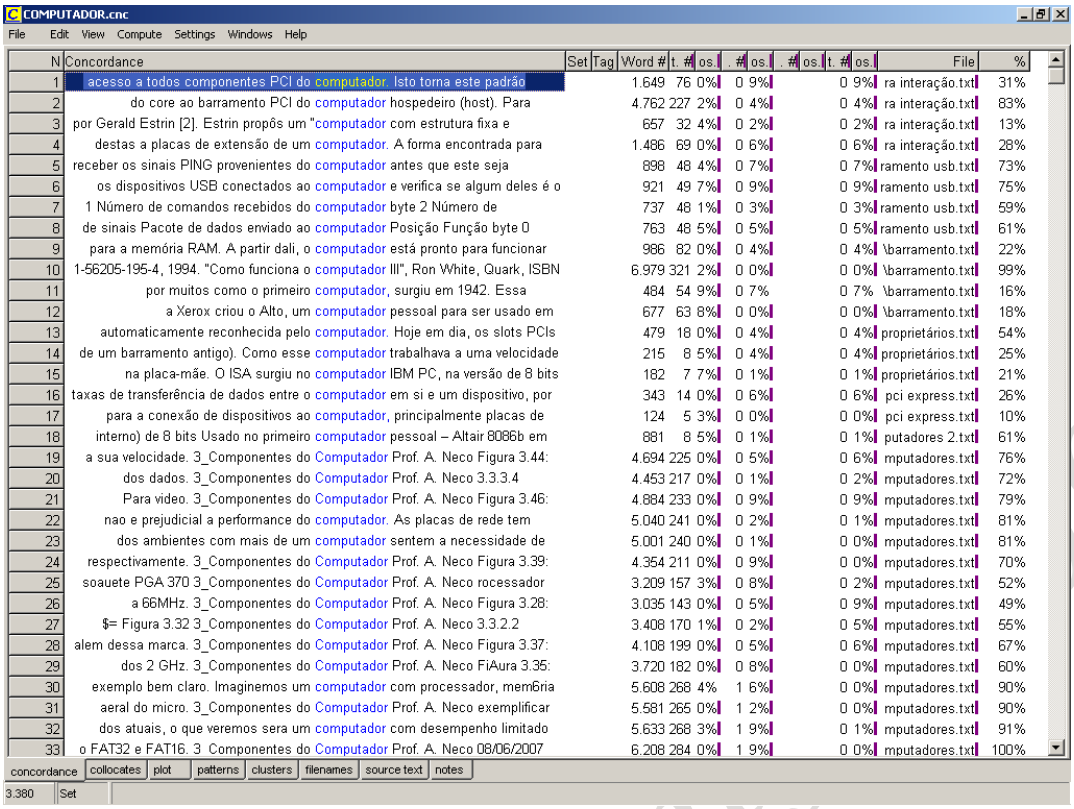
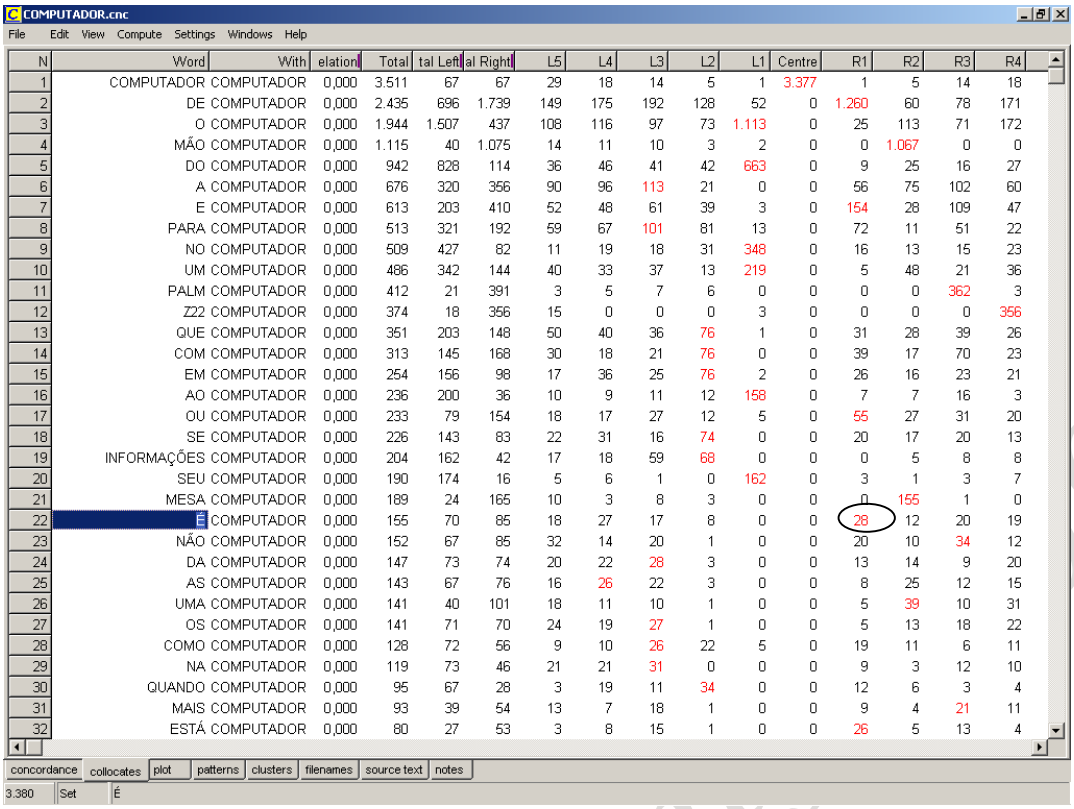


Figura 7. Termo “computador”, em uma tela de concordâncias, totalizando 3.380 delas.

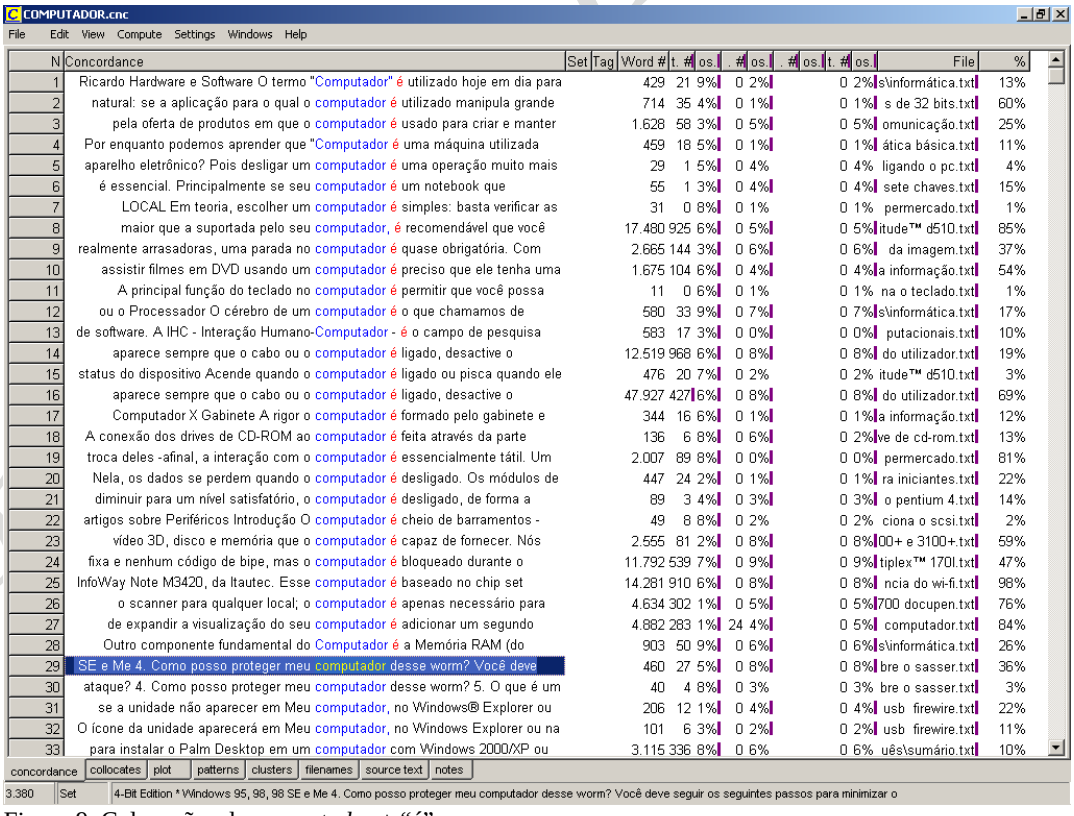
2. uma busca pelos colocados (para o programa, colocado é a combinação de alta freqüência entre o termo selecionado mais um termo a ele associado)[11]. O primeiro termo procurado como colocação foi o verbo *ser* (ou *to be*) em todas as suas formas. Veja na figura 8 as colocações para o termo “computador”: existem vinte e oito colocações com o verbo *ser* (*é*) no primeiro campo à direita de computador (coluna R1, seguinte à coluna “*centre*”). Ao clicar no número vinte e oito (em vermelho, no original; em destaque, aqui), a tela volta para a apresentação das concordâncias e destaca os segmentos de texto que apresentam essa combinação (figura 9). Na linha quatro dessa nova tela, por exemplo, temos um contexto definitório para o termo computador (“... podemos aprender que computador é uma máquina utilizada...”). Para ver todo o parágrafo, basta clicar duas vezes sobre a linha e o programa abre uma nova tela (figura 10).

[11] Para Sardinha (2004, p. 40) é a“ [...] associação entre itens lexicais, ou entre o léxico e campos semânticos”.



N	Word	With	elation	Total	tal Left	al Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4
1	COMPUTADOR	COMPUTADOR	0,000	3.511	67	67	29	18	14	5	1	3.377	1	5	14	18
2	DE	COMPUTADOR	0,000	2.435	696	1.739	149	175	192	128	52	0	1.260	60	78	171
3	O	COMPUTADOR	0,000	1.944	1.507	437	108	116	97	73	1.113	0	25	113	71	172
4	MÃO	COMPUTADOR	0,000	1.115	40	1.075	14	11	10	3	2	0	0	1.067	0	0
5	DO	COMPUTADOR	0,000	942	828	114	36	46	41	42	663	0	9	25	16	27
6	A	COMPUTADOR	0,000	676	320	356	90	96	113	21	0	0	56	75	102	60
7	E	COMPUTADOR	0,000	613	203	410	52	48	61	39	3	0	154	28	109	47
8	PARA	COMPUTADOR	0,000	513	321	192	59	67	101	81	13	0	72	11	51	22
9	NO	COMPUTADOR	0,000	509	427	82	11	19	18	31	348	0	16	13	15	23
10	UM	COMPUTADOR	0,000	486	342	144	40	33	37	13	219	0	5	48	21	36
11	PALM	COMPUTADOR	0,000	412	21	391	3	5	7	6	0	0	0	0	362	3
12	Z22	COMPUTADOR	0,000	374	18	356	15	0	0	0	3	0	0	0	0	356
13	QUE	COMPUTADOR	0,000	351	203	148	50	40	36	76	1	0	31	28	39	26
14	COM	COMPUTADOR	0,000	313	145	168	30	18	21	76	0	0	39	17	70	23
15	EM	COMPUTADOR	0,000	254	156	98	17	36	25	76	2	0	26	16	23	21
16	AO	COMPUTADOR	0,000	236	200	36	10	9	11	12	158	0	7	7	16	3
17	OU	COMPUTADOR	0,000	233	79	154	18	17	27	12	5	0	55	27	31	20
18	SE	COMPUTADOR	0,000	226	143	83	22	31	16	74	0	0	20	17	20	13
19	INFORMAÇÕES	COMPUTADOR	0,000	204	162	42	17	18	59	68	0	0	0	5	8	8
20	SEU	COMPUTADOR	0,000	190	174	16	5	6	1	0	162	0	3	1	3	7
21	MESA	COMPUTADOR	0,000	189	24	165	10	3	8	3	0	0	0	155	1	0
22	É	COMPUTADOR	0,000	155	70	85	18	27	17	8	0	0	28	12	20	19
23	NÃO	COMPUTADOR	0,000	152	67	85	32	14	20	1	0	0	20	10	34	12
24	DA	COMPUTADOR	0,000	147	73	74	20	22	28	3	0	0	13	14	9	20
25	AS	COMPUTADOR	0,000	143	67	76	16	26	22	3	0	0	8	25	12	15
26	UMA	COMPUTADOR	0,000	141	40	101	18	11	10	1	0	0	5	39	10	31
27	OS	COMPUTADOR	0,000	141	71	70	24	19	27	1	0	0	5	13	18	22
28	COMO	COMPUTADOR	0,000	128	72	56	9	10	26	22	5	0	19	11	6	11
29	NA	COMPUTADOR	0,000	119	73	46	21	21	31	0	0	0	9	3	12	10
30	QUANDO	COMPUTADOR	0,000	95	67	28	3	19	11	34	0	0	12	6	3	4
31	MAIS	COMPUTADOR	0,000	93	39	54	13	7	18	1	0	0	9	4	21	11
32	ESTÁ	COMPUTADOR	0,000	80	27	53	3	8	15	1	0	0	26	5	13	4

Figura 8. Lista de colocações do termo *computador*.



N	Concordance	Set	Tag	Word #	t. #	os.	. #	os.	. #	os.	t. #	os.	File	%
1	Ricardo Hardware e Software O termo "Computador" é utilizado hoje em dia para			429	21	9%	0	2%			0	2%	s\informática.txt	13%
2	natural: se a aplicação para o qual o computador é utilizado manipula grande			714	35	4%	0	1%			0	1%	s de 32 bits.txt	60%
3	pela oferta de produtos em que o computador é usado para criar e manter			1.628	58	3%	0	5%			0	5%	omunicação.txt	25%
4	Por enquanto podemos aprender que "Computador é uma máquina utilizada			459	18	5%	0	1%			0	1%	ática básica.txt	11%
5	aparelho eletrônico? Pois desligar um computador é uma operação muito mais			29	1	5%	0	4%			0	4%	ligando o pc.txt	4%
6	é essencial. Principalmente se seu computador é um notebook que			55	1	3%	0	4%			0	4%	sete chaves.txt	15%
7	LOCAL Em teoria, escolher um computador é simples: basta verificar as			31	0	8%	0	1%			0	1%	permercado.txt	1%
8	maior que a suportada pelo seu computador, é recomendável que você			17.480	925	6%	0	5%			0	5%	itude™ d510.txt	85%
9	realmente arrasadoras, uma parada no computador é quase obrigatória. Com			2.665	144	3%	0	6%			0	6%	da imagem.txt	37%
10	assistir filmes em DVD usando um computador é preciso que ele tenha uma			1.675	104	6%	0	4%			0	4%	a informação.txt	54%
11	A principal função do teclado no computador é permitir que você possa			11	0	6%	0	1%			0	1%	na o teclado.txt	1%
12	ou o Processador O cérebro de um computador é o que chamamos de			580	33	9%	0	7%			0	7%	s\informática.txt	17%
13	de software. A IHC - Interação Humano-Computador - é o campo de pesquisa			583	17	3%	0	0%			0	0%	putacionais.txt	10%
14	aparece sempre que o cabo ou o computador é ligado, desactive o			12.519	968	6%	0	8%			0	8%	do utilizador.txt	19%
15	status do dispositivo Acende quando o computador é ligado ou pisca quando ele			476	20	7%	0	2%			0	2%	itude™ d510.txt	3%
16	aparece sempre que o cabo ou o computador é ligado, desactive o			47.927	427	6%	0	8%			0	8%	do utilizador.txt	69%
17	Computador X Gabinete A rigor o computador é formado pelo gabinete e			344	16	6%	0	1%			0	1%	a informação.txt	12%
18	A conexão dos drives de CD-ROM ao computador é feita através da parte			136	6	8%	0	6%			0	2%	ve de cd-rom.txt	13%
19	troca deles -afinal, a interação com o computador é essencialmente tátil. Um			2.007	89	8%	0	0%			0	0%	permercado.txt	81%
20	Nela, os dados se perdem quando o computador é desligado. Os módulos de			447	24	2%	0	1%			0	1%	ra iniciantes.txt	22%
21	diminuir para um nível satisfatório, o computador é desligado, de forma a			89	3	4%	0	3%			0	3%	o pentium 4.txt	14%
22	artigos sobre Periféricos Introdução O computador é cheio de barramentos -			49	8	8%	0	2%			0	2%	ciona o scsi.txt	2%
23	vídeo 3D, disco e memória que o computador é capaz de fornecer. Nós			2.555	81	2%	0	8%			0	8%	00+ e 3100+.txt	59%
24	fixa e nenhum código de bipe, mas o computador é bloqueado durante o			11.792	539	7%	0	9%			0	9%	tiplex™ 170l.txt	47%
25	InfoWay Note M3420, da Itautec. Esse computador é baseado no chip set			14.281	910	6%	0	8%			0	8%	ncia do wi-fi.txt	98%
26	o scanner para qualquer local; o computador é apenas necessário para			4.634	302	1%	0	5%			0	5%	700 docupen.txt	76%
27	de expandir a visualização do seu computador é adicionar um segundo			4.882	283	1%	24	4%			0	5%	computador.txt	84%
28	Outro componente fundamental do Computador é a Memória RAM (do			903	50	9%	0	6%			0	6%	s\informática.txt	26%
29	SE e Me 4. Como posso proteger meu computador desse worm? Você deve			460	27	5%	0	8%			0	8%	bre o sasser.txt	36%
30	ataque? 4. Como posso proteger meu computador desse worm? 5. O que é um			40	4	8%	0	3%			0	3%	bre o sasser.txt	3%
31	se a unidade não aparecer em Meu computador, no Windows® Explorer ou			206	12	1%	0	4%			0	4%	usb firewire.txt	22%
32	O ícone da unidade aparecerá em Meu computador, no Windows Explorer ou na			101	6	3%	0	2%			0	2%	usb firewire.txt	11%
33	para instalar o Palm Desktop em um computador com Windows 2000/XP ou			3.115	336	8%	0	6%			0	6%	uês\sumário.txt	10%

Figura 9. Colocações de *computador* + "é"

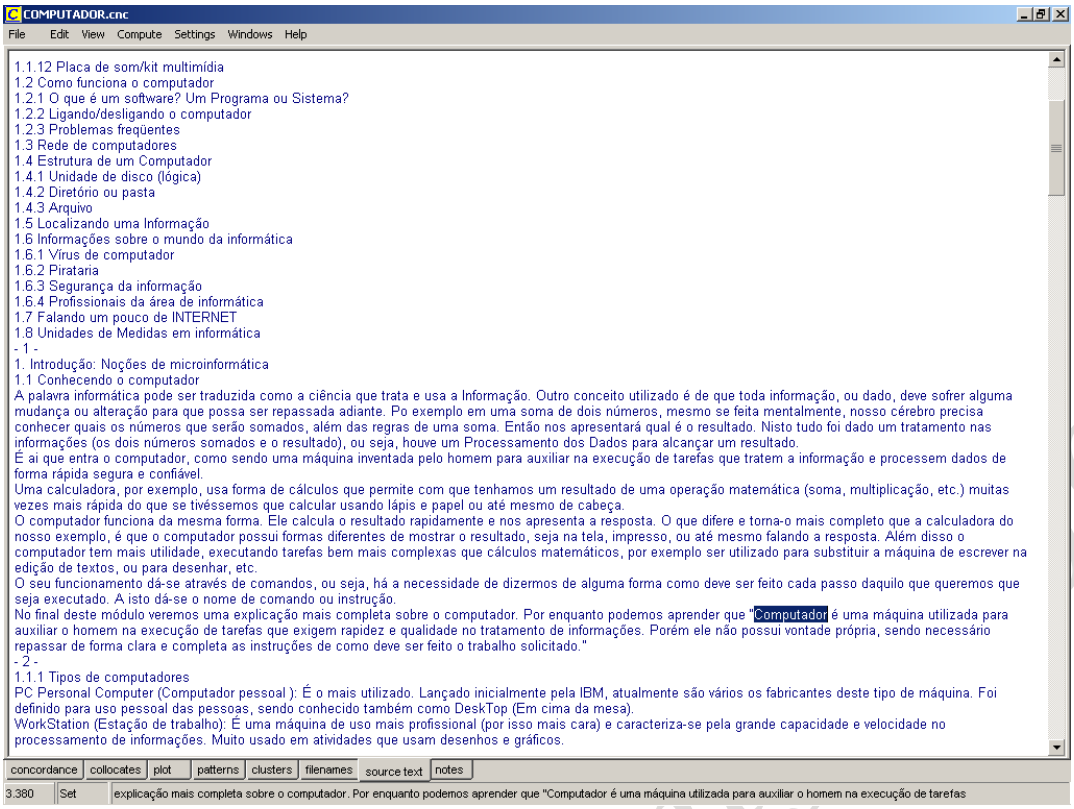


Figura 10. Contexto da quarta linha de concordância (figura anterior).

A busca através desses mecanismos nem sempre retorna contextos definitórios, que são aqueles mais fáceis de serem incluídos no banco de dados. A busca por outras colocações pode fornecer pistas para contextos explicativos que, somados, podem criar uma definição.

Mais *corpora*

Terminada a fase acima, decidiu-se pela elaboração de novos *corpora*, dessa vez no campo da Lingüística, para que não houvesse a necessidade de explicar termos pertinentes desse campo na "Ajuda Online" do site em desenvolvimento. A idéia era que houvesse um sistema de metalinguagem. Cada vez que o aluno tivesse uma dúvida sobre um termo do campo da Lingüística que aparecesse na microestrutura do site, bastaria consultar esse termo no próprio site. Para isso, a construção de novos *corpora* se fez necessária.

Todos os passos descritos nos itens anteriores foram realizados novamente e uma nova árvore de campo foi criada. Dessa vez, contudo, não houve a necessidade de

se desdobrar mais do que três subníveis da árvore, já que o objetivo desses *corpora* é diferente. Essa nova árvore ficou configurada como na figura 11.

Assim como nos *corpora* anteriores, esses contam com, no mínimo, vinte mil palavras em cada domínio. O *corpus* em português totalizou 1.309.967 palavras e o *corpus* em inglês totalizou 1.921.811 palavras.

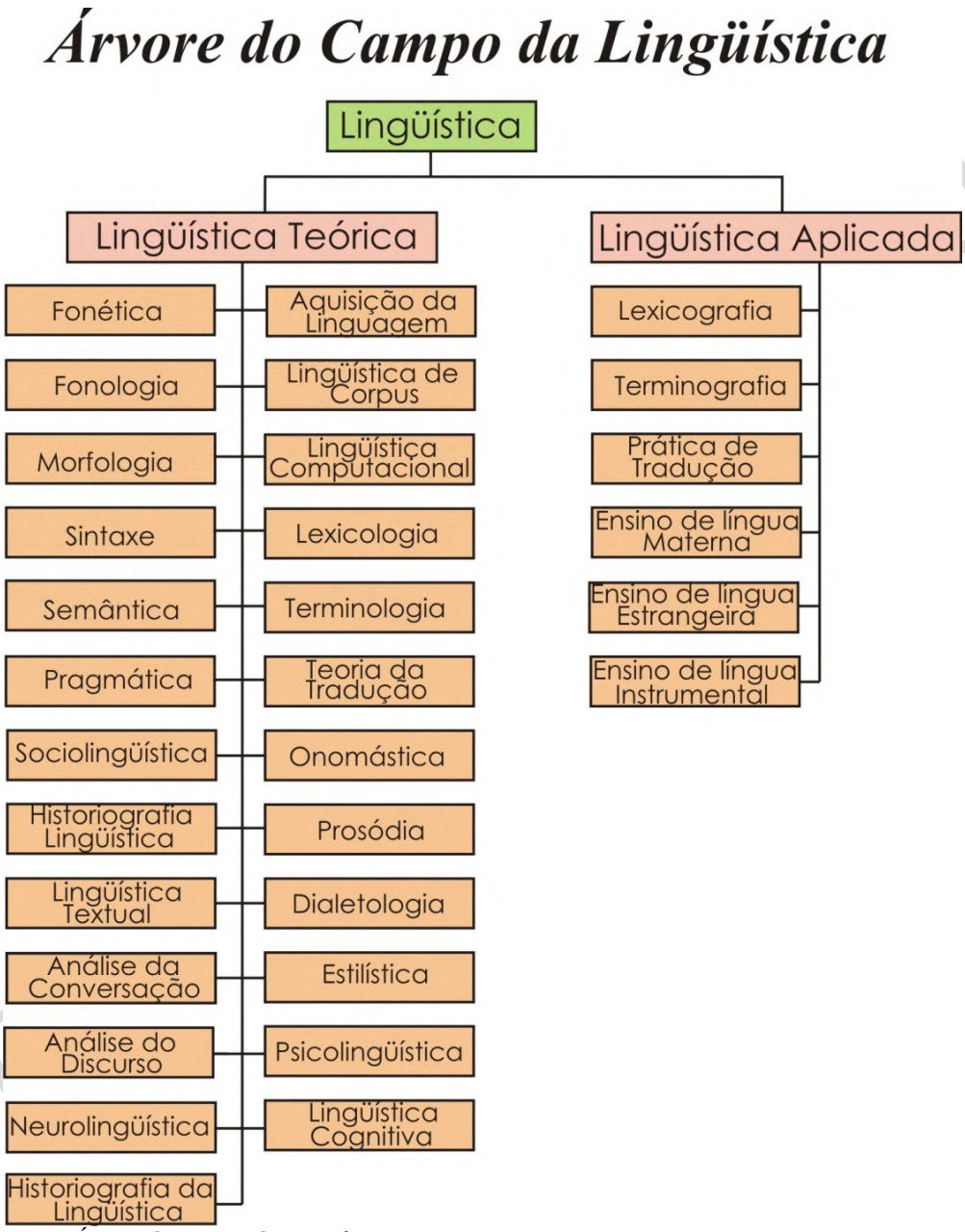


Figura 11. Árvore do Campo da Lingüística

Em suma

O projeto prevê que os *corpora* construídos para alimentar o banco de dados sejam dinâmicos, isto é, novos textos e áreas (com os respectivos domínios e subdomínios) poderão ser acrescentados no futuro para aumentar sua precisão e escopos de análise. O aumento do *corpus* implica, porém, a atualização de dados referentes ao *corpus* para cada termo no banco de dados (freqüência no *corpus* e número total de exemplos encontrados).

No momento foram incluídos somente textos escritos, já que os mesmos representam bem os campos técnicos, mas nada impede que futuramente outros tipos de texto (como os orais) sejam adicionados.

É de extrema importância notar que, diferente de algumas ferramentas disponíveis na Internet (como o Corpógrafo), a presente proposta não oferece uma solução de armazenamento do corpus. O mesmo deve ficar disponibilizado no computador do pesquisador. Somente os contextos dele extraídos é que serão armazenados no banco de dados.

Bibliografia

ALMEIDA, G. M. B.; ALUISIO, S. M.; OLIVEIRA, L. H. M. A terminologia na era da informática. **Ciência e Cultura**, v. 58, n. 2. 2006. Disponível em: <http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200016&lng=en&nrm=iso>.

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ciência da Informação**, Brasília: IBCT, v. 32, n. 3., 2003.

AUBERT, F. H. **Introdução à metodologia da pesquisa terminológica bilíngüe**. São Paulo: Humanitas, 1996.

BERBER SARDINHA, A. **Lingüística de *corpus***. Barueri: Manole, 2004.

BIDERMANN, M.T.C. **Teoria Lingüística**. 2. ed. São Paulo: Martins Fontes, 2001.

FROMM, G. **Proposta para um modelo de glossário de informática para tradutores**. São Paulo, 2002. Dissertação (Mestrado em Lingüística). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

HOUAISS, A. **Dicionário eletrônico Houaiss da língua portuguesa**. São Paulo: Objetiva, 2002.

LARA, M. L. G. de; TÁLAMO, M. F. G. M. *Uma experiência na interface Lingüística Documentária e Terminologia***.** In: **DataGramaZero** - Revista de Ciência da Informação - v.8 n.5 out/07. Disponível em: http://www.dgz.org.br/out07/Art_01.htm. Acessado em: 22/07/2008.

MARINOTTO, O. **Para a elaboração de um vocabulário especializado bilíngüe (inglês/português) da linguagem da aviação**: manutenção de aeronaves, controle de tráfego aéreo e operações aéreas. São Paulo, 1995. Tese (Doutorado em Lingüística) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

SCOTT, M. **WordSmith Tools**. Versão 4. Disponível em: <http://www.lexically.net/wordsmith/>. Acesso em 17 junho 2007.

SOWA, J. F. **Building, sharing and merging ontologies**. Tutorial. 1999. Disponível em: < http://www.jfsowa.com/ontology/ontoshar.htm#s6>. Acesso em: 22 abril 2007.

TAGNIN, S. E. O. ***Corpora***: o que são e para quê servem. Minicurso. São Paulo, 2004.