A QUESTÃO DA TAXONOMIA NUM *CORPUS* COLABORATIVO PARA CONSTRUÇÃO DE UM VOCABULÁRIO NA ÁREA DE LINGUÍSTICA

Guilherme FROMM
Universidade Federal de Uberlândia
guifromm@ileel.ufu.br

Resumo: pretendemos descrever os passos para a elaboração ou reelaboração de uma Árvore de Domínio (ou árvore conceitual) na área de Linguística, representando a taxonomia da mesma. A construção dessa árvore é um passo primordial para a compilação e análise linguística de corpora técnicos nas mais diversas áreas do conhecimento.

Palavras-chave: Árvore de Domínio; Linguística; Linguística de *Corpus*; Terminologia; Terminografia.

1. Introdução

O objetivo deste texto é a descrição do processo de construção de uma taxonomia ("ciência ou técnica de classificação"; HOUAISS, 2009) para a área de Linguística. Esse processo taxonômico é um dos passos básicos para o planejamento e balanceamento de um *corpus* de uma área técnica. No nosso caso, o *corpus* a ser compilado faz parte de um projeto colaborativo entre alunos de graduação e pós-graduação (contando com a supervisão do professor responsável) e temos como objetivo final a construção de um vocabulário (ou dicionário técnico) na área da Linguística.

2. Pressupostos teóricos do projeto geral

O projeto geral visa a construção de verbetes bilíngues, na área de Linguística, a serem disponibilizados gratuitamente na Internet através da plataforma VoTec (disponível em: www.pos.voteconline.com.br). Para chegar ao resultado proposto, o projeto trabalha com os pressupostos teóricos de Cabré (ALMEIDA, 2006) em relação à Teoria Comunicativa da Terminologia, as concepções de dicionário, vocabulário e glossário de Barbosa (2001), os estudos de Ilari (2003), Krieger e Finatto (2004) e Almeida, Pino e Souza (2007) em relação à elaboração das definições dos verbetes, as bases teóricas para o trabalho com a metodologia da Linguística de *Corpus* (BERBER SARDINHA, 2004; TEIXEIRA, 2008; SCOTT, 2012) e a operacionalização da plataforma de gestão terminológica VoTec (FROMM, 2007).

O primeiro passo na construção de um projeto terminográfico moderno é o desenho do *corpus* que vai oferecer subsídios de pesquisa para a construção/exemplificação dos candidatos a termos extraídos desse *corpus*. No caso em questão, algumas decisões foram tomadas nesta fase de planejamento: a. O *corpus* seria na área de Linguística, já que há uma maior proximidade dos alunos com a área; b. Em virtude da característica quase que totalmente acadêmica da área, seriam levantados apenas textos científicos: artigos em periódicos, dissertações e teses; c. Como a proposta do VoTec é bilíngue, os textos teriam que ser levantados nos pares de línguas a serem trabalhadas: inglês e português; d. Cada par de língua e cada subárea deveria contar com um sub*corpus* de quinhentas mil palavras (experiências anteriores nos provaram que esse número é uma quantidade razoável de palavras para a elaboração de termos). Toda essa preocupação na fase de planejamento é essencial para a credibilidade do *corpus*, o qual servirá não só para a pesquisa já mencionada, mas também estará disponível para outros estudos na descrição de línguas.

Como resultado desse desenho, temos a seguinte tipologia para o *corpus* de Linguística:

Tabela 1 - Tipologia do *Corpus* de Linguística (fonte: FROMM; YAMAMOTO, 2013)

Língua	Bilíngue (inglês e português)
Modo	Escrito (textos acadêmicos: artigos científicos, dissertações e teses)
Data de publicação	Sincrônico (levantamento realizado entre 2010 e 2014[1]),
Seleção	Amostragem, Estático
Conteúdo	Especializado (Linguística)
Autoria	Falantes nativos/não nativos (inglês e português), individual/coletivo
Disposição Interna	Comparável
Uso na pesquisa	Estudo (análise terminológica/terminográfica)
Tamanho	Grande (mais de 10 milhões de palavras)
Nível de Codificação	Sem cabeçalhos, sem etiquetas

3. A elaboração da Árvore de Domínios da Linguística

Para o início do projeto, já tínhamos uma Árvore de Domínio (ou mapa conceitual) da Linguística - aquela proposta por Fromm (2008). Essa árvore, no entanto, carecia de estudos mais aprofundados, já que foi elaborada para outros propósitos.

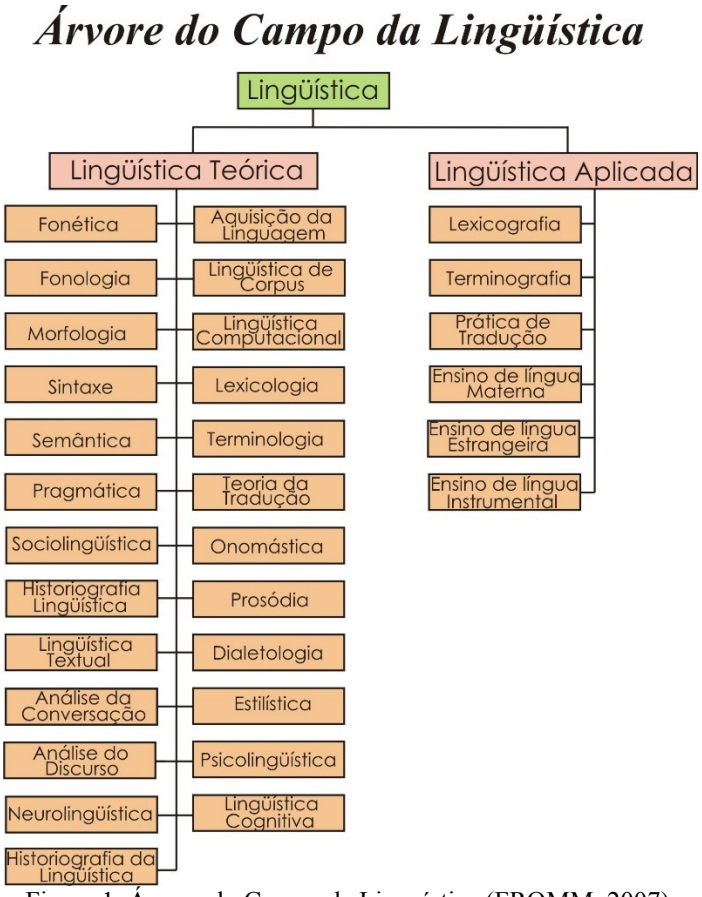

Figura 1. Árvore do Campo da Linguística (FROMM, 2007).

[1] Período estimado para a compilação completa do *corpus*.

O primeiro passo na reorganização dessa árvore foi a reformulação terminológica de uma das duas subáreas principais, por meio de sugestão de colegas: de Linguística Teórica passou-se para Linguística (simplesmente).

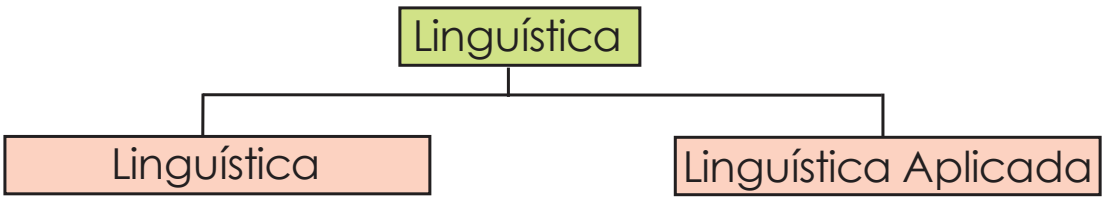

Figura 2. Nova divisão das subáreas principais.

O segundo passo foi a verificação de como se dá a taxonomia atual da CAPES para a grande área de Letras, Linguística e Artes:

Tabela 1. Áreas de Conhecimento. Fonte: CAPES. Disponível em: http://www.capes.gov.br/avaliacao/tabela-de-areas-de-conhecimento

80000002	**LINGUÍSTICA, LETRAS E ARTES**
	ÁREA DE AVALIAÇÃO: LETRAS / LINGUÍSTICA
80100007	**LINGUÍSTICA**
80101003	TEORIA E ANÁLISE LINGUÍSTICA
80102000	FISIOLOGIA DA LINGUAGEM
80103006	LINGÜÍSTICA HISTÓRICA
80104002	SOCIOLINGUÍSTICA E DIALETOLOGIA
80105009	PSICOLINGUÍSTICA
80106005	LINGUÍSTICA APLICADA
80200001	**LETRAS**
80201008	LÍNGUA PORTUGUESA
80202004	LÍNGUAS ESTRANGEIRAS MODERNAS
80203000	LÍNGUAS CLÁSSICAS
80204007	LÍNGUAS INDÍGENAS
80205003	TEORIA LITERARIA
80206000	LITERATURA BRASILEIRA
80207006	OUTRAS LITERATURAS VERNÁCULAS
80208002	LITERATURAS ESTRANGEIRAS MODERNAS
80209009	LITERATURAS CLÁSSICAS
80210007	LITERATURA COMPARADA

A CAPES e o CNPq já tentaram abranger mais áreas de conhecimento: há uma versão preliminar de proposta (tabela 2), datada de 2005, que pretendia atualizar a estrutura dessa árvore. No caso da nossa área, ficaria configurada desta maneira:

Tabela 2. Áreas de Conhecimento – proposta 2005. Fonte: CNPq. Disponível em: http://memoria.cnpq.br/areasconhecimento/docs/cee-areas_do_conhecimento.pdf

08. Grande Área: Linguagens e Artes
1. Área – Linguagem
Teoria da Linguagem Verbal
Teoria e Análise do Discurso
Teoria e Análise do Texto
Linguagem Verbal Não Oral
Linguagens Não Verbais
Linguagens Sincréticas
Teoria e Prática da Tradução
Filosofia da Linguagem
História das Idéias Lingüísticas

2. Área – Línguas
Fonética e Fonologia
Morfologia e Sintaxe
Semântica
Lexicologia, Lexicografia e Terminologia
Variação Lingüística
Mudança Lingüística
Uso Lingüístico
Aquisição da linguagem
Patologias da Linguagem
Tratamento Automático das Línguas
Língua Portuguesa
Línguas Clássicas
Línguas Estrangeiras Modernas
Línguas Indígenas
Outras Línguas

3. Área – Literatura
História da Literatura
Teoria da Literatura
Literatura Comparada
Literaturas Vernáculas
Literaturas Clássicas
Literaturas Estrangeiras Modernas
Literatura Infantil

Essa nova proposta, no entanto, não teve prosseguimento e a configuração atualmente em uso é aquela da tabela 1.

O terceiro passo foi a consulta a obras que funcionam como manuais da área de Linguística (por exemplo: LOPES, 1995; MARTELOTTA, 2008) e séries que procuram discutir vários aspectos desta ciência, como os livros de introdução à Linguística (por exemplo: MUSSALIM; BENTES, 2001; FIORIN, 2002)

Tendo como base as árvores apresentadas nas tabelas e figuras acima, além da consulta aos manuais e livros da área, e a necessidade de estabelecer parâmetros mais

amplos para a compilação equilibrada de um *corpus* de especialidade, desenvolvemos[2] a proposta de uma nova árvore, a ser discutida na sequência.

4. Objeto de pesquisa, abordagem, metodologia

A nossa proposta para a Árvore de Domínio da Linguística, na verdade, foi dividida em três: os objetos de pesquisa, as abordagens e as metodologias.

Quando nos referimos aos objetos de pesquisa, estamos pensando na árvore em si. Em relação à construção dessa árvore, podemos fazer a seguinte pergunta: em quantas subáreas se divide a Linguística? Respostas rápidas para esta pergunta costumam girar em torno das subáreas tradicionais, derivadas dos estudos gramaticais: Fonética, Fonologia, Morfologia, Sintaxe e Semântica. A Fonética tem como objeto de pesquisa os sons, a Morfologia tem como objeto de pesquisa os morfemas, a Semântica tem como objeto de pesquisa o significado, e assim por diante. O próprio processo de compilação do *corpus* vai nos fornecendo pistas de outras subáreas com as quais podemos trabalhar – ao serem identificadas, basta fazer uma procura detalhada no Google para saber a quantidade de textos nela produzida.

No momento[3], essa árvore (figura 3) conta com cinquenta subáreas assim divididas:

[2] Ressalta-se, na elaboração desta árvore, a grande contribuição que os colegas vem apresentando nos últimos quatro anos. Gostaríamos de agradecer as sugestões propostas pelos pesquisadores Heliana Mello (UFMG), Deise Prina Dutra (UFMG), Ariel Novodvorski (UFU), Maria José Bocorny Finatto (UFRGS) e, sobretudo, pela professora Vera Lúcia Menezes de Oliveira e Paiva (UFMG), sem a qual o procedimento de elaboração desta nova árvore ficaria bastante prejudicado.

[3] Importante notar que, por ainda não estar finalizado, o projeto permite que alteremos a estruturação da árvore e a compilação do *corpus*.

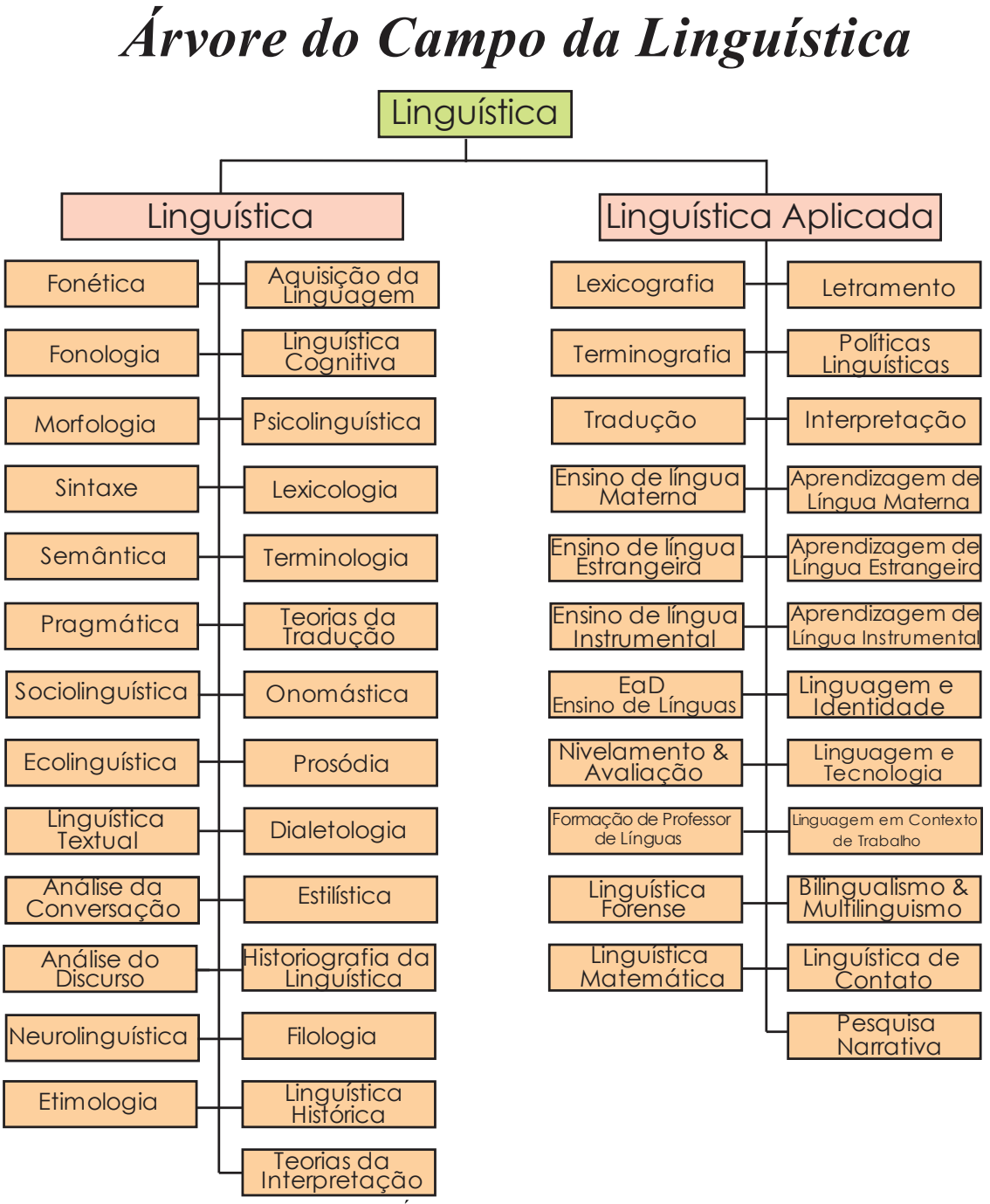

Figura 3. Proposta para uma Árvore de Domínio da Linguística – versão 2013.

Esta proposta consta, portanto, de uma árvore elaborada em três níveis: Linguística > Linguística/Linguística Aplicada > subáreas. Podemos notar que ainda caberia mais um subnível em determinadas subáreas: a Onomástica poderia contar com a Toponímia e a Antroponímia. Discussões futuras a respeito desta questão podem propor um maior detalhamento na elaboração da árvore.

Além da árvore propriamente dita, também estamos elaborando, para uma maior compreensão dos alunos envolvidos no projeto, as abordagens que podem ser trabalhadas em estudos linguísticos. Partindo, inicialmente, das classificações propostas por Paveau e Sarfati (2006), e também contando com a colaboração dos colegas, temos a seguinte situação:

Figura 4. Proposta para abordagens usadas na área de Linguística.

É importante notar que certas abordagens estão tradicionalmente mais associadas com algumas subáreas da Linguística: um caso clássico é a associação entre o Gerativismo e a Sintaxe; um caso mais recente é a associação entre as subáreas do Léxico (especialmente a Lexicografia e a Terminografia) e as abordagens Baseada em *Corpus* e Estatística. Costumamos frisar aos alunos que nem todas as abordagens podem ser usadas em todas as subáreas, ou seja, a associação não é livre.

Um terceiro ponto no desenvolvimento da árvore seriam as metodologias de pesquisa. Essa parte do estudo ainda precisa de mais desenvolvimento, mas já podemos apresentar algumas propostas:

Figura 4. Proposta inicial para metodologias usadas na área de Linguística.

A sequência do trabalho com os alunos é a seguinte: escolha de uma subárea do *corpus* de Linguística > compilação de quinhentas mil palavras em cada língua (trabalho em dupla: um aluno levanta os textos em português, o outro, em inglês) > escolha dos candidatos a termos, via programa de análise lexical (usando suas três ferramentas principais: lista de palavras, palavras-chave e concordanciador) > inserção de cinco termos no banco de dados do VoTec. Como resultado, os termos podem ser visualizados online, como o exemplo da figura 5.

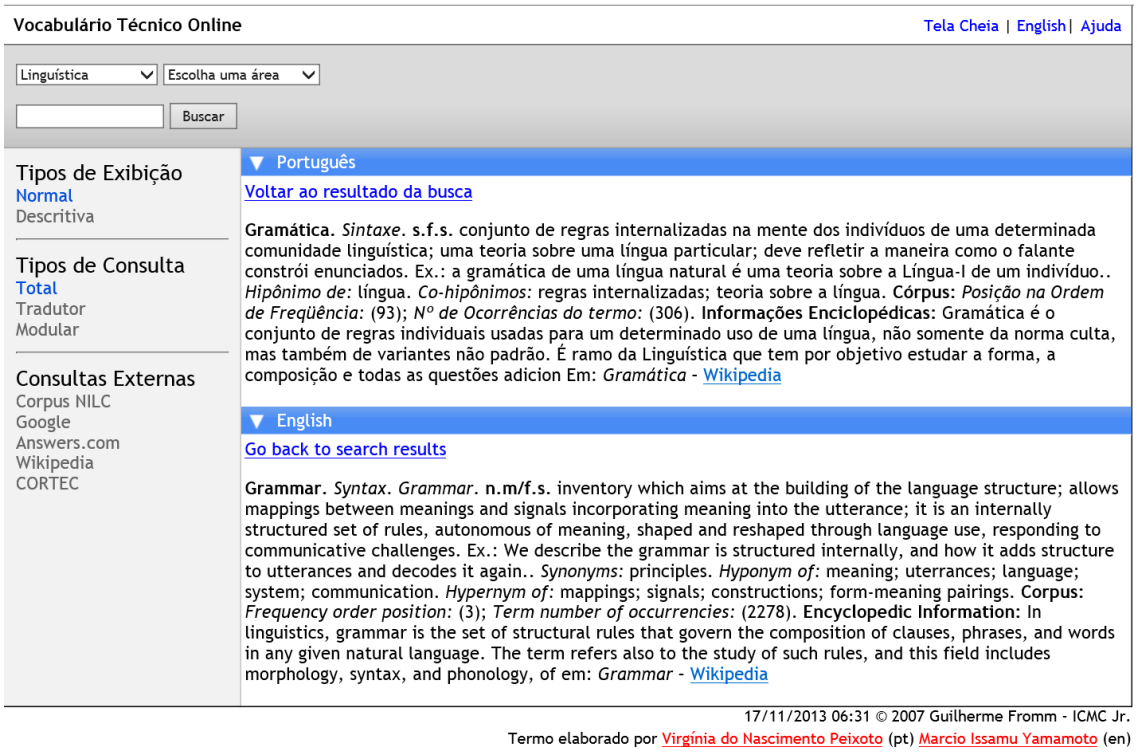

Figura 5. VoTec, termo *gramática*.

5. Considerações finais

O objetivo deste texto foi mostrar parte de um processo normal de planejamento de um *corpus* para análise linguística. Quando pretendemos trabalhar com a descrição terminológica/terminográfica de alguma área de especialidade, é essencial a elaboração de uma Árvore de Domínio da mesma. Embora essa árvore apareça (provavelmente) em apenas em uma página do trabalho, como resultado de pesquisa e ponto de partida para análises posteriores, sua elaboração consome muito tempo na leitura de documentos da área, entrevistas com especialistas, consulta à bibliografia especializada e aos sites sobre assuntos relatos, etc. Consideramos importante a valorização deste trabalho.

Referências bibliográficas

ALMEIDA, G. M. B. A Teoria Comunicativa da Terminologia e a sua prática. In: **ALFA**: revista de Linguística. V. 50, n. 2. São Paulo: UNESP, 2006.

ALMEIDA, G. M. B., PINO, D. H. P., SOUZA, D. S. L. A definição nos dicionários especializados: proposta metodológica. **RITerm – Debate Terminológico**, n. 3, janeiro 2007. Disponível em: <http://www.riterm.net/revista/n_3/Art_Barcellos_Almeida.pdf>. Acesso em: 15 abril 2007.

BARBOSA, M. A. Dicionário, vocabulário, glossário: concepções. In: ALVES, I. M. (org.). **A constituição da normalização terminológica no Brasil**. São Paulo: FFLCH/CITRAT, 2001.

BERBER-SARDINHA, A. **Lingüística de *Corpus***. São Paulo: Manole, 2004.

FIORIN, J. L. (org.). **Introdução à Lingüística**: objetos teóricos. São Paulo: Contexto, 2002.

FROMM, G**. VoTec**: a construção de vocabulários eletrônicos para aprendizes de tradução. São Paulo, 2007. Tese (Doutorado – Programa de Pós-Graduação em Estudos Linguísticos e Literários em Inglês – Departamento de Letras Modernas). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

FROMM, G; YAMAMOTO, M. I. Terminologia, Terminografia, Tradução e Linguística de *corpus*: a criação de um vocabulário bilíngue sobre Linguística. In: TAGNIN, S.; BEVILACQUA, C. **Corpora na Terminologia**. São Paulo: Hub Editorial, 2013.

HOUAISS, A. **Dicionário eletrônico Houaiss da língua portuguesa** 1.0. Rio de Janeiro: Objetiva, 2009.

ILARI, R. **Introdução ao estudo do léxico**: brincando com as palavras. 2ª. ed. São Paulo: Contexto, 2003.

KRIEGER, M. das G.; FINATTO, M. J. B. **Introdução à terminologia**: teoria e prática. São Paulo: Contexto, 2004.

LOPES, E. **Fundamentos da Lingüística Contemporânea**. São Paulo: Cultrix, 1995.

MARTELOTTA, M. E. (org.). **Manual de Lingüística**. São Paulo: Contexto, 2008.

MUSSALIM, F.; BENTES, A. C. (orgs.). **Introdução à lingüística**: domínios e fronteiras. São Paulo: Cortez, 2001.

PAVEAU, M. A.; SARFATI, G. E. **As grandes teorias da Linguística**: da gramática comparada à pragmática. São Carlos: Claraluz, 2006.

SCOTT, M. **WordSmith Tools version 6**. Liverpool: Lexical Analysis Software, 2012.

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