Análise do desempenho de extratores automáticos de candidatos a termos: proposta metodológica para tratamento de filtragem dos dados

Rosana de Barros Silva e Teixeira[*]

Abstract: This article aims to present one aspect of the masters dissertation entitled (Onco)mastology terms: a corpus-mediated approach (2011). This work will explore one of the goals that guided the study, namely, verifying the success rates of four computational tools for automatic extraction of term candidates: Corpógrafo 4.0, WordSmith Tools 3.0, e-Termos and ZExtractor. Two corpora were used in the investigation: the study corpus (MAMAtex), with a total of 563,482 words, and a reference corpus (Banco de Português 1.0), with 125,927,624 words. The first, which is specialized, consists of some of the genres of scientific discourse, of scientific dissemination and instruction in (Onco)mastology, while the second, a general-language text, includes various genres. Two approaches were chosen to support this analysis from the theoretical and methodological standpoint: the Communicative Theory of Terminology (CABRÉ 1993) and Corpus Linguistics (SINCLAIR 1991; BERBER SARDINHA 2004, 2005). As revealed by the data, Corpógrafo 4.0 ranks highest, with 27.56% accuracy, followed by ZExtractor (26.05%), WordSmith Tools 3.0 (21.77%) and e-Terms (14.44 %). In order to make feasible the examination of candidates, given that the lists generated by the programs included thousands of words, a methodology was developed using Microsoft Office Excel 2007 for filtering candidates common to all the tools and unique to each one. This cut in the data served as a possibly feasible "methodological shortcut" for optimizing the selection of term candidates from lists processed by two or more programs.

Keywords: Terminology - Corpus Linguistics - Computational tools - Automatic extraction of term candidates.

[*] Mestra em Linguística Aplicada e Estudos da Linguagem pela Pontifícia Universidade Católica de São Paulo (PUC-SP).

Resumo: Este artigo pretende apresentar um aspecto da dissertação de mestrado intitulada Termos de (Onco)mastologia: uma abordagem mediada por corpus (2011). Nesta ocasião, explorarei um dos objetivos que norteou a pesquisa, qual seja, a verificação do índice de acerto de quatro ferramentas computacionais para extração automática de candidatos a termo: Corpógrafo 4.0, WordSmith Tools 3.0, e-Termos e ZExtractor. Dois corpora prestaram-se à investigação: o de estudo (MAMAtex), que totaliza 563.482 palavras, e o de referência (Banco de Português 1.0), com 125.927.624 palavras. O primeiro, especializado, é composto de alguns dos gêneros dos discursos científico, de divulgação científica e instrucional da (Onco)mastologia, enquanto o segundo, de linguagem geral, compreende gêneros discursivos variados. Para subsidiar a análise do ponto de vista teórico-metodológico, foram eleitas duas abordagens, a Teoria Comunicativa da Terminologia (CABRÉ 1993) e a Linguística de Corpus (SINCLAIR 1991; BERBER SARDINHA 2004, 2005). Conforme apontam os dados, o Corpógrafo 4.0 lidera o ranking, com 27,56% de acerto, seguido, respectivamente, pelo ZExtractor (26,05%), WordSmith Tools 3.0 (21,77%) e e-Termos (14,44%). Com vistas a tornar factível o exame dos candidatos, posto que as listas geradas pelos programas abrangiam milhares de palavras, foi desenvolvida uma metodologia com o auxílio do Microsoft Office Excel 2007 para filtragem dos candidatos comuns entre todas as ferramentas e exclusivos de cada uma. Esse recorte nos dados funcionou como um "atalho metodológico" possivelmente viável no sentido de otimizar a seleção de candidatos a termo a partir de listas processadas por dois ou mais programas.

Palavras-chave: Terminologia - Linguística de Corpus - Ferramentas computacionais - extração automática de candidatos a termo.

Introdução

Ferramentas computacionais para análise de *corpus* textual parecem ter auxiliado cada vez mais o trabalho (árduo) de pesquisadores de linguagens especializadas na última década. Algumas pesquisas voltadas a questões terminológicas atestam o uso recorrente e crescente de apoio informatizado: TEIXEIRA (2005, 2008a, 2008b); FROMM (2004, 2008); MATUDA (2009); MOREIRA (2010), para citar apenas alguns exemplos. Entre os recursos disponibilizados por aplicativos de análise lexical, estão os extratores automáticos de candidatos a termo. Trata-se de ferramentas que, baseadas em abordagens de ordem estatística, linguística ou concebidos de forma híbrida, retornam uma lista de possíveis termos da área de especialidade.

É sabido que ferramentas híbridas[1] são as mais eficazes para extração de candidatos verdadeiro-positivos (termos, propriamente). Isso porque extratores estatísticos apresentam uma porcentagem de ruído em torno de 75% (BAGOT 2001), enquanto nos linguísticos essa porcentagem pode variar entre 55% e 75% (BAGOT 1999 apud LOPES et al. 2010).

Entretanto, se esse tipo de extrator pode apresentar acuidade até 20% mais precisa que o estatístico, por outro lado ele impõe uma limitação: aplica-se a uma única língua ou mesmo uma variante desta, já que é necessário um estudo linguístico prévio para levar a cabo essa abordagem. Por isso, mecanismos de extração automática que combinam os dois métodos tendem a apresentar os melhores resultados.

Para a análise que trago à tona, foram cotejados quatro programas de natureza estatística — Corpógrafo 4.0[2] (MAIA et al. 2005), e-Termos[3] (OLIVEIRA

[1] Segundo OLIVEIRA (2009), na abordagem estatística, é a frequência de ocorrência da palavra que funciona como vetor de seleção dos candidatos; na linguística, informações morfológicas, morfossintáticas, sintáticas, semânticas e pragmáticas contribuem para seleção dos termos candidatos; na híbrida há o casamento das abordagens anteriores.

[2] Disponível em <http://193.137.34.101/ferramentas/gc/>.

2009), ZExtractor[4] (LOPES 2009) e WordSmtih Tools 3.0[5] (SCOTT 1999) — por dois motivos: os três primeiros por serem gratuitos, o que tende, em tese, a contribuir para o aumento da demanda por parte dos pesquisadores, e, o último, por conter o programa de extração de palavras-chave.[6]

Sendo assim, o primeiro passo foi processar, nos programas mencionados, os *corpora* de estudo (de análise) e, quando possível, o de referência (de contraste). Em seguida, os candidatos (unigramas) comuns a todos os programas e específicos de cada um foram filtrados; na sequência, cada candidato foi submetido a um concordanciador para que pudessem ser constatados tanto o estatuto de termo da (Onco)mastologia[7] quanto a acuidade de cada ferramenta no quesito índice de acerto terminológico.

Nos próximos itens, procurarei explicitar cada um desses passos.

[3] Disponível em <http://www.etermos.cnptia.embrapa.br/index.php>.

[4] De autoria de José Lopes Moreira Filho. Disponível em <http://www.x*corpus*.net/downloads/ZExtractor-ptbr.zip>.

[5] De Mike Scott. Oxford. Oxford University Press. Disponível em <http://www.lexically.net/wordsmith>, o programa encontra-se na versão 5.0 e, embora não tenha sido concebido para pesquisa terminológica, muitos pesquisadores da área o têm utilizado para esse fim, o que vem justificar uma análise de sua *performance*.

[6] Palavras-chave caracterizam um *corpus* sob alguns aspectos, como o das escolhas lexicais típicas, ponto de vista ao qual me ative.

[7] A (Onco)mastologia encontra-se no entroncamento da Oncologia com a Mastologia. Trata-se de uma (sub)área que está, segundo especialistas, em via de ser oficializada — por isso os parênteses , posto que já nomeia cursos de especialização no setor de Mastologia da Escola Paulista de Medicina (Unifesp), entre outras, além de congressos e clínicas especializadas. Essa (sub)área vem ganhando visibilidade (inclusive na mídia) devido ao crescente número de casos de câncer de mama (INCA); daí a motivação para a pesquisa terminológica.

1 Constituição dos *corpora*

Dois *corpora* foram utilizados para análise das ferramentas: o de estudo (MAMAtex), de cunho especializado, e o de referência (Banco de Português 1.0),[8] de caráter geral.

Abrangendo 563.482 palavras, o *corpus* de estudo, assim denominado por compreender a amostra de dados efetivamente utilizada para análise, é composto de textos escritos oriundos de gêneros dos discursos científico, de divulgação científica e instrucional (ALMEIDA 2006), tais como: artigos, dissertações, laudos, teses, entrevistas, fôlderes, notícias, resumo, livros acadêmicos e manual de medicina. É sincrônico (compreende a década de 1998 a 2008) e de amostragem (possui uma amostra finita da linguagem da (Onco)mastologia).

O *corpus* de referência, por sua vez, integra 125.927.624 palavras. Empregado para fins de contraste com o *corpus* de estudo, seu tamanho supera o mínimo estabelecido[9] para o propósito em questão. No caso do Banco de Português 1.0, o conteúdo compreende textos escritos e transcritos da fala; é orgânico, pois está em constante compilação, e diversificado do ponto de vista dos gêneros discursivos: é composto de conversas, entrevistas, aulas, reuniões, conversas telefônicas, notícias, editorial, reportagem, artigos, teses, documentos de negócios etc.

[8] No momento em versão 2.0, o Banco de Português encontra-se disponível em CD somente para alunos do departamento de Linguística Aplicada e Estudos da Linguagem (LAEL) da Pontifícia Universidade Católica de São Paulo (PUC-SP).

[9] Para mais informações, vide BERBER SARDINHA (2005).

2 Procedimentos metodológicos

Os quatro programas analisados partilham determinados traços, ao mesmo tempo em que apresentam aspectos que os diferenciam. Entretanto, como não faz parte do escopo deste artigo retratar as generalidades e as peculiaridades de cada um detidamente, exponho, na tabela 1, alguns pontos convergentes e divergentes entre eles a fim de oferecer uma visão geral das potencialidades que os caracterizam.

aspectos	**CORPÓGRAFO 4.0**	**WST 3.0**	**E-TERMOS**	**ZEXTRACTOR**
precisão gráfica (caracteres alfanuméricos)	SIM	NÃO	SIM	NÃO
ajuste de parâmetros estatísticos: valor de corte/ log likelihood[10]	NÃO	SIM	SIM/NÃO	SIM
palavras-chave (comparação estatística entre corpora)	NÃO	SIM	NÃO	SIM
extração precisa de n-gramas	NÃO	SIM	SIM	SIM
interface gráfica	SIM	SIM	SIM	SIM

Tabela 1 – Descrição parcial das características constituintes dos programas

Com base nessas características, adentrarei, a seguir, nos procedimentos efetuados em cada ferramenta para obtenção da listas de candidatos a termo.

[10] *Log-likelihood* é, grosso modo, a medida estatística que determina o valor de significância (*p value*). Para mais detalhes sobre este item, vide seção 7.1. Nas ciências sociais o mínimo aceitável para *log-likelihood* é 0,5.

2.1.1 Primeiro passo: geração automática das listas de candidatos a termo

Embora alguns dos procedimentos efetuados nos programas para geração das listas coincidam, julguei conveniente dividir as etapas de trabalho de acordo com cada um, uma vez que, caso o pesquisador já tenha em mente quais aplicativos vai utilizar, o seguimento da metodologia proposta tornar-se-á facilitado.[11]

Cabe frisar, de antemão, que um valor de corte (sete) foi determinado no sentido de tornar factível a metodologia. Para tanto, segui as orientações de BAGOT (1999) *apud* LOPES et al. (2010), segundo a qual o tamanho do *corpus* é levado em conta a partir da fórmula: valor do corte = (<tamanho do *corpus*>/100.000 + 1).

No entanto, saliento que existem outros cálculos por meio dos quais é possível estipular um valor de corte, como a medida F (*F-measure*), resultado do equilíbrio entre precisão (capacidade do programa de identificar candidatos verdadeiro-positivos) e abrangência (quantidade de candidatos verdadeiro-positivos que, de fato, o programa extraiu), considerando-se os índices de frequência absoluta e relativa dos dados.[12]

2.1.2 Corpógrafo 4.0

Para que a lista de candidatos a termos seja gerada por esta ferramenta, após o carregamento do *corpus* (1), é necessário gerar uma base de dados terminológica (2). A seguir, aciona-se a execução da lista de

[11] Para detalhamento dos programas e do passo a passo de inserção de dados neles, vide dissertação de mestrado *Termos de (Onco)mastologia: um abordagem mediada por corpus* em: <http://www4.pucsp.br/pos/lael/lael-inf/def_teses.html.>.

[12] Para mais detalhes e exemplos, vide LOPES et al. 2009 e LOPES et al. 2010).

frequência das palavras (candidatos) (3). Os candidatos (com frequência ≥ 7, neste caso) devem ser copiados na folha 1 de uma planilha do Microsoft Office Excel[13] (4). Nela, candidatos do tipo composto, como "autoexame",[14] ou complexo, tal qual "linfonodo sentinela", precisam ser desmembrados em unigramas (uma lexia):[15] menu "Dados" > subgrupo "Ferramentas de Dados" > botão "Texto para colunas", onde deve-se clicar no campo "Delimitado" e, depois, em "Avançar" (5). Itens duplicados (*linfonodo* sentinela/ *linfonodo* axilar) devem ser eliminados (6), de modo que a lista final componha-se somente de unigramas. Para tano, basta ir ao menu "Início" > subgrupo "Estilo" > botão "Formatação Condicional" > "Realçar Regras das Células" > "Valores duplicados".[16]

2.1.3 WordSmith Tools 3.0 (WST)

Vale esclarecer, previamente, que a suíte WordSmith Tools 3.0 agrega três programas: *WordList, KeyWords* e *Concordance*. Cada um deles foi utilizado de acordo com as necessidades emergentes da pesquisa.

Primeiramente, os *corpora* de estudo e de referência foram carregados (1) para que o programa *WordList* gerasse uma lista de frequência de palavras de cada um dos *corpora*. (2). A seguir, o parâmetro do *KeyWords* foi ajustado para ≥ 7 (3). Do cruzamento da lista de frequência de palavras dos dois *corpora* pelo programa (4), uma lista de palavras-chave (unigramas) foi

[13] Para que a análise dos dados seja levada a cabo, é aconselhável efetuar as operações propostas neste artigo em várias planilhas. Aqui, por motivos de didaticidade, concentrei todas as etapas metodológicas em uma só.

[14] Cumpre esclarecer que à época em que a compilação do *corpus* de estudo foi concluída (2008), a grafia desta palavra ainda era feita com hífen, diferentemente de hoje, em decorrência do Novo Acordo Ortográfico aprovado em 2009.

[15] Embora 70% das terminologias sejam complexas (KRIEGER e FIINATTO 2004), parti de unigramas por acreditar que, juntando os termos um a um, nenhum dado seria perdido.

[16] Para detalhamento completo dessas etapas, vide dissertação de mestrado *Termos de (Onco)mastologia: uma abordagem mediada por corpus* em <http://www4.pucsp.br/pos/lael/lael-inf/def_teses.html>.

apresentada (5). As positivas[17] resultantes foram selecionadas e coladas na planilha do Excel (folha 2) (6).

2.1.4 e-Termos

Após carregamento do *corpus* de estudo nesta ferramenta (1), acionei a compilação dele (2) por ser esta uma exigência do programa. Na sequência, ajustei o parâmetro para frequência ≥ 7 (3). Uma lista de candidatos (unigramas) foi gerada (4). Esses candidatos foram copiados para a planilha (folha 3) (5). Como alguns dos candidatos listados eram compostos, a exemplo de "autoexame", eles foram desmembrados em um único item (6): menu "Dados" > subgrupo "Ferramentas de Dados" > botão "Texto para colunas", no qual deve-se clicar no campo "Delimitado" e, depois, em "Avançar". Itens duplicados foram eliminados (7), de modo que a lista final apresentasse somente candidatos simples, ou seja, unigramas: menu "Início" > subgrupo "Estilo" > botão "Formatação Condicional" > "Realçar Regras das Células" > "Valores duplicados".

2.1.5 ZExtractor

Iniciei os procedimentos ajustando a frequência para ≥ 7 (1). A seguir, carreguei no programa os *corpora* de estudo e de referência (2), para que, após cotejamento de ambos (3), uma lista de palavras-chave (unigramas) fosse produzida (4). Como nos demais programas, selecionei essa lista de unigramas (candidatos) e a copiei para a folha 4 da planilha (5).

[17] O conceito de palavra-chave positiva nada tem a ver com palavra importante. Ele diz respeito àquelas cuja frequência é maior no *corpus* de estudo que no de referência.

3 Segundo passo: extração dos candidatos comuns[18] a todas as listas

Nesta etapa do trabalho, fiz uso da função PROCV,[19] disponível no Microsoft Office Excel 2007 ("Fórmulas" > subgrupo "Biblioteca de Funções" > botão "Pesquisa e Referência" > PROCV). Esse recurso localiza um determinado valor (dado) em uma lista de origem. Com o auxílio dele, foi feita uma varredura nos dados, de modo que os candidatos comuns a todos os programas pudessem ser extraídos e colados na folha 5 da planilha do Excel.

A título de certificação, uma contraprova foi realizada com o objetivo de garantir que os dados resultantes da filtragem eram, de fato, comuns a todos os programas.[20] Para esta finalidade fiz uso também da função "SE": menu "Fórmulas" > subgrupo "Biblioteca de Funções" > botão "Lógica"> "SE".

4 Terceiro passo: extração dos candidatos exclusivos de cada lista[21]

Posteriormente, utilizando-me da mesma função PROCV (vide seção anterior), busquei selecionar os candidatos exclusivos de cada programa.

[18] Parti dos candidatos comuns por considerar que, dentre esses, provavelmente estaria a maioria dos termos (candidatos verdadeiro-positivos).

[19] O correspondente em inglês é VLOOKUP.

[20] Para mais detalhes, vide dissertação em: <http://www4.pucsp.br/pos/lael/lael-inf/def_teses.html>.

[21] Resolvi investigar também os exclusivos para saber o que cada ferramenta em particular traria como termo (candidato verdadeiro-positivo).

Da mesma forma como foi feito com os candidatos comuns, uma contraprova foi realizada com os (supostamente)[22] exclusivos, a fim de atestar se os dados resultantes da varredura eram fidedignos.

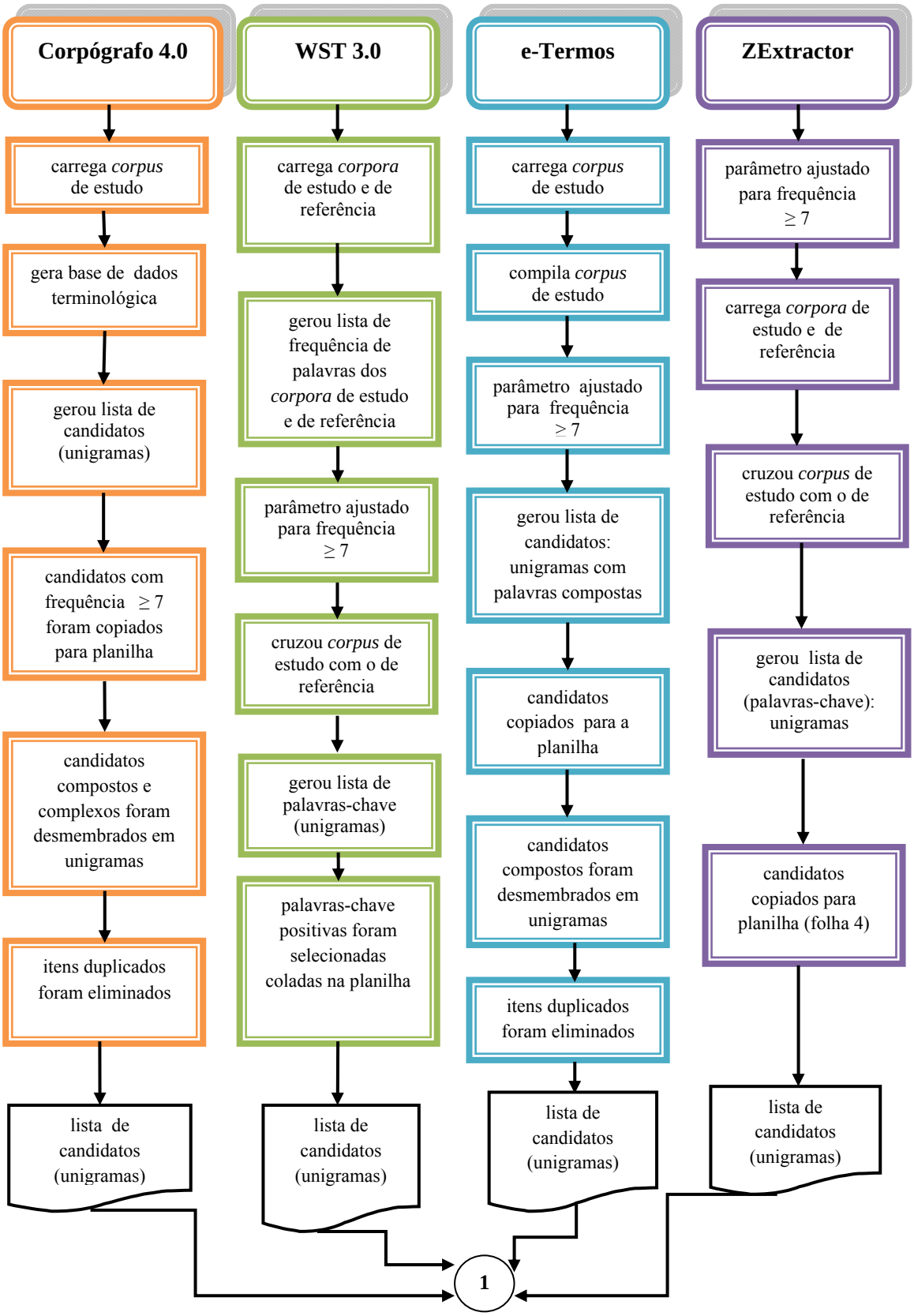
Além disso, no intuito de comprovar se os candidatos filtrados eram exclusivos *de fato*, procedi a uma análise qualitativa dos dados. Isso porque, dependendo do programa — vide primeiro aspecto exposto na Tabela 1 —, determinados caracteres são omitidos. Dessa forma, a representação do mesmo candidato, por vezes, deu-se de maneira gráfica diferente conforme o aplicativo.

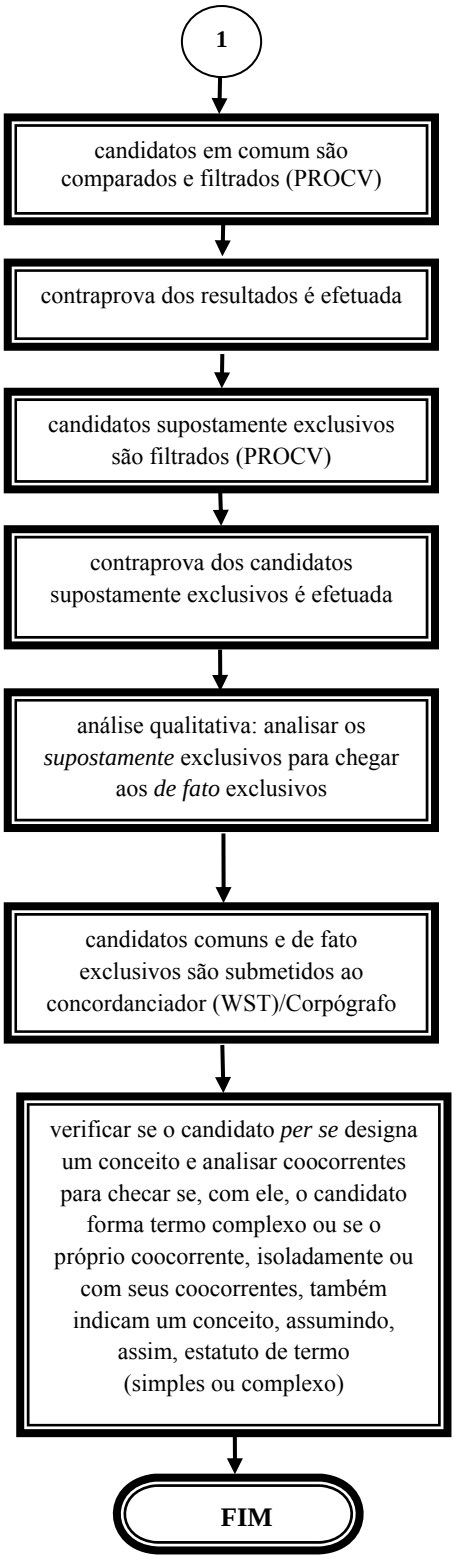
Por exemplo: o candidato "BRCA1" fora listado como presente unicamente na lista do Corpógrafo; contudo, ao buscar por somente "BRCA" nas listas dos outros programas, este foi localizado e, uma vez submetido à condição de palavra de busca no concordanciador, o item "1" apareceu como colocado[23] de "BRCA" (que se apresentou no *corpus* de estudo ora separado de 1 por espaço, ora por hífen). Portanto, o candidato "BRCA1" não era, com efeito, privilégio do Corpógrafo; ao contrário, foi identificado em todas as listas representado por "BRCA" e passou a figurar na lista dos candidatos comuns a todos os programas que, ao final da análise qualitativa, contou com 728 ocorrências.

A fim de retomar as operações executadas, o fluxograma, na página seguinte, sintetiza os passos trilhados até o momento, enquanto a tabela 2, na sequência, mostra os resultados obtidos.

[22] Os parênteses são para indicar que a exclusividade não fora ainda qualitativamente comprovada.

[23] Refiro-me a *colocado*, por ser esta a nomenclatura usada no programa WordSmith Tools 3.0, embora se trate de concorrente, já que colocação é uma característica linguística cuja comprovação é feita por meio de medidas estatísticas de associação, como Razão Observado/Esperado, *T-score* e/ou *Mutual Information*.


Corpógrafo 4.0
carrega *corpus* de estudo
gera base de dados terminológica
gerou lista de candidatos (unigramas)
candidatos com frequência ≥ 7 foram copiados para planilha
candidatos compostos e complexos foram desmembrados em unigramas
itens duplicados foram eliminados
lista de candidatos (unigramas)
WST 3.0
carrega *corpora* de estudo e de referência
gerou lista de frequência de palavras dos *corpora* de estudo e de referência
parâmetro ajustado para frequência ≥ 7
cruzou *corpus* de estudo com o de referência
gerou lista de palavras-chave (unigramas)
palavras-chave positivas foram selecionadas coladas na planilha
lista de candidatos (unigramas)
e-Termos
carrega *corpus* de estudo
compila *corpus* de estudo
parâmetro ajustado para frequência ≥ 7
gerou lista de candidatos: unigramas com palavras compostas
candidatos copiados para a planilha
candidatos compostos foram desmembrados em unigramas
itens duplicados foram eliminados
lista de candidatos (unigramas)
ZExtractor
parâmetro ajustado para frequência ≥ 7
carrega *corpora* de estudo e de referência
cruzou *corpus* de estudo com o de referência
gerou lista de candidatos (palavras-chave): unigramas
candidatos copiados para planilha (folha 4)
lista de candidatos (unigramas)
1


1
candidatos em comum são comparados e filtrados (PROCV)
contraprova dos resultados é efetuada
candidatos supostamente exclusivos são filtrados (PROCV)
contraprova dos candidatos supostamente exclusivos é efetuada
análise qualitativa: analisar os *supostamente* exclusivos para chegar aos *de fato* exclusivos
candidatos comuns e de fato exclusivos são submetidos ao concordanciador (WST)/Corpógrafo
verificar se o candidato *per se* designa um conceito e analisar coocorrentes para checar se, com ele, o candidato forma termo complexo ou se o próprio coocorrente, isoladamente ou com seus coocorrentes, também indicam um conceito, assumindo, assim, estatuto de termo (simples ou complexo)
FIM

Total de candidatos após filtragem	CORPÓGRAFO	WST	E-TERMOS	ZEXTRACTOR
exclusivos (de fato)	12	241	705	51

Tabela 2 - Contagem de candidatos exclusivos de cada programa

5 Quarto passo: submissão dos candidatos comuns e exclusivos ao concordanciador

De posse dos candidatos comuns a todos os programas (728) e de fato exclusivos de cada um, a etapa subsequente foi submetê-los, um a um, ao concordanciador do programa *Concordance*, da suíte WordSmith Tools 3.0, ou do Corpógrafo 4.0 para constatar se designavam um conceito da sub(área) em questão, ou seja, se possuíam "um valor singularmente específico" (CABRÉ 1999: 124).

Nesse processo de submissão, foram geradas listas de concordâncias[24] da palavra de busca (nódulo) — no caso, os candidatos a termo. Elas são imprescindíveis para a análise em contexto das palavras que coocorrem (SINCLAIR 1991) com os candidatos, propiciando, adiante, a detecção de termos complexos graças a medidas estatísticas de associação, como Razão Observado/Esperado, Escore T (*T-Score*) *ou Mutual Information.*[25]

Outros termos (simples ou complexos) que figuraram na condição de coocorrente[26] dos candidatos da lista também puderam ser apurados via linha de concordância.

[24] As linhas de concordância correspondem a uma lista de sentenças que contêm e trazem, alinhadas ao centro, a palavra de busca (nódulo).

[25] Para detalhes, vide BERBER SARDINHA (2004).

[26] Friso que o valor de corte não impede que se encontrem, no momento da pesquisa em linhas de concordância, outros candidatos a termo via coocorrência. Isso porque determinadas palavras, ainda que não tenham atingido o valor de corte, podem manifestar indícios conceituais que põem à vista seu estatuto terminológico. No caso do ramo das ciências médicas ao qual o *corpus* de estudo estava vinculado, propriedades semânticas de sufixos

Ao final, foi possível indicar o percentual de acerto das ferramentas e, a partir dos termos apurados, proceder a um glossário de termos da (Onco)mastologia.

Na próxima seção, exponho, na tabela 3, a contagem de candidatos verdadeiro-positivos (termos) e falso-positivos (não-termos) por programa.

6 Apresentação dos resultados

Os apontamentos feitos nesta seção parecem indicar o quão viável pode ser o cumprimento das quatro etapas metodológicas esboçadas anteriormente, tendo em vista a deficiência das ferramentas no quesito por ora explorado: índice de acuidade no que tange à extração de termos, especificamente.

A tabela 3 e os gráficos 1 e 2 apresentam o desempenho atingido por cada ferramenta.

Candidatos	CORPÓGRAFO	WST	E-TERMOS	ZEXTRACTOR
verdadeiro-positivos	204	211	207	203
falso-positivos	536	758	1.226	576
Total de candidatos (comuns + exclusivos)	740	969	1.433	779

Tabela 3 - Contagem de candidatos verdadeiro-positivos e falso-positivos por programa sobre o total resultante da soma de candidatos comuns com exclusivo.

Graficamente e em termos percentuais, o resultado obedece a esta escala de valores:

(-ectomia,-oma) e radicais (terapia, grafia) eruditos de origem grega, em conjunto com outros morfemas de natureza similar (oofor-) ou latina (ultra-) (ALVES 2006) foram responsáveis pela nomeação de processos por meio de derivação e composição: ooforectomia, carcinoma, hormonioterapia, tomossínte, entre tantos outros. Para Delvizio; Barros (2008: 1409), "na terminologia médica é comum que a estrutura morfológica do termo deixe transparecer traços do fenômeno designado". Por isso, vali-me também desses recursos linguísticos como "pista" para encontrar os termos da (Onco)mastologia, além dos critérios já descritos.

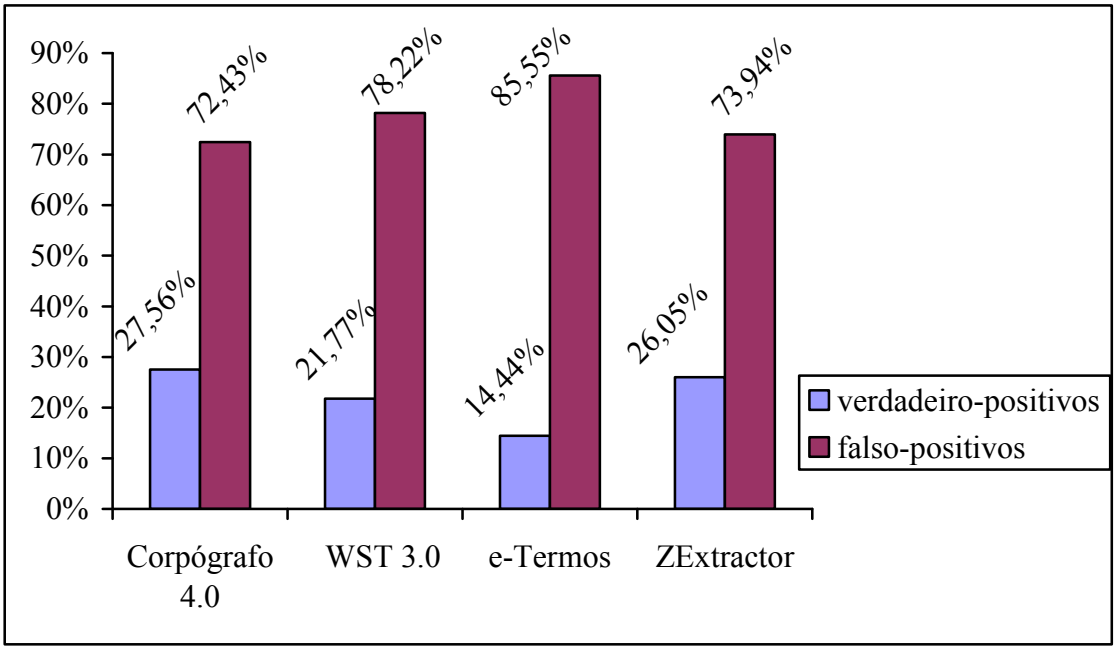


Gráfico 1 - Índice percentual de acerto e de erro das ferramentas investigadas

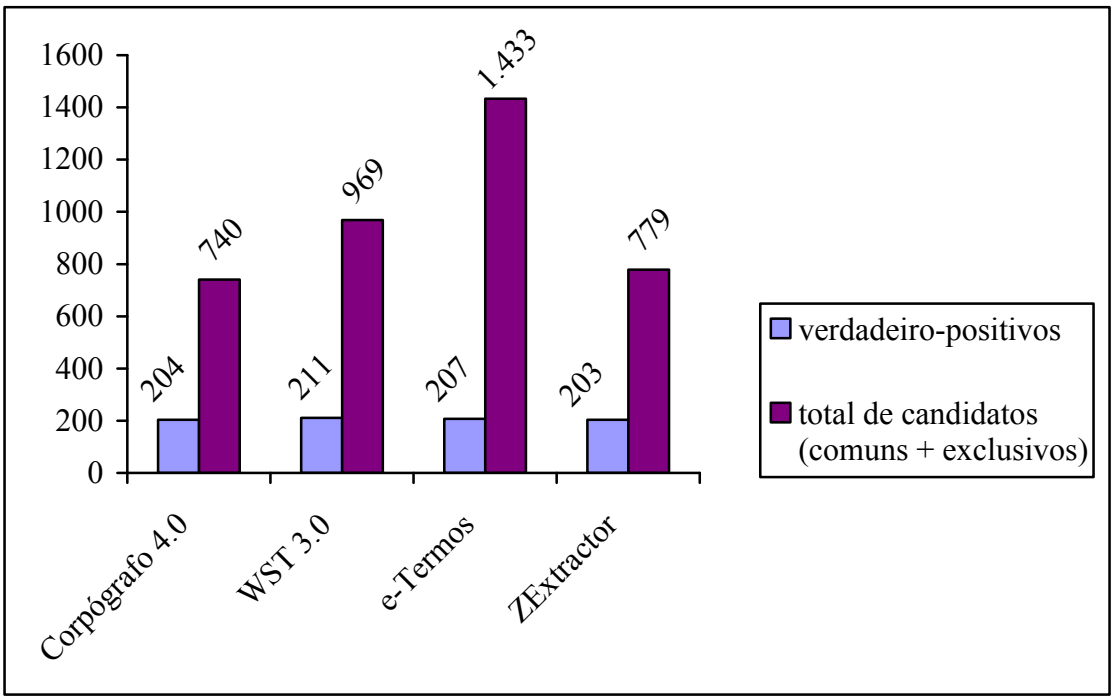


Gráfico 2 – Total de candidatos verdadeiro-positivos (termos) sobre o total de candidatos (comuns + exclusivos) arrolados por programa

6.1 Teste estatístico

A fim de comprovar que a diferença entre os valores atingidos pelos programas poderia ser considerada relevante, apliquei o teste estatístico do qui-quadrado (x^2). Ele baseia-se na fórmula:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

em que *O* equivale à frequência observada; *E* corresponde à frequência esperada; *Σ* indica somatória; *i*, condição e *j*, grupo.

A figura 1, a seguir, reproduz a tela da calculadora *online* <http://www.people.ku.edu/~preacher/chisq/chisq.htm>, em que os valores expostos na tabela 3 foram inseridos.

	Gp 1	Gp 2	Gp 3	Gp 4	Gp 5	Gp 6	Gp 7	Gp 8	Gp 9	Gp 10	
Cond. 1:	204	211	207	203							825
Cond. 2:	740	969	1433	779							3921
Cond. 3:											0
Cond. 4:											0
Cond. 5:											0
Cond. 6:											0
Cond. 7:											0
Cond. 8:											0
Cond. 9:											0
Cond. 10:											0
	944	1180	1640	982	0	0	0	0	0	0	4746

Output:

Calculate | Reset all

Chi-square:	45.232
degrees of freedom:	3
p-value:	0
Yates' chi-square:	44.349
Yates' *p*-value:	0

Status: Status okay

Figura 1 - Tela da calculadora online com o resultado do teste estatístico de comparação (qui-quadrado)

A partir do qui-quadrado, medida estatística de comparação que testa a associação significativa entre variáveis, foi possível saber se havia diferença significativa entre os valores obtidos. Para tanto, o resultado do valor de significância (*p-value*) precisaria ser inferior a 0,05.

Como é possível observar na figura 1, o resultado do *p-value* foi igual a zero; portanto, existe diferença, do ponto de vista estatístico, entre os programas e também entre os termos verdadeiro-positivos e falso-positivos.

7 Discussão dos resultados

Pelo gráfico 1, nota-se que três das quatro ferramentas utilizadas (Corpógrafo 4.0, WST 3.0 e ZExtractor) apresentaram índice de acerto (candidatos verdadeiro-positivos) na faixa dos 20% e índice de erro (candidatos falso-positivos) na faixa dos 70%.

Destes para o e-Termos, que ficou em quarto lugar no *ranking* de acuidade, há uma queda de, aproximadamente, sete pontos percentuais no que se refere ao índice de acerto, considerando-se o programa que atingiu a terceira posição, o WST 3.0, e uma elevação do mesmo percentual no que concerne ao índice de erro em relação ao mesmo programa (WST 3.0).

No gráfico 2, a percepção do índice de acerto das ferramentas fica ainda mais visível ao se aferir os resultados sobre o total de candidatos inventariados pelos aplicativos — a partir do que foi, como outrora mencionado, comum entre todos e exclusivos de cada um.

Esses números sugerem que, enquanto o Corpógrafo 4.0, o WST 3.0 e o ZExtractor mantêm uma média de acerto e de erro, o e-Termos ficou aquém do previsto no quesito acerto de candidatos verdadeiro-positivos.

Entretanto, o índice de acuidade dos aplicativos que ocuparam os três primeiros lugares é considerado baixo se comparado a outros tipos de ferramentas estatísticas, como etiquetadores morfossintáticos, que possuem índice de confiabilidade acima de 90%.

Segundo BICK (2007), o PALAVRAS, por exemplo, etiquetador morfossintático para Língua Portuguesa que também atribui características semânticas às palavras — o que a torna mais complexa — possui mais de 80% de confiabilidade até o momento; logo, esses dados ensejam que ainda há muito por fazer quando se refere à precisão de ferramentas voltadas à extração de candidatos a termo.

8 Considerações finais

Retomando o esquema metodológico desenvolvido nas páginas anteriores, convém dizer que se o caminho em busca de termos via *corpora* não parece convergir para uma única direção, a da plena eficiência de uma única ferramenta computacional, o uso combinado de dois ou mais programas pode ser proveitoso, se somar recursos.

Ora, posto que o manejo de *corpora* também deve estar aliado a conhecimentos gerais sobre a área que se pretende examinar, uma forma de otimizar o acesso a eles pode ser através do "atalho metodológico" proposto aqui: de posse de noções do domínio, compara-se dois ou mais aplicativos com base nos candidatos comuns e exclusivos extraídos.

A despeito da perda de dados em que se pode incorrer, pois o que não é comum a todos os programas nem específico de cada, mas encontra-se em somente três deles, por exemplo, é descartado, cumpre argumentar que ela não impede de chegar a um número substancial de termos,[27] se o *corpus* de

[27] No caso da dissertação de mestrado aqui citada, foram localizados 237 termos.

estudo possuir dimensão, no mínimo, média, segundo a classificação proposta por BERBER SARDINHA (2004: 26).

Vale frisar ainda que, caso a comparação se der somente entre dois aplicativos, essa perda será inexistente e, se a opção for por três ferramentas, ela seria minimizada, fato que, além de favorecer a credibilidade dessa metodologia, em adição, parece validá-la.

Acrescenta-se a isso o fato de os procedimentos metodológicos sugeridos neste artigo não exigirem do pesquisador conhecimentos aprofundados de programação, como elaboração linhas de *scripts*, por fazer uso de um programa (Excel) que, além de integrar o sistema operacional Windows, plataforma francamente popularizada, possui outras vantagens, como interface gráfica e um tutorial riquíssimo, com exemplos dos recursos disponíveis.

Outro aspecto que merece atenção é o ponto de corte que, na pesquisa da qual deriva este artigo, esteve baseado em BAGOT (1999 apud LOPES et al. 2010) e considera como resultado o tamanho do *corpus*/100.000 + 1.

Ao analisar os dados, observei que o valor obtido a partir dessa fórmula pode ser aplicável a unigramas, mas caso se desejasse partir de bi, trigramas[28] etc. seria prudente ter como referência a medida F (*F-measure*). Notei que, de forma geral, quanto maior o tamanho do termo, menor a sua frequência (com exceção de "câncer de mama", obviamente, por ser este o tema em torno do qual o *corpus* de estudo se formou).

Por exemplo: enquanto o termo simples "biópsia" apresentou frequência 276, "biópsia cirúrgica", somou 54 ocorrências. Com "biópsia por agulha grossa" a frequência caiu ainda mais, atingindo somente 3. Isso não significa que a frequência seja inversamente proporcional ao número de

[28] É preciso salientar que caso seja usado um extrator de palavras-chave nos moldes apresentados aqui para obter lista de bi ou trigramas, a lista de palavras do *corpus* de referência e do estudo devem estar no mesmo formato, ou seja, ambas as listas devem ser de bi ou trigramas para que o cruzamento dos dados seja feito com êxito.

palavras que compõem o termo complexo, mas foi visível, no *corpus* de estudo, especialmente com aqueles termos que derivaram de um genérico (lexema-base), como hormonioterapia, mastectomia, mamografia, metástase, quimioterapia, radioterapia, entre outros, o fato de a frequência dos termos baixarem à medida que se tornaram complexos — o que parece lógico, se considerarmos que "biópsia" (para citar só um caso) dá origem a várias formações poliléxicas, enquanto "cirúrgica" apenas restringe seu significado.

Revisitando o percurso traçado até aqui, por meio do qual o índice de acerto de quatro programas de extração e estudo do léxico (especializado) foi trazido à baila, espero ter contribuído para as pesquisas em Terminologia de foro descritivo, principalmente no que concerne à melhoria dos aspectos de seleção e análise dos dados.

Referências bibliográficas

ALVES, I. M. A *renovação lexical nos domínios de especialidade*. Ciência e Cultura, São Paulo, v. 58, n.º 2, p. 32-34, 2006.

BERBER SARDINHA, T. *Linguística de Corpus*. Barueri (SP): Manole, 2004, 410 p.

_____. *A influência do tamanho do corpus de referência na obtenção de palavras-chave usando o programa computacional WordSmith Tools*. The ESPecialist, São Paulo, v. 26, n. 2, 2005: 183-204.

BAGOT, R. E. *Extracción de Terminologia: elements per la construcción de un extractor*. TradTerm — Revista do Centro Interdepartamental de Tradução e Terminologia, 7 (1), Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, 2001.

BICK, E. *Automatic Semantic Role Annotation for Portuguese*. In: Proceedings of TIL 2007, V Workshop on Information and Human Language Technology / Anais do XXVII Congresso da SBC, Rio de Janeiro, jul. de 2007: 1713-1716.

CABRÉ, M.T. *La Terminología: teoria, metodologia, aplicaciones*. Barcelona: Editorial Antártida/Empúries, 1993, 521 p.

FROMM, G. *Ferramentas de análise lexical computadorizadas: uma aplicação prática*. Factus, Taboão da Serra 1(3), 2004, p. 153-164. Disponível em <http://www.fflch.usp.br/dlm/comet/>. Acesso em 24/03/2011.

_____. *VoTec: uma ferramenta para terminógrafos, tradutores e alunos de Letras*. In: XI Mini-Enapol, 2008, São Paulo. Tratamentos do Léxico: diversidade cultural, a multiconceptualização do mundo, 2008: 13-13. Disponível em: <http://www.ileel.ufu.br/guifromm/producao_resumos.asp?core=pagina1_sub5>. (24/3/2011).

KRIEGER, M. G.; FINATTO, M. J. B. *Introdução à Terminologia: teoria & prática*. São Paulo: Contexto, 2004, 223 p.

LOPES, L. *et al*. *Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area*. Reciis – Eletronic Journal of Communication, Information & Innovation in Health. Rio de Janeiro, v. 3, n. 1, mar. de 2009: 76-88. Disponível em: <http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/244/259>. (10/2010).

MATUDA, S. *Fraseologia no futebol: um estudo bilíngue baseado em corpus*. Domínios da Linguagem — Revista Eletrônica de Linguística do ILEEL/UFU, n. 4, dez. 2008. Disponível em: <http://www.dominiosdelinguagem.org.br/pdf/09-07-09/Texto%205.pdf>. (24/3/2011).

MOREIRA, A. C. S. *Terminologia e Tradução: criação de uma base dados terminológica do turismo baseada num corpus paralelo português-inglês*. Vigo (Espanha), 2010. Tese de doutoramento. Faculdade de Filologia e Tradução da Universidade de Vigo. 641 p.

SINCLAIR, J. M. *Corpus, Concordance, Collocation*. London: Oxford University Press, 1991, 179 p.

OLIVEIRA, L. H. M. *e-Termos: um ambiente colaborativo web de gestão terminológica*. São Carlos (SP), 2009a, 321 p. (Tese de Doutoramento) Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (IMC-USP).

PREACHER, K. J. (2001, April). *Calculation for the chi-square test: an interactive calculation tool for chi-square tests of goodness of fit and independence* [Computer software]. Disponível em: <http://quantpsy.org>. (26/3/2011).

SILVA E TEIXEIRA, R. B. *Termos de (Onco)mastologia: uma abordagem mediada por corpus*. 2011. Dissertação de mestrado. Pontifícia Universidade Católica de São Paulo.

SINCLAIR, J. M. *Corpus, Concordance, Collocation*. London: Oxford University Press, 1991, 179 p.

TEIXEIRA, E. D. *A Linguística de Corpus a serviço do tradutor: proposta de um dicionário de Culinária voltado para padronização textual*. São Paulo, 2008, 400 p. (Tese de Doutoramento). Programa de Pós-Graduação em Estudos Linguísticos e Literários em Inglês do Departamento de Letras Modernas. Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

_____. *Tradução e Terminologia plurilíngue — a Línguística de Corpus como proposta de aproximação*. In: Grupo de Estudos Linguísticos, 53.° Seminário, 2005. Disponível em <http://www.fflch.usp.br/dlm/comet/artigos/GEL%202005%20Elisa.pdf>. (24/3/2011).

_____. *Tradução culinária e ensino: um exemplo de metodologia de avaliação utilizando etiquetagem e o WordSmith Tools*. Domínios da Linguagem — Revista Eletrônica de Linguística do ILEEL/UFU, n. 4, dez. 2008. Disponível em: <http://www.dominiosdelinguagem.org.br/pdf/09-07-09/Texto%202.pdf>. (24/3/2011).