

VOCABULÁRIO BILÍNGUE PORTUGUÊS/INGLÊS DE LINGUÍSTICA GERAL

Márcio I. Yamamoto (UFU E UFG)
issamu2009@gmail.com

RESUMO

Apesar da difusão da Linguística desde 1960 na academia brasileira, a maioria dos dicionários de Linguística é limitada às traduções de obras estrangeiras. Objetivando sanar a falta de um vocabulário nessa área, estamos desenvolvendo um vocabulário bilíngue, português – inglês, na área de Linguística Geral. A metodologia do trabalho advém da Linguística de *Corpus* e as teorias basilares para o desenho do produto terminológico são as de Barbosa (1990), para o conceito de vocabulário, e a de Cabré (1999), a Teoria Comunicativa da Terminologia – TCT. O vocabulário baseia-se em *corpora* acadêmicos e em manuais de Linguística com 47 milhões de palavras. Neste artigo, são trazidos a proposta geral de organização do Vocabulário, seu desenho inicial, as etapas já cumpridas e as perspectivas para sua finalização.

Palavras-chave:

Terminologia Bilíngue. Linguística de *Corpus*. Vocabulário de Linguística Geral.

1. Introdução

A Linguística é a ciência da linguagem, historicamente iniciada a partir de Saussure no século XIX na Europa, precedida pelos períodos denominados pré-linguísticos e paralinguísticos. Em oposição à escola de gramáticos que a antecedeu, a Linguística se consolidou como uma ciência de caráter mais descritivo que prescritivo. Vários autores europeus e americanos contribuíram para a consolidação da Linguística na Europa e América, e a escola linguística brasileira usufruiu do conhecimento produzido por esse grupo de cientistas para promover a consolidação e crescimento dessa disciplina no Brasil. Como ciência, ela chega ao Brasil e é difundida na academia brasileira, a princípio, por meio dos trabalhos de Joaquim Mattoso Camara Jr. Nos anos 60, a Linguística passa a ser disciplina ministrada nos cursos de Letras das universidades brasileiras (BAGNO; RANGEL, 2005; CASTILHO, 2007).

Atualmente há grandes centros que se dedicam ao estudo desta área no Brasil com produções relevantes a nível nacional e internacional. Apesar da difusão da disciplina no âmbito universitário e científico há mais de meia década, a maioria dos dicionários técnicos disponíveis, para

uso dos acadêmicos e profissionais da área, são limitados a traduções de obras estrangeiras em geral, ou obras que objetivam a especialistas, não a leigos. Nosso objetivo é propor um vocabulário bilíngue, português-inglês, de termos da Linguística Geral, intitulado “Vocabulário da Linguística”. Ele será criado a partir de *corpora* autênticos e disponibilizado de forma gratuita na Internet no *site* <http://pos.votec.ileel.ufu.br/> a partir de 2020, de acordo com o cronograma do projeto.

Este artigo compõe-se de duas partes: (1) Sobre o Vocabulário de Linguística e (2) Sobre a organização do *corpus*. Na primeira parte, apresentaremos a justificativa, os referenciais teóricos, o público-alvo deste trabalho. Na segunda parte, apresentaremos a organização do *corpus*, sua padronização, seu processamento pelo console do *WordSmithTools* 6.0 (SCOTT, 2012), doravante WST, e um breve explicação da plataforma VoTec (Vocabulário Técnico) (FROMM, 2007).

PARTE I – Sobre o Vocabulário de Linguística

2. *Justificativa e referenciais teóricos do trabalho*

Assim como outras áreas da Ciência, a área da Linguística detém um estatuto que garante sua especificidade, o que faz com que uma linguagem de especialidade se faça necessária para expressão de conceitos que a diferenciam de outras Ciências Humanas.

Tal especificidade obriga alunos, estudiosos da área da Linguística e outros profissionais a dominarem um conjunto de conceitos, significados e terminologias que permite a compreensão da Linguística, suas subáreas e ramificações. Sem que haja esse domínio na área conceitual e terminológica, a compreensão da disciplina é comprometida e até mesmo pode tornar-se impossível, especialmente para alunos ingressantes em cursos de graduação.

Em um mundo globalizado, no qual a ciência se populariza mais rapidamente, os cientistas brasileiros podem ter acesso ao conhecimento produzido no exterior, divulgar e difundir sua produção intelectual em nível internacional usando a língua inglesa. Para justificar esta afirmação, partimos do conceito de **língua franca**, entendendo que, no presente século, a língua que cumpre essa função é a língua inglesa. Portanto, acre-

ditamos ser pertinente a construção de um vocabulário terminológico que disponibilize os dados em inglês, já que uma obra bilíngue português-inglês, atenderia bem o público lusófono, anglófono e internacional.

Além da justificativa quanto à escolha das línguas para a constituição de um vocabulário bilíngue, é necessário explicitar o porquê do uso do termo Vocabulário e não Dicionário. Esta escolha se deu por uma questão teórica, já que esta proposta de trabalho se insere na perspectiva de Barbosa (1990) quanto à definição de **vocabulário**. Primeiramente há o **vocabulário**, o léxico de uma língua, cuja unidade seria o **vocabulo**. Além do **vocabulário**, concebido como um conjunto, inserido no léxico das línguas, também existiria o **Conjunto Vocabulário** ou **Vocabulários das Ciências**, o **vocabulário de especialidades**, conjunto terminológico ou terminologia. É dentro dessa perspectiva que optamos por denominar o produto desta pesquisa terminográfica como “Vocabulário da Linguística” e não Dicionário de Linguística.

Este trabalho envolve várias áreas da Linguística, portanto será necessário tratarmos de pressupostos teóricos concernentes à Terminologia/Terminografia, a Teoria Comunicativa da Terminologia (TCT) de Cabré (1999), e a Linguística de *Corpus* (LC). (SARDINHA, 2004; VIANA e TAGNIN, 2010).

2.1. O Breve reflexão sobre a Terminologia e a Teoria Comunicativa da Terminologia (TCT)

A Terminologia é uma disciplina que tem como objeto o termo técnico-científico, bem como as fraseologias especializadas e a definição terminológica. Ela pode ter dois enfoques: (i) o desenvolvimento teórico e as análises descritivas e (ii) as aplicações terminológicas. Suas bases foram estabelecidas pelo austríaco Wüster (1898-1977), na Universidade de Viena em 1972, objetivando a univocidade comunicacional em nível internacional e o lançamento da Teoria Geral da Terminologia (TGT), de caráter prescritivo (KRIEGER E FINATTO, 2004; TALAVÁN, 2016.).

Posteriormente a Wüster, temos Alain Rey que traz uma reflexão crítica da TGT na década de 80 e propõe a compreensão da unidade terminológica, objeto da Terminologia, de um ponto de vista descritivo, como uma unidade da língua natural. É a corrente da socioterminologia, que busca resgatar a dimensão social da terminologia e a variação terminológica na linguagem de especialidade (KRIEGER E FINATTO,

2004; GAUDIN, 2005).

Nos anos 1990, Maria Teresa Cabré lança as bases da Teoria Comunicativa da Terminologia (TCT), e a unidade terminológica é concebida como parte da linguagem natural. Os aspectos comunicativos das linguagens especializadas são mais valorizados pela TCT que os normativos da TGT e a unidade terminológica passa a existir pelo contexto no qual se insere, o de comunicação especializada (KRIEGER E FINATTO, 2004; CABRÉ, 1999). É dentro da perspectiva dessa teoria que este trabalho se insere.

A TCT concebe o termo como uma UL inserida em um sistema linguístico, ou seja, a UT é concebida como integrante de um sistema morfológico, fonológico, semântico e sintático das línguas naturais e não como uma etiqueta artificial cunhada para denominar conceitos.

O objeto de estudo da Terminologia é a **unidade terminológica** que, segundo Dornelles (2015), é “uma unidade lexical da língua natural que adquire valor especializado dentro de um contexto especializado, segundo critérios semânticos, discursivos e pragmáticos” (DORNELLES, 2015, p.13).

Na perspectiva da TCT, a unidade lexical pode ser, ora uma unidade lexical, ora uma unidade terminológica. O que distingue um *status* do outro é o contexto no qual ela se insere, o que Cabré (2011) denomina de “caráter poliédrico do termo” (Cabré, 2011, p.149). Isto se dá, uma vez que o termo pode ser, ao mesmo tempo, uma unidade linguística cognitiva e sociocomunicativa: linguística por ser proveniente do contexto real de uso da língua, a partir do qual a unidade lexical adquire um valor especial em relação às outras; cognitiva por promover a distinção entre o conhecimento especializado do geral ou não especializado; e sociocomunicativa, pois consideram-se os aspectos pragmáticos que identificam a conformação e uso das unidades lexicais/terminológicas em determinadas situações comunicativas.

Cabré (1994) e Rey (1977) são citados por Finatto (2001) ao tratar da conceituação e definição na Terminologia, propondo que a definição terminológica não busca representar ou reproduzir, mas evocar um conceito; logo, ela busca ressaltar uma noção, dentro de um conjunto de outras noções (FINATTO, 2001). Se considerarmos o princípio da Lexicologia, em que uma unidade lexical, quando se difere de outra, produz

sentidos diferentes, compreenderemos que há fronteiras distintas entre o conceito, o significado e a definição¹.

2.2. *A linguística de corpus e taxonomia do corpus*

A perspectiva de uma obra baseada em *corpus* advém da Linguística de Corpus – (LC) (SARDINHA, 2004; VIANA e TAGNIN, 2010), de sua concepção de língua, e da abordagem que ela propõe para a construção de obras lexicográficas ou terminográficas.

O conceito que adotamos de *corpus*, dentro da LC, é o de Viana (2010): uma compilação de textos de ocorrência natural que representa uma certa língua ou seus aspectos específicos, possibilitando uma análise linguística pré-estabelecida; e Tagnin (2015), quem define *corpora* (*corpus* no plural) da seguinte forma: “bancos de textos de linguagem autêntica, criteriosamente construídos, destinados à pesquisa e legíveis por computador.” (TAGNIN, 2015, p. 20).

A LC concebe a língua como um sistema probabilístico (HALLIDAY, 1991), logo, a metodologia usada por essa subárea da Linguística propõe que as obras sejam construídas a partir de um *corpus* representativo da língua, sejam textos de ocorrência natural e autênticos, que permitam seu processamento por computador e que sejam criteriosamente selecionados para atingir o objetivo da pesquisa. (SARDINHA, 2004; VIANA e TAGNIN, 2010).

Neste artigo, usaremos o termo *corpus* para designar o conjunto de textos compilados para cada língua na área da Linguística, sendo eles quatro: *corpus* de artigos acadêmicos em português, em língua inglesa, manuais de Linguística em português e em inglês. Ao conjunto de dois ou mais desses subgrupos de *corpus*, denominaremos *corpora*.

A LC foi escolhida por atender ao rigor teórico-metodológico da Terminologia/Terminografia, observando-se os princípios de frequência e representatividade terminológica de natureza probabilística (BERBER

¹ Na lexicologia/lexicografia, busca-se registrar todas as acepções de um verbete, os significados denotativos e conotativos. Na Terminologia/Terminografia isto não deve acontecer, já que o termo/vocábulo/UT só poderia representar um conceito em contextos de comunicação especializada (o termo técnico-científico deve ser monossêmico para garantir a univocidade).

SARDINHA, 2004; VIANA; TAGNIN, 2010). As vantagens dessa metodologia são: (i) possibilitar a descrição da língua em uso, (ii) possibilitar a análise quantitativa e qualitativa de dados linguísticos, (iii) disponibilizar resultados estatisticamente confiáveis e precisos dos termos usados pelos especialistas. Contudo, é importante elucidar a taxonomia dos *corpora* nesta pesquisa.

O *corpus* de estudo pode ser classificado de diversas formas e a partir de perspectivas diferentes. Abaixo segue a classificação dos *corpora* de Linguística, de acordo com a proposta de Teixeira (2008).

Quadro 1 - Taxonomia do *corpus* de Linguística (português-ínglês)

Língua	Bílingue (ínglês e português)
Modo	Escrito (textos acadêmicos: manuais, artigos científicos, dissertações e teses)
Data de publicação	Sincrônico (levantamento realizado entre 2010 e 2016),
Seleção	Estático
Conteúdo	Especializado (Linguística e Linguística Aplicada)
Autoria	Falantes nativos (ínglês e português), individual/coletivo
Disposição Interna	Comparável
Uso na pesquisa	Estudo (análise terminológica/terminográfica)
Nível de Codificação	Com cabeçalhos, sem etiquetas

O Quadro 1 traz a taxonomia do *corpus*² que usamos neste trabalho, explicada a seguir. Ele é um *corpus*:

- (i) **bílingue**, por abarcar a língua portuguesa e inglesa;
- (i) **escrito** contrastando com o oral;
- (ii) **sincrônico**, pois traz o registro de uma dada língua num momento específico da história, enquanto que o diacrônico cobre períodos diferentes dessa língua; os textos

² Excepcionalmente, neste caso, concebemos o conceito de *corpus* como a representação do conjunto de todos os *subcorpora* de uma forma singular, como pertencentes à Linguística, de língua inglesa ou portuguesa.

- desses *corpora* abarcam o período de 2010 a 2018, considerado o período de busca e digitalização dos manuais;
- (iii) **estático**, o que significa que o *corpus* não permite ser alterado, ou seja, a inclusão ou exclusão de textos não pode ser realizada;
 - (iv) **especializado**, já que objetiva a linguagem de especialidade da Linguística e da Linguística Aplicada e contrasta com o de língua geral;
 - (v) A autoria é caracterizada por **falantes nativos** de língua portuguesa de português de Portugal ou do Brasil; e a autoria de língua inglesa é de falantes de inglês de várias partes do globo;
 - (vi) É **comparável**, pois o *corpus* de português não é uma tradução do *corpus* de inglês e *vice-versa*. Tagnin (2015) o define da seguinte forma, “textos originais em duas (ou mais) línguas, numa determinada área de domínio.” (TAGNIN, 2015, p.26);
 - (vii) Com relação ao aspecto de uso, o *corpus* é de **estudo**, pois será usado para estudo nesta pesquisa, quanto ao nível de codificação, **com cabeçalhos, sem etiquetas**.

Na próxima seção, apresentaremos brevemente o console do WST, usado para processamento dos dados quantitativos da pesquisa, e a plataforma terminológica do VoTec, na qual os dados, frutos da análise qualitativa, serão alocados e disponibilizados para o público via internet.

2.3. *Word smith tools e votec*

O console utilizado para processamento dos *corpora* será o WST, do qual usaremos as ferramentas lista de palavras, lista de palavras-chave e o concordanciador³. O WST permite o processamento e análise dos *corpora* que seriam impossíveis se dependêssemos da leitura humana dado o período desta pesquisa, além de fornecer resultados estatisticamente confiáveis e precisos da terminologia e conceitos usados pelos especialistas.

O uso desses *corpora* proporcionará a disponibilização de dados

³ Ferramenta que disponibiliza a palavra-chave no contexto em que ela se insere.

oriundos de fontes confiáveis aos usuários. Essa disponibilização será realizada por meio do **VoTec**⁴ (FROMM, 2007; FROMM; YAMAMOTO, 2013), já que as definições dessa plataforma serão construídas a partir de traços conceituais provenientes desses *corpora* acadêmicos e dos manuais de Linguística.

As vantagens proporcionadas pelo uso do WST que nos levaram à sua escolha são: identificação do candidato a termo por meio da ferramenta lista de palavras-chave; a localização dos contextos definitórios⁵ e explicativos⁶ (AUBERT, 1996; DUBUC, 1992; PAVEL; NOLET, 2002; BARROS, 2007.) no *corpus* de estudo; o salvamento dessas listas no computador, o que nem todos os programas disponibilizam. A desvantagem é que este é um programa pago.

2.4. Definição lexicográfica vs. Terminográfica

Quando pensamos em obras lexicográficas e terminográficas, especialmente glossários e dicionários, a maneira como as definições são construídas nos levam a classificá-las em subtipos diferentes.

A forma como um terminólogo/lexicólogo organiza as informações usadas para se construir uma definição lexicográfica/terminográfica denomina-se paradigma definicional. Segundo Finatto (2001), paradigmas definicionais podem ser enciclopédicos, lexicográficos ou terminológicos.

O paradigma definicional enciclopédico é mais detalhado, trazendo em si a definição terminológica, o histórico do termo, dados geográficos, culturais, históricos e sociais, entre outros.

O paradigma definicional de cunho lexicográfico pode trazer da-

⁴ Maiores detalhes sobre a plataforma e seu funcionamento encontram-se no artigo intitulado *Linguística Histórica e Linguística de Corpus: Uma Proposta Para Dicionário Bilingue Português-Ingês*, disponível em em <<http://www.filologia.org.br/rph/ANO19/55SUP/058.pdf>>. Acesso em 15 maio 2018.

⁵ Contextos definitórios são aqueles usados pelos autores para definir algo ou alguém, usando para tal uma rede de traços conceituais para delimitar o objeto, de forma que o sentido fique claro para seu interlocutor.

⁶ Neste tipo de contexto, encontramos alguns dados sobre o objeto, tais quais: materialidade, finalidade, funcionamento, constituição.

dos gramaticais, morfológicos e fonológicos do termo, sua definição e exemplos, dependendo da abordagem do terminólogo.

Finalmente, o paradigma definicional de cunho terminológico/terminográfico valoriza os traços conceituais, é mais sucinto e as “características restritivas são apresentadas de modo que se diferencie o conceito descrito de todas as outras pertencentes ao mesmo nível de abstração do sistema de conceitos, dando-lhes um caráter único” (BARROS, 2004, p. 171).

Neste trabalho, o paradigma definicional a ser construído será definido após pesquisa com o público-alvo, usuários desta obra, descrito no próximo item.

2.5. Público-alvo do vocabulário e modelos de paradigmas definicionais

Nesta seção apresentaremos o público-alvo desta obra, bem como apresentaremos paradigmas definicionais de algumas obras terminológicas pesquisadas.

Considerando a especificidade da Linguística como área da Ciência, acreditamos que uma obra terminológica, direcionada a alunos iniciantes, viria a suprir a necessidade de conceituação desse vocabulário de especialidade.

Dessa forma, tanto o acadêmico quanto os profissionais da área poderiam adquirir o domínio necessário da Linguística. Isto é, nosso público-alvo abrangeria alunos e professores da área de Letras, e outros profissionais que viriam a se interessar pela área. Segundo Talaván (2016), eles podem ser chamados de usuários diretos ou indiretos.

Os alunos de Letras seriam os usuários **diretos**⁷, principalmente aqueles dos períodos iniciais, que buscam dominar os conceitos relativos à Linguística, recém ingressados em uma universidade.

Os professores, em especial, pesquisadores e aqueles que desenvolvem trabalhos interinstitucionais com universidades do exterior, também seriam **usuários diretos**, os quais poderiam se beneficiar de tal obra para a tradução e leitura de obra escritas em inglês. Esses professores também recorrem à bibliografia proveniente do exterior para ministrar seus cursos, desenvolver pesquisas e publicar em eventos

⁷ Especialistas

nacionais e internacionais.

Outro perfil de usuários **indiretos**, seriam os profissionais da área de Tradução que trabalham com obras em língua portuguesa e língua inglesa, ou mesmo estudantes desses cursos. Finalmente, outros usuários comuns que utilizam a Internet também poderiam usufruir de tal vocabulário, pois o acesso é livre a todos os públicos.

PARTE II – Organização do *corpus*

Nesta parte apresentaremos os *corpora* de análise deste trabalho e a metodologia. Os *corpora* estão parcialmente compilados e a metodologia adotada baseia-se na LC e em princípios da Terminologia e Terminografia.

2. *Corpora: bases do Vocabulário da Linguística*

Os *corpora* que serão analisados para a construção do Vocabulário da Linguística serão compostos por artigos científicos: teses, dissertações e artigos, e manuais de Linguística.

Uma parte dos *corpora* bilíngue de Linguística Geral, português-ínglês, composto por artigos científicos, já foi compilado, em sua grande maioria, por alunos de graduação e pós-graduação, como parte do trabalho desenvolvido pelo professor Guilherme Fromm (FROMM; YAMAMOTO, 2013). Este acervo totaliza aproximadamente 45 milhões de palavras em cada língua. Como metodologia deste trabalho, tivemos as seguintes etapas:

a) uma “árvore de domínio⁸” da Linguística foi exposta aos alunos;

b) os alunos foram orientados a optar por uma subárea da Linguística, cujo *corpus* não tinha sido coletado;

c) sob a supervisão do professor Fromm, os alunos pesquisaram na Internet os textos disponíveis nessa subárea (em português e inglês), salvando-os em formato .txt. Os textos em .txt foram processados pelo

⁸ A árvore de domínio é um diagrama esquemático que permite ao pesquisador ter uma visão geral das áreas e subáreas que compõem sua pesquisa, e do dimensionamento do trabalho.

WST, sendo usadas as ferramentas lista de palavras, lista de palavras-chave e concordanciador;

d) procedida a análise qualitativa dos dados (escolha dos candidatos a termo, identificação de contextos definitórios e explicativos, seleção final dos termos), os dados foram carregados (*uploaded*) no VoTec (FROMM, 2007; FROMM; YAMAMOTO, 2013), disponível em <<http://pos.votec.ileel.ufu.br/>>, os alunos tiveram seus nomes registrados como crédito do trabalho realizado.

O segundo grupo de *corpora*, também bilíngue, português-inglês, é composto por Manuais de Linguística, compilados em formato PDF e transformados em .txt, para serem lidos eletronicamente pelo console do WST. Os manuais compilados em inglês e português estão elencados abaixo, no **Quadro 3**.

Quadro 1. Manuais de Linguística em inglês e português

Nome do Manual		Autor
1.	<i>General Linguistics: An Introductory Survey Third Edition</i>	Robins e Leech
2.	<i>The Handbook of Linguistics</i>	Mark Aronoff e Janie Rees-Miller
3.	<i>The Handbook of Linguistics</i>	Wiley & Black
4.	<i>The Handbook of English Linguistics</i>	Aarts & McMahon
5.	<i>Linguistics</i>	Farmer Harnish
6.	Fundamentos da Linguística Contemporânea	Edward Lopes
7.	Introdução à Linguística	José Luiz Fiorin (Org.)
8.	Introdução a Linguística I: objetos teóricos	José Luiz Fiorin (Org.)
9.	Introdução à linguística: domínios e fronteiras, v. I	Fernanda Mussalim, Anna Christina Bentes (Org.)
10.	Introdução à linguística: domínios e fronteiras, v. II	Fernanda Mussalim, Anna Christina Bentes (Org.)

11.	Introdução à linguística: domínios e fronteiras, v. III	Fernanda Mussalim, Anna Christina Bentes (Org.)
12.	Introdução à Linguística Teórica	John Lyons
13.	Linguística e comunicação	Roman Jakobson
14.	Manual de Linguística: subsídios para a formação de professores	MEC e UNESCO
15.	Manual de Linguística Românica	Benedek Elemer Vidos
16.	Manual de Linguística: fonologia, morfologia e sintaxe	Luiz Carlos Schwindt (Org.)

O tamanho desses *corpora* está equilibrado em aproximadamente 1,3 milhões de palavras/*tokens* em cada língua.

Os objetivos que buscamos atingir com a compilação desses *corpora* de manuais são:

- (i) compará-lo ao anterior;
- (ii) identificar qual deles trará mais contextos definitórios; e
- (iii) usá-los para construção de definições na plataforma do VoTec.

Como objetivo geral deste trabalho, pretendemos criar um banco de dados com aproximadamente 400-500 verbetes a partir dos *corpora* disponíveis na rede mundial de computadores e dos manuais de Linguística, provenientes das produções científicas de LP e inglesa. As definições serão extraídas de *corpora* autênticos e serão organizadas a partir de termos técnicos em português e inglês. Essa abordagem é justificada por Krieger e Finatto (2004).

[...] destaca-se o processo de globalização que, incrementando as transações comerciais entre as nações, propiciou o surgimento dos atuais blocos econômicos, bem como de uma série de intercâmbios que ultrapassaram o âmbito comercial, expandindo-se para o mundo científico, tecnológico e cultural. Tal situação fez crescer a preocupação com a utilização e a tradução adequadas das terminologias, posto que os protagonistas dos processos de alargamento de fronteiras passaram a perceber o importante papel dos **termos técnicos** para uma comunicação mais eficiente, uma adequada transferência de

tecnologia e um correto estabelecimento de contratos comerciais entre outras ações de cooperação. (KRIEGER e FINATTO, 2004, p. 27) (Grifo nosso)

Após o tratamento dos dados (construção de lista de palavras, lista de palavras-chave, análise dos cotextos⁹ *via* Concordanciador), eles serão sistematizados em fichas terminológicas e o vocabulário será posteriormente disponibilizado para o público usuário do VoTec.

Como objetivo deste trabalho temos a construção de um vocabulário bilíngue português-inglês, em contraste¹⁰, na área de Linguística que atenda alunos dos anos ou semestres iniciais do Curso de Letras.

Na próxima seção, abordaremos os aspectos metodológicos deste trabalho, desde a compilação de *corpus* até a disponibilização dos dados para os consulentes do Vocabulário bilíngue de Linguística.

2.1. Metodologia: compilação e padronização dos corpora

Este trabalho terminológico/terminográfico tem os *corpora* de Linguística como um dos objetos de estudo. As etapas do trabalho são: entrevista com o público-alvo para definição do paradigma definicional mais adequado para atendê-lo; compilação, seleção e estudo/análise dos *corpora*; entrevista com especialistas da Linguística; alimentação dos dados na plataforma do VoTec; e a escrita da tese.

Nosso ponto de partida será uma pesquisa que buscará definir o padrão de paradigma definicional para ser adotado na nova plataforma do VoTec (FROMM, 2007; FROMM; YAMAMOTO, 2013), disponível em < <http://pos.votec.ileel.ufu.br/>>.

Essa nova plataforma conta com um novo *layout* que admitirá modificações, como a inserção de NOTAS, permitindo que dados definicionais extras sejam disponibilizados aos consulentes.

Nessa primeira fase da pesquisa, disponibilizaremos vários

⁹ Cotexto em Linguística de *Corpus* refere-se ao ambiente linguístico, isto é, os itens lexicais que estão à direita e à esquerda de uma dada palavra. (VIANA, 2010, p. 71).

¹⁰ A definição terminológica bilíngue em contraste é aquela na qual a definição em L1 não é a tradução de L2 e vice-versa. No caso desta pesquisa, as definições são distintas, pois elas são provenientes de *corpora* diferentes.

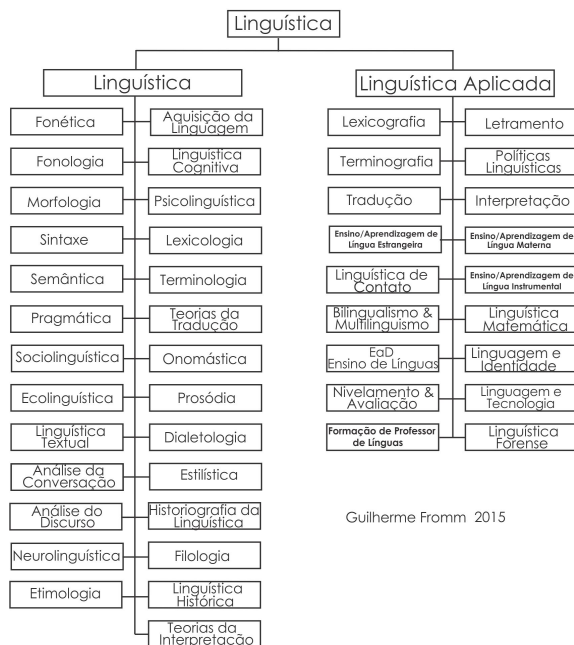
padrões de paradigmas definicionais, *via GoogleDocs*, para que o público-alvo possa escolher a que melhor atenda ao objetivo de conceituar e definir o vocabulário de especialidade da Linguística com clareza e objetividade.

Os padrões disponibilizados serão o lexicográfico, o enciclopédico, o terminológico, ou outro padrão que poderá ser construído no decorrer do trabalho. O questionário disponibilizado aos alunos será do tipo estruturado e os alunos responderão às perguntas selecionando, por fim, um paradigma padrão.

Na fase final do trabalho, o público-alvo e especialistas serão consultados quanto à clareza conceitual e objetividade das definições. Nessa avaliação final, o objetivo será verificar aspectos conceituais e não estruturais.

Outro procedimento metodológico será a organização das subáreas da Linguística, na árvore de domínio, a qual é fruto do trabalho desenvolvido pelo professor Fromm nos cursos de Letras e Tradução. Veja na Figura 1 a árvore de domínio da Linguística conforme proposta por Fromm (2015),

Árvore do Campo da Linguística



Guilherme Fromm 2015

Figura 1 – Árvore do Campo da Linguística com as subáreas de estudo, reformulada a partir da árvore de 2015. (Fonte: FROMM, 2015)

De acordo com a Figura 1, podemos visualizar as subáreas da Linguística que serão analisadas neste trabalho. Isto é, para cada subárea da Linguística, partimos do número de 1 milhão de palavras para dimensionamento do *corpus*, sendo metade dessas correspondentes a cada uma das línguas: português e inglês.

Haja visto que temos 27 subáreas na grande área da Linguística e 18 subáreas na área de LA, objetivamos ter um total de aproximadamente 45 milhões de palavras compondo os *corpora* de artigos científicos. A quantia de 45 milhões de palavras, em português, será a mesma em inglês, caso consigamos levantar 500 mil palavras em cada língua.

A título de exemplificação, desse fazer terminográfico, citamos a Ecolinguística como uma subárea em crescimento no Brasil. A pesquisa com alunos da graduação, em anos anteriores, revelou que ainda não havia publicações acadêmicas na internet que atingissem as 500 mil

palavras.

Os *corpora* de artigos acadêmicos, armazenados em uma pasta do *OneDrive*, estão em processo de padronização. Para que tal procedimento fosse sistematizado, partimos das subáreas propostas na árvore de domínios da Linguística de Fromm (2015). O que descrevemos a seguir são os procedimentos realizados rumo à esta padronização:

(1) as pastas foram organizadas de acordo com a árvore de domínio (FROMM, 2015); as pastas das subáreas que não estavam presentes nessa árvore foram excluídas. No caso de pastas com nomes semelhantes, preservamos as que estavam nomeadas de acordo com a árvore de domínio e os arquivos foram transferidos para a pasta mantida.

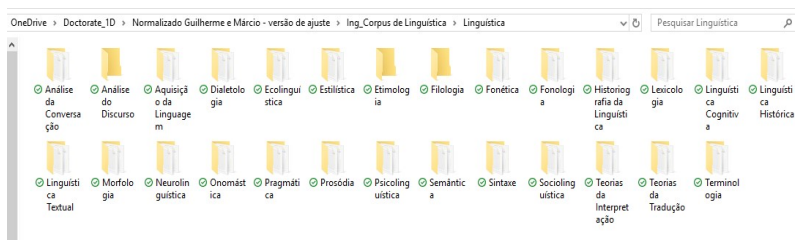


Figura 2. Visão parcial das subáreas da Linguística e suas pastas (Fonte: Autor)

(2) O passo seguinte foi o dimensionamento do *corpus* contido em cada subárea: baixamos os arquivos e fizemos seu processamento com o WST para determinar o número de palavras contido em cada pasta.

Na execução do passo 1 ocorreu que tínhamos a pasta da subárea Pesquisa Narrativa na pasta de Linguística Aplicada (LA)/Língua Inglesa (LI), mas esta não consta na árvore de domínios. Além desta, havia a subárea de Linguística de Contato na árvore de domínios, cuja pasta é inexistente nos arquivos compartilhados. Diante desta realidade, foi necessário decidir pela inclusão ou não da subárea da Pesquisa Narrativa na árvore de domínio, bem como será feita a coleta de *corpus* da subárea de Linguística de Contato.

No que tange à área da Pesquisa Narrativa, essa pasta foi excluída. A exclusão se deu, pois, em consulta aos especialistas dessa área, sou-

bemos que ela não pertence à Linguística, mas sim à Educação e é de caráter interdisciplinar, como a Terminologia.

Com relação ao passo 2, no qual dimensionamos o tamanho dos *corpora*, identificamos 4 tipos de pastas, quais sejam:

- (i) pastas com aproximadamente 500 mil palavras;
- (ii) pastas com mais de um milhão de palavras;
- (iii) pastas com menos de 500 mil palavras; e
- (iv) pastas sem nenhum arquivo contido nelas.

Identificadas os tipos de pastas, três ações serão necessárias: a primeira será a de reduzir a quantidade de arquivos nas pastas que contém mais de 1 milhão de palavras. Para tal ação, uma análise dos textos será feita utilizando-se o WST e manteremos os arquivos nos quais a presença de contextos definitórios é maior. A segunda ação será a de aumentar o *corpus* até que ele atinja ao número de 500 mil palavras; e a terceira ação será a de compilar os textos para aquelas áreas que ainda não os possuem. O número de 500 mil palavras por subárea se deve à formatação original do projeto, na qual este número foi considerado como representativo da subárea, já que responderia pela dimensão de um *corpus* médio, conforme Teixeira (2008, p. 184). O objetivo de balancear os *corpora* em 500 mil palavras visa atender ao princípio de balanceamento do *corpus*, princípio metodológica da LC. Dessa forma, os dados resultantes da análise computacional dos textos são balanceados e corresponderão a uma amostra realista do uso da língua, o princípio da representatividade por meio do balanceamento.

Quanto ao dimensionamento desses *corpora* de manuais, na primeira parte da pesquisa ainda não era possível prever, já que não podíamos prever a quantidade daqueles disponíveis na Internet ou se todos teriam que ser digitalizados. Até aquele momento, tínhamos compilado aproximadamente 320 mil palavras em português e 1,2 milhões em inglês. Estes dados mostravam que o *corpus* em português era menor que o inglês, o que demandou que equilibrássemos estes *corpora* após o processo de compilamento que estivesse finalizado.

O que já foi feito até o momento é: compilamos os manuais em inglês e em português já digitalizados, a partir do *site* Library Genesis, <<http://gen.lib.rus.ec/>>. O que observamos é que há muito mais manuais em inglês do que em português. Isto significou que era necessário escanear as obras em português, o que fizemos por várias horas na

biblioteca da Universidade Federal de Uberlândia (UFU).

As etapas de compilação dos manuais de linguísticas foram:

1) seleção dos manuais que continham claramente as subáreas da Linguística tais quais: morfologia, fonética, fonologia, sintaxe, semântica, variação linguística, pragmática, psicologia da linguagem, aquisição da linguagem, neurolinguística, dentre outros;

2) que eram direcionados para um público-alvo de iniciantes (checagem da introdução).

Após a seleção, seguiu-se a digitalização das obras em formato .PDF, já que este era o único tipo de formato de arquivo disponível, à mão, que escaneasse as obras rapidamente. Neste caso, cabe explicar que as obras foram compiladas por meio de um escâner planetário, disponível na biblioteca da UFU. A capacidade de escaneamento é de 100 páginas para cada dez minutos, aproximadamente.

Em seguida, as obras foram processadas pelo programa *OmniPage* 16, que transformou os arquivos do formato .PDF para .txt. Neste caso, a obras eram carregadas por completo e então processadas pelo programa. A justificativa para esta alteração de formato se deve ao fato de que os arquivos precisam estar em formato .txt para ser processado pelo WST.

No que tange à compilação, lançamos mão de duas estratégias: (1) compilar manuais digitalizados diretamente da internet; e (2) digitalizar manuais impressos e disponibilizá-los em formato .txt.

No caso de compilarmos os manuais em inglês diretamente da internet, optamos pelo *site Library Genesis*, < <http://gen.lib.rus.ec/> >, usando as palavras-chaves *Handbook of Linguistics*. Nesta busca, a busca resultou em 8 páginas, cada uma com 25 resultados, com vários tipos de manuais de Linguística, LA, de línguas europeias e orientais, como árabe, hebraico, grego, japonês, coreano, entre outros. Para selecionar as obras mais próximas dos manuais já compilados, usei o padrão dos índices que já havíamos compilado, descritos no início desta seção.

Em língua portuguesa, usamos o *site Library Genesis* também, mas só havia dois manuais organizados por Fiorin (vide lista na seção 3, Quadro 3). Os outros foram escaneados de livros impressos. Nesta seleção de manuais impressos em português, um detalhe nos chamou a atenção. O Manual de Linguística de Jakobson, intitulado Linguística e

Comunicação, é um manual com poucos contextos definitórios. Contudo, ao analisar a estrutura textual, observamos que este manual é rico em conceitos. Esses conceitos podem ser úteis na seleção de contextos explicativos, conseqüentemente para extrairmos traços conceituais e posteriormente construirmos as definições dos termos. Por isso, optamos por mantê-lo em nossa lista de *corpora* de manuais compilados

Com a compilação dos *corpora* concluída, faremos uma análise comparativa para observar qual deles traz mais contextos definitórios, úteis para a construção das definições, e a possível eleição dos candidatos a termos. A seleção de candidatos a termos será norteada por textos que tragam o signo terminológico inserido em contextos explicativos ou definitórios, gerando assim o vocabulário de especialidade. Como exemplos de contexto explicativo e definitório para os termos *language* e *Linguistics* em inglês temos:

1. *Language is one of the social possessions which most obviously reflects the internal differentiation of human societies.* (em negrito)
2. *Linguistics is also one of the social sciences, in that the phenomena forming its subject-matter are part of the behaviour of men and women in society, in interaction with their fellows.* (em negrito)

Para termos em português citamos os exemplos de **língua** e **linguística**:

3. A língua é o **mecanismo que permite ao emissor da mensagem a associação de um conteúdo mental (a idéia) a uma expressão material (letras, sinais, sons).**
4. A linguística, como se sabe, **trata dos textos em prosa, ou seja, dos que se servem do plano da expressão apenas como porta de acesso ao plano do conteúdo.**

Os exemplos acima exemplificam dois tipos de contextos definitórios, com os verbos SER e TRATAR. Sendo que os exemplos 1-3 definem por meio do verbo ser, enquanto que o que usa o verbo TRATAR define explica um dos tipos de textos dos quais a Linguística trata: os textos em prosa.

Quanto ao trabalho de digitalização das obras é mais árduo do que a busca na internet, por ser (i) mais mecânico e pelo (ii) fato de que a tecnologia disponível até o momento não é tão rápida. A variável que temos que considerar quanto ao tempo é se a imagem escaneada ficou

legível ou embaçada ou distorcida. Caso a imagem estivesse imperfeita, era necessário refazer o procedimento, para que, posteriormente, as obras digitalizadas, em formato .PDF, pudessem ser decodificadas e transformadas em arquivos .txt pelo programa *OmniPage* 16.

2.2. Escolha de candidatos a termo

Os textos e manuais serão organizados no computador, em diretórios distintos, separados por idioma e seguindo a ordem proposta pela árvore de domínio, sendo essa ordem estabelecida em domínio e subdomínios; ou pelo nome do manual. Eles terão cabeçalhos, nos quais constarão os endereços de origem dos textos, bem como a data de coleta ou digitalização dos manuais.

Assim feito, os *corpora* provenientes de manuais e dos artigos científicos serão lidos artificialmente, por meio do programa WST, e uma **lista de palavras** será gerada. O gerador de lista de palavras do WST será usado para se criar uma lista de todas as palavras diferentes com suas respectivas frequências. Essa ferramenta listará as palavras mais recorrentes dentro dos *corpora* em ordem decrescente.

Em seguida, essa lista de palavras será comparada com um corpus de referência de língua geral para ser refinada e chegaremos ao vocabulário de especialidade por meio de uma **lista de palavras-chave**. A lista de palavras-chave permite o levantamento de índices linguísticos que caracterizam um *corpus* de estudo quando este é comparado a outro de referência, no mínimo cinco vezes maior.

Caso a palavra-chave esteja inserida em contextos definitórios ou explicativos, ela poderá ser útil para compor a **lista de candidatos a termos**.

Os candidatos a **termos** servem para dar forma ao banco de dados de verbetes do vocabulário em português e em inglês. Farão parte desse banco: **substantivos** e **unidades sintagmáticas** ou **fraseologias** (*clusters*), cujos traços conceituais serão extraídos dos próprios *corpora* para comporem as definições dos verbetes do vocabulário nas fichas terminológicas *online*. A criação de **fichas terminológicas** no VoTec¹¹ será monolíngue equivalente, e serão usadas posteriormente para formar

¹¹ Vide nota de rodapé 4.

o banco de dados de verbetes.

Neste momento da pesquisa será usado o **Concordanciador**, que no WST é chamado *Concord*, para buscar em qual contexto ou co-texto o candidato a termo se encontra, como se insere nele e se atende aos quesitos anteriores: ser apresentado em um contexto definatório/explicativo claro e objetivo.

Na parte final do desenvolvimento do trabalho, os **traços conceituais** são organizados em fichas terminológicas. Os contextos nos quais as unidades terminológicas se inserem serão copiados e salvos. Posteriormente serão visualizados pelos usuários. As relações de hiperonímia, hiponímia e sinonímia serão observadas e igualmente registradas nas fichas terminológicas, caso existam. Após o lançamento desses dados, dados externos são pesquisados e agregados ao banco de dados. São eles: definição do termo, extraída de uma obra lexicográfica ou terminológica de acesso livre *online*, e dados enciclopédicos (Veja Figura 3). Também serão registrados nessas fichas dados possíveis como: informações morfológicas, sintáticas, semânticas e relativas ao *corpus* de origem. Esses dados poderão ser comparados com outros já existentes em dicionários.

Quando o pesquisador parte da do **signo terminológico** para a identificação dos traços distintivos que o constituem, temos o chamado **percurso semasiológico** (BARROS, 2004). Contextualizando o percurso semasiológico neste trabalho, podemos dizer que as listas de palavra-chave, produzidas pelo WST, permitem que o pesquisador possa identificar os termos. Em seguida, é possível identificar e analisar os contextos nos quais esses se inserem para a identificação dos traços conceituais. A extração desses traços conceituais, organizados nas fichas terminológicas, leva à construção da **definição terminológica final**.

O VoTec é uma ferramenta que oferece várias formas de visualização: **normal** e **descritiva**, e as consultas podem ser nos módulos total, tradutor e modular, dependendo do objetivo do consultante. A visualização **normal total** é o formato que segue o padrão dos dicionários impressos; dados como **NOTA** e exemplos e as relações de sinonímia, hiponímia, antonímia, bem como dados do termo no *corpus* vêm em ordem ininterrupta. O modo **descritivo** apresenta os dados em forma hierárquica, conforme mostra a Figura 3 a seguir.

<p>Tipos de Exibição Normal Descritiva</p> <hr/> <p>Tipos de Consulta Total Tradutor Modular</p> <hr/> <p>Consultas Externas Corpus NILC Google Answers.com Wikipedia CORTEC</p>	<p>▼ Português</p>
	<p>Voltar ao resultado da busca</p> <p>Gramática</p> <p>ciência que estuda e registra os fatos de uma língua falada ou escrita, em tempo e espaço específicos, podendo ser relacionada com outra análoga.</p> <p>NOTA: modo justo de falar e escrever uma língua, dividido em Fonética, Morfologia e Sintaxe</p> <p>Abreviação/Acrônimo: <i>Nada encontrado</i></p> <p>Categoria Gramatical: substantivo</p> <p>Gênero: feminino</p> <p>Número: singular</p> <p>Cópus</p> <ul style="list-style-type: none"> • Posição na Ordem de Freqüência: 158 • N° de Ocorrências do termo: 563 <p>Ontologia: Linguística Histórica</p> <p>Variações Morfossintáticas: <i>Nada encontrado</i></p> <p>Sinônimos: ciência de letras</p> <p>Antônimos: <i>Nada encontrado</i></p> <p>Hipônimo de: processos mentais universais; linguagem; língua</p> <p>Co-hipônimos: <i>Nada encontrado</i></p> <p>Hiperônimo de: categorias gramaticais; sons e letras; palavras; Fonética; Morfologia; Sintaxe</p> <p>Exemplos</p> <ol style="list-style-type: none"> 1. Na acepção mais tradicional da linguística oitocentista, como vimos, as línguas engendram suas mudanças (portanto, a mudança é fundamentalmente interna, e o sistema, instável); aqui, o objeto-língua (a gramática) é visto como estável - e a mudança, portanto, como externamente motivada. 2. Vamos ressaltar um último ponto importante: na perspectiva gerativista, a gramática é o lugar da estabilidade; a língua, não; ainda segundo D. Lightfoot, As gramáticas, portanto, são entidades reais, biológicas, representadas nas mentes de cada indivíduo.

Figura 3. Visualização parcial do termo **Gramática** no modo **Descritivo**. (Fonte: o autor)

Na Figura 3, podemos ver os dados apresentados de forma separada, facilitando ao leitor a localização de dados específicos como categoria gramatical, posição no *corpus*, entre outros, mais rapidamente.

2.3. Macro e microestrutura

Em relação à macro e à microestrutura do Vocabulário da Linguística que temos em mente, os verbetes serão buscados de forma eletrônica e serão disponibilizados para visualização na página inicial do VoTec.

A organização de dados como categoria gramatical, a definição, posição do termo no *corpus* farão parte da microestrutura. A organização dos conceitos que comporão a estrutura da definição terminológica poderá ser feita seguindo-se o padrão GPDE – **gênero próximo** e **diferença específica** (FINATTO, 2001, p. 119) e uma possibilidade é a

definição terminológica.

O padrão GPDE rege que os conceitos sejam organizados de uma relação de hiperonímia para a de hiponímia, ou seja, que se parta das informações mais gerais para as mais específicas em uma oração. Nesse caso, para o termo **gramática**, temos “ciência que estuda e registra (hiperonímia) os fatos de uma língua falada ou escrita (hiponímia), em tempo e espaço específicos, podendo ser relacionada com outra análoga (hiponímia)”.

Caso o *corpus* traga mais informações do que as que serão usadas na definição de padrão terminológico, disponibilizaremos um campo à parte reservado para **NOTA**. Isto se realizará se o paradigma definicional selecionado pelo público-alvo, os alunos iniciantes do Curso de Letras, for o terminológico. No campo de NOTAS, organizaremos a definição que se aproximará da definição lexicográfica ou enciclopédica, sendo que a primeira foca em uma descrição linguística do termo, e a segunda, a descrição da coisa ou fenômeno. Essa é uma possibilidade surgida da necessidade de implementação das definições, caso o *corpus* traga esse tipo de informação.

Após o registro dos termos e definições, faremos a pesquisa com o público-alvo e com os profissionais da área da Linguística, por meio de questionários com perguntas específicas. Essa pesquisa será realizada, via Internet, por meio do formulário do *Google Docs*, pois este sistema permite a coleta de dados que podem ser categorizados em uma planilha. Os dados consistirão no parecer técnico dos especialistas quanto à disposição dos conceitos dentro dos paradigmas definicionais. Dependendo da opinião dos profissionais, esses conceitos podem sofrer algumas alterações, mas não de forma a invalidar os dados provenientes dos *corpora* nem o paradigma definicional escolhido pelo público-alvo na fase inicial da pesquisa.

3. *Considerações Finais*

Neste capítulo, apresentamos o trabalho em desenvolvimento da construção do Vocabulário da Linguística Geral. Descrevemos os passos que tem composto nossa prática terminológica e como temos feito a padronização dos *corpora* acadêmicos já compilados, seguindo as subáreas da árvore de domínios da Linguística de Fromm (2015). Até o momento, temos o *corpus* de manuais em inglês e português compilados

e já disponíveis para processamento linguístico pelo WST6.0. Compilados o *corpus* das subáreas em inglês e português, concluiremos a padronização e será dado andamento à construção do vocabulário de Linguística Geral como descrito neste trabalho. A padronização de *corpora* – é parte do processo metodológico na criação de obras terminográficas. Por padronização, entende-se que os corpora sejam equilibrados em termos de tamanho, tipo e que suas características sejam as mesmas. Isto é, número de palavras, mesmo tipo (artigos científicos, dissertações, teses, manuais de Linguística), e no nosso caso atendam às características listadas no Quadro 1 deste artigo.

Esperamos contribuir para a comunidade de alunos ingressantes e professores do curso de Letras, bem como profissionais da área de Linguística e Tradução. Embasados na LC e na TCT, acreditamos que a metodologia e teoria que embasam este trabalho atendam ao rigor teórico-metodológico da Terminologia, bem como do fazer terminográfico para ser útil ao público lusófono e anglófono usuários de obras terminográficas bilíngues.

REFERÊNCIAS BIBLIOGRÁFICAS

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

BARROS, L. A. *Curso básico de terminologia*. São Paulo: EDUSP, 2004.

_____. *Conhecimentos de terminologia geral para a prática tradutória*. São José do Rio Preto: NovaGraf, 2007.

BAGNO, M.; RANGEL, E. O. Tarefas da educação linguística no Brasil. In: *Revista Brasileira de Linguística Aplicada*. V. 5, n. 1, 2005. p. 63-81. Disponível em: <<http://www.scielo.br/pdf/rbla/v5n1/04.pdf>>. Acesso em 01 maio 2018.

BARBOSA, M. A. Lexicologia, Lexicografia, Terminologia, Terminografia, Identidade científica, Objeto, Métodos, Campos de atuação. In: *Simpósio latino-americano de terminologia e I Encontro brasileiro de terminologia técnico-científica*. 2., 1990, Brasília, Anais. p. 152-158

CABRÉ, M. T. *La Terminología: representación y comunicación*. Barcelona: IULA/UPF, 1999.

_____. El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en terminología (I). In: *Organon*. Universidade Federal do Rio Grande do Sul, Instituto de Letras, vol.1, n.1 (1956). Porto Alegre: UFRGS, semestral, 2011.

CASTILHO, A. T. As letras no ensino e na pesquisa. In: *Veredas on line – ensino*. PPG – Linguística / UFJF: V. 2, 2007. P.05-21. Disponível em: <<http://www.ufjf.br/revistaveredas/files/2009/12/artigo011.pdf>>. Acesso em 01 maio 2018.

DORNELLES, M. S. *Bases teórico-metodológicas para elaboração de um glossário bilíngue (português-inglês) de treinamento de força: subsídios para o tradutor*. Disponível em <<http://hdl.handle.net/10183/117567>>. Acesso em 01 maio 2018.

DUBUC, Robert. *Manuel pratique de terminologie*. 3e éd. Montréal : Linguatex, 1992.

FINATTO, M. J. B. Elementos lexicográficos e enciclopédicos na definição terminológica: questões de partida. In: KRIEGER, M. G.; MACIEL, A. M. B. (org.) *Temas da Terminologia*. Porto Alegre/São Paulo: Ed. Universidade/UFRGS/Humanitas/USP, 2001.

FROMM, G. *VoTec: a construção de vocabulários eletrônicos para aprendizes de tradução*. São Paulo, 2007. Tese (Doutorado em Estudos Linguísticos e Literários em Língua Inglesa). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

_____; 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.

GAUDIN, F. La sociterminologie. In: *Languages*, n. 157, Larousse, p.80-92, 2005.

HALLIDAY, M. A. K. Corpus studies and probabilistic grammar. In: K. AIJMER; B. ALTENBERG (org.). *English corpus linguistics: Studies in honour of Jan Svartvik*. London: Longman, 1991.

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

PAVEL, S.; NOLET, D. *Manual de terminologia*. Tradução de Enilde Faulstich. Canadá: Departamento de Tradução, 2002.

SARDINHA, Tony Berber. *Linguística de Corpus*. São Paulo: Manole, 2004.

SCOTT, M. *WordSmith Tools*. Versão 6, 2012. Disponível em: <<http://lexically.net/wordsmith/version6/>>. Acesso em 01 maio 2018.

TAGNIN, S. E. O. *Corpora na tradução*. São Paulo: Hub Editorial, 2015.

TALAVÁN, Noa. *A University Handbook on Terminology and Specialized Translation*. Madrid: Universidad Nacional de Educación a Distancia, 2016.

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

VIANA, V. e TAGNIN S. E. O. *Corpora no ensino de línguas estrangeiras*. São Paulo: Hub Editorial, 2010.

YAMAMOTO, M.I. Linguística Histórica e Linguística de Corpus: uma proposta para dicionário bilingue português-inglês. In: *Revista Philologus*. Rio de Janeiro, Ano 19, n.55 Suplementos, p.691-705. Disponível em <<http://www.filologia.org.br/rph/ANO19/55SUP/058.pdf>>. Acesso em 01 maio 2018.