



**UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE LETRAS E LINGUÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM
ESTUDOS LINGUÍSTICOS**



MÁRCIO ISSAMU YAMAMOTO

**VOBLING - VOCABULÁRIO BILÍNGUE DE LINGUÍSTICA,
PORTUGUÊS-INGLÊS, GUIADO POR *CORPUS***

**UBERLÂNDIA
2019**

MÁRCIO ISSAMU YAMAMOTO

**VOBLING - VOCABULÁRIO BILÍNGUE DE LINGUÍSTICA,
PORTUGUÊS-INGLÊS, GUIADO POR *CORPUS***

Tese de doutorado apresentada ao Programa de Pós-Graduação em Estudos Linguísticos do Instituto de Letras e Linguística da Universidade Federal de Uberlândia como requisito parcial para obtenção do título de Doutor em Estudos Linguísticos.

Área de concentração: Linguística Teórica e Descritiva

Linha de pesquisa: Teoria, descrição e análise linguística

Orientador: Prof. Dr. Guilherme Fromm

UBERLÂNDIA
2020

MÁRCIO ISSAMU YAMAMOTO

**VOBLING - VOCABULÁRIO BILÍNGUE DE LINGUÍSTICA,
PORTUGUÊS-INGLÊS, GUIADO POR *CORPUS***

Tese de doutorado apresentada ao Programa de Pós-Graduação em Estudos Linguísticos do Instituto de Letras e Linguística da Universidade Federal de Uberlândia como requisito parcial para obtenção do título de Doutor em Estudos Linguísticos.

Área de concentração: Linguística Teórica e Descritiva

Linha de pesquisa: Teoria, descrição e análise linguística

Uberlândia, 28 de agosto de 2020.

Banca Examinadora:

Profa. Dra. Maria José Bocorni Finatto – UFRGS/RS

Profa. Dra. Stella Esther Ortweiler Tagnin – USP/SP

Profa. Dra. Valeska Virgínia Soares Souza – UFU /MG

Prof. Dr. Ariel Novodvorski – UFU /MG

Prof. Dr. Guilherme Fromm – UFU/MG – Orientador

AGRADECIMENTOS

Ao professor Guilherme, orientador, mestre, mentor e amigo sempre presente. Ao professor Ariel Novodvorski, que desde 2009 me acompanha nesta carreira acadêmica. Aos professores Maria José Bocorny Finatto, Stella Esther Ortweiler Tagnin, Valeska Virgínia Soares Souza, Eliana Dias e Eduardo Batista da Silva que se dispuseram a compor esta banca de defesa.

À Universidade Federal de Goiás e à Universidade Federal de Jataí que me concederam o afastamento para cursar este doutorado. À Universidade Federal de Uberlândia, ao Instituto de Letras e Linguística, ao Programa de Pós-graduação em Linguística que há 20 anos têm contribuído para a minha formação acadêmica e profissional. À UFU, à UFG e à UFJ, pois, através delas, tive acesso ao acervo bibliográfico de bibliotecas internacionais excelentes.

Aos professores de Linguística da UFU, prof. Frederico de Sousa Silva e prof. Guilherme Fromm, e da UFJ, prof. Fabiano Silvestre Ramos e prof. Silvio Ribeiro da Silva dos Cursos de Letras-Português e Letras-Inglês, e aos seus alunos, pois me permitiram conduzir a pesquisa inicial em seus ambientes de ensino e aprendizagem.

Ao colega Fernando Oliveira pela permissão e assistência no uso da plataforma ToGatherUp. Aos membros do grupo GPELC e PLEX, sempre presentes nesta caminhada rumo ao conhecimento, pesquisa e descobertas.

À CAPES/MEC pelo apoio financeiro, por meio da bolsa de doutorado bastante útil para a condução desta pesquisa em sua fase final.

A todos os familiares, aos amigos e colegas que, de uma forma ou de outra, contribuíram para a realização desta pesquisa.

Muito obrigado!

RESUMO

Neste trabalho objetivamos a elaboração de um protótipo de Vocabulário Bilíngue na área de Linguística, denominado VoBLing. Esta obra foi realizada em português e em inglês, baseada em *corpus* comparável, posteriormente disponibilizada na plataforma do VoTec – Vocabulário Técnico On-line (FROMM, 2007, 2018), de acesso livre. As teorias que embasam este trabalho são as de Barbosa (1990), sobre o conceito de Vocabulário; de Cabré (1999), com a Teoria Comunicativa da Terminologia, doravante TCT; Finatto (2001a), Krieger e Finatto (2004) e Barros (2004) sobre a Terminologia em relação à elaboração das definições dos verbetes. Na metodologia, adotamos os pressupostos teórico-metodológicos da Linguística de *Corpus* (BERBER SARDINHA, 2004; 2009; MCENERY; HARDIE, 2012; SCOTT, 2016; 2020) e os parâmetros da plataforma terminológica do VoTec (FROMM, 2007; CARDOSO, 2017). Os *corpora* utilizados como fonte para extração de dados para escolha dos termos e construção das definições são acadêmicos, escritos na área de Linguística: teses, dissertações, artigos científicos, manuais e coletâneas de Linguística. Parte dos *corpora* foi compilada por discentes de graduação, pós-graduação e pelo pesquisador, em disciplinas ministradas pelo professor Fromm, e como parte do processo metodológico desta pesquisa. Já os manuais foram compilados pelo pesquisador, desde a digitalização dos manuais em língua portuguesa (LP), até a disponibilização desse material em arquivos txt; e a compilação da internet para os manuais em língua inglesa (LI). O processamento dos textos foi feito pelo console do *WordSmith Tools* 7.0 e 8.0 (SCOTT, 2016, 2020), por meio de suas ferramentas de lista de palavras, lista de palavras-chave e concordanciador. A proposta original da árvore de domínio da Linguística é de Fromm (2008) a qual foi atualizada por Fromm e Yamamoto (2020), com 47 subáreas. A organização dos traços distintivos na plataforma do VoBLing foi baseada na Semântica de *Frames*, de Charles J. Fillmore (1975; 1976; 2003 e 2005) e Souza (2019) e na Terminologia de *Frames*, segundo Faber Benítez et al. (2005). Como resultado, o VoBLing disponibiliza 200 verbetes bilíngues, metade dos quais correspondem à área da Linguística, da Linguística Descritiva, da Linguística Aplicada e suas subáreas; a outra metade corresponde aos termos-chave da Linguística e suas subáreas.

Palavras-chave: VobLing. Vocabulário bilíngue de Linguística. Terminologia. Terminografia. Linguística de *Corpus*.

ABSTRACT

In this research we aimed to elaborate the prototype of a Bilingual Vocabulary in the fields of Linguistics, named VoBLing. This work was conducted in Portuguese and in English, based on comparable *corpus*, and made it available on VoTec - Online Technical Vocabulary (FROMM, 2007, 2018) platform available on the internet, for free access. The theories that support this research are those of Barbosa (1990), on the concept of Vocabulary; Cabré (1999), on the Communicative Theory of Terminology, henceforth TCT; Finatto (2001a), Krieger and Finatto (2004) and Barros (2004), on Terminology in relation to the elaboration of entry definitions. For the methodology, we adopted the theoretical-methodological assumptions of *Corpus Linguistics* (BERBER SARDINHA, 2004, 2009; MCENERY; HARDIE, 2012; SCOTT, 2016; 2020) and VoTec terminology platform parameters (FROMM, 2007). The *corpora* used as source for data extraction to choose the terms and definition constructions are specialized, written, subdivided into two different types in Linguistics: (1) PhD theses, master's dissertations and scientific articles; and (2) Linguistics manuals and Linguistics collections. Part of these texts were compiled by undergraduate students, graduate students and by this researcher, in disciplines taught by Professor Fromm, and as part of the methodological step of this research. Data were processed by the *Wordsmith Tools* 7.0 and 8.0 console (SCOTT, 2016, 2020), using its Wordlist, keyword list, and concordance tools. The original proposal of the domain tree of Linguistics is by Fromm (2008), updated by Fromm and Yamamoto (2020), comprising 47 subdomains. The organization of distinctive features on VoBLing was based on Frame Semantics, by Fillmore (1975; 1976; 2003 and 2005), Souza (2019) and on Frame-based Terminology, according to Faber Benítez et al. (2005). As a result, VoBLing makes available 200 bilingual entries, half of them corresponds to Linguistics, Descriptive Linguistics, Applied Linguistics, and its subareas; the other half corresponds to the key terms of Linguistics and its subareas.

KeyWords: VoBLing. Bilingual Vocabulary of Linguistics. Terminology. Terminography. *Corpus Linguistics*.

LISTA DE FIGURAS

Figura 1 - Percurso onomasiológico (POTTIER, 1992).....	43
Figura 2 - Árvore de domínio da Linguística alterada com base em Fromm (2015)	70
Figura 3 - Proposta 1 - PD enciclopédico baseado em <i>Frames: ecolinguística</i>	80
Figura 4 – Proposta 2 - PD enciclopédico baseado em <i>Frames: língua</i>	81
Figura 5 - Proposta 3 - PD enciclopédico baseado em <i>Frames: biolinguística</i>	81
Figura 6 - Proposta 4 - PD enciclopédico baseado em <i>Frames: análise da conversação</i>	82
Figura 7 - Dados estatísticos dos <i>corpora</i> de manuais.	87
Figura 8 - Plataforma <i>ToGatherUp</i> : página de acesso.	96
Figura 9 - Plataforma <i>ToGatherUp</i> : página inicial.	97
Figura 10 - Plataforma <i>ToGatherUp</i> : cadastro de texto.	97
Figura 11 - Plataforma <i>ToGatherUp</i> : visualização do <i>corpus</i> de português.	99
Figura 12 - Plataforma <i>ToGatherUp</i> : organização dos arquivos baixados.	100
Figura 13 - <i>ToGatherUp</i> : visualização final dos arquivos em txt.	100
Figura 14 – <i>ToGatherUp</i> : arquivo com cabeçalho corrompido pós-download de base Windows.	103
Figura 15 – <i>ToGatherUp</i> : arquivo carregado e baixado em UTF-16 LE de base Windows..	104
Figura 16 - Programa Notepad++ - <i>layout</i> da página inicial.	105
Figura 17 - Notepad++ abertura de arquivo.	105
Figura 18 - Notepad++ visualização do cabeçalho e do texto ANSI de base Unix.....	106
Figura 19 – <i>ToGatherUp</i> : arquivo com texto corrompido pós-download de base Unix.....	107
Figura 20 – <i>ToGatherUp</i> : arquivo com texto corrigido pós-download de base Unix ANSI para UTF-8.	107
Figura 21 – Linguística: pastas com o <i>corpus</i> em português.	108
Figura 22 - Visão parcial da codificação do <i>subcorpus</i> de Linguística em português: subárea da Aquisição da Linguagem.	111
Figura 23 - <i>Subcorpus</i> com problemas de decodificação em português.	112
Figura 24 - Lista de palavras com arquivos padronizados: Unicode.....	113
Figura 25 - <i>Plot</i> do termo <i>discurso</i> no <i>subcorpus</i> de Análise do Discurso em português.	115
Figura 26 - Lista de palavras-chave da Análise do Discurso (parcial – português).	116
Figura 27 - Nome do arquivo identificado a partir da palavra-chave <i>Words</i> no concordanciador.	118
Figura 28 - <i>Subcorpus</i> da Análise da Conversação redimensionado: pós-limpeza (visão parcial).	122
Figura 29 - Listas de palavras-chave da Linguística, LD e LA: pré-limpeza (visão parcial).	125
Figura 30 - Termos chave da Linguística (10 termos).....	126
Figura 31 - Lista de palavras-chave da LD e LA (visão parcial).	126
Figura 32 - VoBLing: visualização parcial do ST <i>linguística</i> no modo Descritivo.	127
Figura 33 - VoBLing: tela inicial de registro de termos.....	128
Figura 34 - VoTec: tela de cadastro e ontologia do termo <i>etymology</i> em 2015.	128
Figura 35– VoBLing: tela de cadastro e ontologia dos termos em 2020.	129
Figura 36 - VoBLing: cadastro da ontologia do termo <i>ecossistema</i>	129
Figura 37 – VoBLing: tela de cadastro de Contextos.....	130
Figura 38 - VoBLing: visualização da aba Dados – <i>linguística descritiva</i>	131
Figura 39 - VoBLing: visualização da aba Traços distintivos – <i>linguística descritiva</i>	133
Figura 40 – VoBLing: visualização da aba Semântica - <i>linguística descritiva</i>	134
Figura 41 – VoBLing: termo equivalente - <i>linguística descritiva</i>	134
Figura 42 – VoBLing: aba Termos remissivos do ST - <i>linguística descritiva</i>	135
Figura 43 – VoBLing: aba Informações enciclopédicas - <i>linguística descritiva</i>	136

Figura 44 - VoBLing: visualização da aba Multimídia - <i>linguística descritiva</i> .	136
Figura 45 - VoBLing: visualização da aba Dados/áudio - <i>linguística descritiva</i> .	137
Figura 46 – VoBLing: imagens da Linguística e LD.	138
Figura 47 - VoBLing: Campo Nota.	138
Figura 48 - VoBLing: visualização normal do termo <i>linguística</i> .	140
Figura 49 – VoBLing: traços distintivos de <i>linguística</i> (em colunas).	141
Figura 50 – VoBLing: visualização parcial do STF Linguística Descritiva no modo normal.	142
Figura 51 – VoBLing: definição terminológica do ST <i>biolinguística</i> .	144
Figura 52 – VoBLing: definição terminológica do ST <i>etymology</i> .	145
Figura 53 – VoBLing: definição enciclopédica do ST <i>etymology</i> .	146
Figura 54 – VoBLing: definição enciclopédica do ST <i>ecolinguistics</i> .	147

LISTA DE QUADROS

Quadro 1 -Taxonomia de <i>corpora</i> segundo Teixeira (2008).....	55
Quadro 2 - Dicionários de Linguística analisados (físicos, eletrônicos ou virtuais) em ordem de publicação.....	58
Quadro 3 - Quantificação de dados na área da Linguística disponíveis na internet (2015).	63
Quadro 4 - Quantificação de dados na área da Linguística disponíveis na internet: artigos científicos/ <i>scholarly articles</i> (2019).....	63
Quadro 5 - Proposta 1 - PD terminológico baseado em <i>Frames: linguística/linguistics</i>	76
Quadro 6 - Proposta 2 - PD terminológico baseado em <i>Frames: linguística aplicada/applied linguistics</i>	77
Quadro 7 - Proposta 3 - PD terminológico baseado em <i>Frames: teletandem/tandem language learning</i>	79
Quadro 8 - Taxonomia dos <i>corpora</i> de artigos científicos: Linguística (português-inglês)	85
Quadro 9 - Manuais de Linguística em inglês e português	86
Quadro 10 - Problemas ortográficos corrigidos no <i>corpus</i> de manuais em português – junho/2018.....	93
Quadro 11 - Dimensionamento do tempo de digitalização dos <i>corpora</i> (parcial).	95
Quadro 12 - Dimensionamento dos <i>subcorpora</i> : significados e cores.....	109

LISTA DE TABELAS

Tabela 1 - Resultado da pesquisa dos paradigmas definicionais com discentes da UFU e da UFJ	68
Tabela 2 - Dimensionamento do <i>corpus</i> de manuais em LP e em LI.....	72
Tabela 3 - Dimensionamento dos <i>corpora</i> em LP e em LI.	73
Tabela 4 – Dimensionamento dos subcorpora de Linguística em LP.	124

LISTA DE ABREVIATURAS E SIGLAS

DT – Definição Terminológica
GPDE – Gênero Próximo e Diferença Específica
ISO – *International Organization for Standardization*
LA – Linguística Aplicada
LC – Linguística de *Corpus*
LD – Linguística descritiva
LN – Língua natural
LP – Língua portuguesa
LI – Língua inglesa
PD – Paradigma Definicional
PE – Paradigma Enciclopédico
PN – Pesquisa Narrativa
ST – Signo Terminológico
STF – Signo Terminológico Fraseológico
TCT – Teoria Comunicativa da Terminologia
TGT – Teoria Geral da Terminologia
TSCT – Teoria Sociocognitiva da Terminologia
UL – Unidade lexical
UT – Unidade terminológica
VoBLing – Vocabulário Bilíngue de Linguística
VoTec – Vocabulário Técnico On-line
WST – *WordSmith Tools*

SUMÁRIO

1. INTRODUÇÃO	14
1.1. Objetivos	18
1.1.1. Objetivo geral	18
1.1.2. Objetivos específicos	18
1.2. Justificativa	18
1.3. Organização da tese	19
2. FUNDAMENTAÇÃO TEÓRICA.....	22
2.1. Vocabulário versus Dicionário.....	22
2.2. Terminologia e Terminografia: princípios teórico-metodológicos	23
2.2.1. Termo: concepções	25
2.2.2. Teoria Geral da Terminologia – TGT.....	29
2.3. Novas Teorias da Terminologia	31
2.3.1. Teoria Comunicativa da Terminologia – TCT	31
2.3.2. Semântica e Terminologia de <i>Frames</i>	33
2.3.2.1 Semântica de <i>Frames</i>	33
2.3.2.2 Terminologia de <i>Frames</i>	36
2.4. Terminografia.....	38
2.4.1. Definições	39
2.4.2. Paradigma definicional	41
2.4.3. Percorso onomasiológico e semasiológico	41
2.4.4. Contextos definitório, explicativo e associativo.....	45
2.4.5. O signo terminológico (ST).....	46
2.4.6. Homonímia e polissemia	48
2.4.7. Árvore de domínio: princípios teóricos	51
2.5. Linguística de Corpus e corpus.....	52
2.5.1. Linguística de <i>Corpus</i> : histórico	53
2.5.2. <i>Corpus</i> : definição e taxonomia.....	55
3. OBRAS DE REFERÊNCIA: BREVE ESTADO DA ARTE	57
3.1. Algumas considerações.	62
4. METODOLOGIA.....	63
4.1. Etapas da pesquisa	64
4.1.1. Contextualização	65
4.1.2. Perfil do público-alvo	65
4.1.3. Realização da pesquisa	66
4.1.4. Resultado da pesquisa.....	67
4.2. Elaboração da árvore de domínio da Linguística	68
4.3. Compilação e tratamento do corpus.....	70
4.4. Extração de dados do corpus	73
4.5. Elaboração dos paradigmas definicionais.....	74
4.5.1. Proposta de PD baseado em <i>Frames</i>	76
4.5.1.1 PD terminológico baseado em <i>Frames</i>	76
4.5.1.2 PD enciclopédico baseado em <i>Frames</i>	80
5. COMPILAÇÃO E TRATAMENTO DO CORPUS.....	83
5.1. Corpora de textos acadêmicos.....	83
5.1.1. Digitalização e correção dos <i>corpora</i>	93
5.1.2. A plataforma <i>ToGatherUp</i>	95
5.1.3. A plataforma <i>ToGatherUp</i> vs WST: codificação textual	101
5.2. Normalização dos corpora.....	108

5.2.1.	Padronização do <i>corpus</i>	111
5.2.2.	Padronização do tamanho dos <i>corpora</i>	113
5.2.3.	Limpeza dos <i>corpora</i>	120
5.2.4.	Linguística Descritiva – dimensionamento dos <i>corpora</i> em português e em inglês 123	
5.2.5.	Linguística Aplicada – dimensionamento dos <i>corpora</i> em português e em inglês 123	
5.3.	Extração e tratamento de dados do corpus: WST	123
5.4.	Tratamento do corpus - VoBLing.....	127
6.	O PARADIGMA DEFINICIONAL (PD) BASEADO EM FRAMES.	140
7.	RESULTADOS	144
8.	CONSIDERAÇÕES FINAIS.....	150
	REFERÊNCIAS	153
	REFERÊNCIA DE MANUAIS DE LINGÜÍSTICA.....	161
	GRUPO 1: DICIONÁRIOS IMPRESSOS E ELETRÔNICOS.....	162
	GRUPO 2: DICIONÁRIOS DISPONÍVEIS ON-LINE.....	163
	APÊNDICES	164
	APÊNDICE A - Questionário de pesquisa para escolha do paradigma definicional.	164
	APÊNDICE B - Resultado da pesquisa com os alunos de Letras da UFU e da UFJ.	166
	APÊNDICE C - Dimensionamento de <i>corpora</i> de filologia, linguística histórica, linguística diacrônica e etimologia (2018).	170
	APÊNDICE D - Dimensão e codificação inicial do <i>corpus</i> da subárea da Linguística Descritiva (LD) em português.	171
	APÊNDICE E - Dimensão e codificação inicial do <i>corpus</i> da subárea da Linguística Aplicada (LA) em português.	172
	APÊNDICE F - Dimensão e codificação inicial do <i>corpus</i> da subárea da LD em inglês.	173
	APÊNDICE G - Dimensão e codificação inicial do <i>corpus</i> da subárea da LA em inglês.	174
	APÊNDICE H - Redimensionamento e codificação do <i>corpus</i> final da subárea da LD em português.	175
	APÊNDICE I - Redimensionamento e codificação do corpus final da subárea da LA em português.	176
	APÊNDICE J - Redimensionamento e codificação do <i>corpus</i> final da subárea da LD em inglês.	177
	APÊNDICE K - Redimensionamento e codificação do <i>corpus</i> final da subárea da LA em inglês.	178
	APÊNDICE L - VoBLing - lista de verbetes definidos.....	179
	APÊNDICE M - Roteiro para avaliação de dicionários ou glossários de Linguística impressos ou eletrônicos segundo Faulstich (2011).	182
	APÊNDICE N - Resultado da análise dos dicionários ou glossários de Linguística impressos e eletrônicos segundo Faulstich (2011).	183
	APÊNDICE O - Roteiro para avaliação de dicionários ou glossários de Linguística disponíveis on-line segundo Faulstich (2011).	196
	APÊNDICE P - Resultado da análise dos dicionários ou glossários de Linguística disponíveis on-line segundo Faulstich (2011).	197
	ANEXOS: Árvores de domínio da Linguística propostas por Fromm	204
	ANEXO A - Árvore de campo da Linguística - Fromm (2008).....	204
	ANEXO B - Árvore de campo da Linguística em discussão (2013).....	205
	ANEXO C - Árvore de campo da Linguística, segundo Fromm, atualizada até 2014.....	206
	ANEXO D - Árvore de domínio da Linguística (2018).	207

1. INTRODUÇÃO

Os registros da reflexão sobre a linguagem humana, na civilização ocidental, datam dos primórdios da cultura grega com os pré-socráticos e a escola de Alexandria. Rey (1995, p. 11) e Ducrot e Todorov (1972, p. 65) apontaram Platão e sua obra *Cratylus* como registro das discussões filosóficas sobre a linguagem na civilização helenística. Em contrapartida, no Oriente, há o registro da gramática de Panini (IV a.C.) com as regras dos procedimentos de derivação e de composição do sânscrito (DUCROT; TODOROV, 1972, p. 65-66). A partir de uma perspectiva científica e positivista, a Linguística consolidou-se com os estudos realizados por Saussure no século XX na Europa.

Vários cientistas e pensadores europeus e estadunidenses contribuíram para a consolidação da Linguística no Velho Continente e nas Américas. Da mesma forma, a escola linguística brasileira usufruiu desse conhecimento rumo ao estabelecimento, à consolidação e ao crescimento da Linguística como Ciência e disciplina acadêmica. Como Ciência, essa disciplina foi difundida no Brasil por meio dos trabalhos de Joaquim Mattoso Câmara Jr., a princípio.

Como disciplina acadêmica, ela teve seu início nas universidades brasileiras nos idos de 1950-1960, nos Cursos de Letras (VANDRESEN, s/d; BAGNO; RANGEL, 2005; CASTILHO, 2007; FIORIN, 2007, p. 145). Atualmente, há grandes centros como USP, UNICAMP, UFU, UFRGS, UNB que se dedicam ao estudo desta área no Brasil, com produções relevantes em nível nacional e internacional. Apesar da difusão da disciplina nas universidades brasileiras há mais de meia década, a maioria dos vocabulários¹ disponíveis, para uso dos acadêmicos e profissionais da área, são obras traduzidas, em geral, ou obras direcionadas a especialistas, não a leigos (FROMM; YAMAMOTO, 2020, p. 224).

Na comunicação cotidiana, usamos o léxico de língua geral, enquanto na comunicação especializada usamos a linguagem de especialidade ou vocabulário (BARBOSA, 1990). O domínio do vocabulário de áreas de especialidade permite uma comunicação mais eficiente do conhecimento técnico-científico por meio de sua terminologia entre profissionais (BARROS, 2004, p. 21). Já o registro dessa linguagem de especialidade, que faz parte do nível da **norma**

¹ Nesta tese, o conceito de vocabulário adotado é: o conjunto de vocábulos das áreas de especialidade, cujo significado é restrito e de alta frequência; quanto à conceituação, o vocabulário apresenta todas as acepções de um verbete dentro de uma área de especialidade ou grupo, de perspectiva sincrônica e sinfásica. Mais detalhes estão expostos no item 2.1 da fundamentação teórica.

e não do nível do **sistema** de uma língua (COSERIU, 1979, p. 56-69), é concretizado nos dicionários técnicos ou vocabulários.

Os vocabulários e os dicionários técnicos ou terminológicos definem os termos, vocábulos ou unidades lexicais especializadas. Nessas obras, os traços distintivos dos termos registram as características semânticas ou os semas que delimitam conceitualmente uma área da ciência ou do conhecimento.

Dada a apresentação inicial sobre o tema desta pesquisa, é importante dizer que a escolha desse tema não foi fortuita. A centelha de motivação que me levou a trabalhar com o vocabulário da Linguística foi plantada ainda na minha infância. Aos sete anos, meu pai me selecionou para ser seu leitor, já que ele era parcialmente cego. Isso aconteceu nos idos de 1981, quando eu, recém-alfabetizado, fui apresentado às obras lexicográficas e terminológicas da Língua Portuguesa, doravante LP, por um pai nipo-brasileiro. Devido à sua alfabetização em colônia japonesa no Brasil, onde todo o ensino se fazia em Língua Japonesa, seu conhecimento de português acadêmico era restrito. Tal conhecimento não lhe era suficiente para que pudesse acompanhar a leitura do material acadêmico na área de Filosofia e Teologia, necessária ao exercício da profissão de teólogo. Por consequência, grande parte da leitura que ele fizera então era de dicionários de Língua Portuguesa e vocabulários de Teologia. Hoje, em retrospectiva, posso compreender que meu pai foi meu mentor no manuseio das obras lexicográficas e terminográficas, fato este que me rendeu uma considerável familiarização com essas obras e com a arte de definir e conceituar os termos e o léxico da LP.

Os ventos da formação acadêmica e da atuação profissional me levaram a ser professor de inglês e, atualmente, como docente do curso de Letras, observo que os discentes enfrentam o mesmo dilema que eu havia experimentado quando infante: “Como lidar com os novos termos, estranhos à realidade do então limitado conhecimento de mundo que tinha?” Refletindo sobre esta indagação, acredito que, assim como os dicionários me propiciaram a construção dos conhecimentos outrora desconhecidos, uma obra terminográfica poderá servir aos discentes de Letras para se familiarizarem com os conceitos da área da Linguística.

Em função do meu interesse por esse tema e da familiaridade com que transito por ele, filiei-me como membro dos seguintes grupos de pesquisa: GPELC - Grupo de Pesquisa e Estudos em Linguística de Corpus - coordenado pelos prof. Dr. Guilherme Fromm e Prof. Dr. Ariel Novodvorski, e do PLex - Pesquisas em Léxico -, coordenado pelos Prof. Dr. Guilherme Fromm e Profa. Dra. Eliana Dias, a fim de dar sequência aos estudos e às investigações sobre esses temas.

Levando em consideração algumas questões, anteriores inclusive à pergunta inicial desta pesquisa, desenvolvemos um estudo analítico para responder se: haveria (há) vocabulários de Linguística direcionados para discentes dos Cursos de Letras? As obras disponíveis seriam (são) direcionadas para especialistas ou para leigos? Suas definições são de cunho lexicográfico, enciclopédico, terminológico ou de forma mista? De acordo com o estudo analítico conduzido por Fromm e Yamamoto (2020), a resposta a essas perguntas é negativa. Por conseguinte, o questionamento a ser respondido por esta pesquisa é: que tipo de vocabulário os alunos iniciantes do Curso de Letras precisam?

Nesse estudo analítico, Fromm e Yamamoto (2020) analisam obras terminológicas da área da Linguística e constataam a carência de dicionários nessa área direcionados para discentes leigos da Linguística (FROMM; YAMAMOTO, 2020, p. 220). Em virtude disso, optamos por desenvolver, neste estudo, uma obra que pudesse atender esses discentes leigos. Baseando-nos em princípios da Terminologia como fundamentação teórica e na Linguística de *Corpus*, doravante LC, como metodologia e abordagem, almejamos produzir uma obra com definições terminológicas e enciclopédicas concisas e objetivas que atenda à comunidade acadêmica de discentes de Letras e profissionais da Tradução.

Quanto ao caráter bilíngue da obra, português-inglês, o objetivo é propiciar aos acadêmicos o acesso às definições e exemplos em inglês, o que poderá contribuir para a redação de resumos e de textos na língua inglesa, doravante LI. A escolha por incluir definições em inglês deve-se à minha atuação profissional como docente de LI, o que me proporciona um certo conforto em trabalhar com essa língua.

Dessa maneira, nesta pesquisa, o objetivo foi construir o protótipo de um vocabulário bilíngue, em português e em inglês, na área da Linguística Geral. Essa obra incluirá a Linguística Descritiva, a Linguística Aplicada e suas subáreas. O embasamento teórico-metodológico dessa pesquisa terá contribuições da (i) Teoria Comunicativa da Terminologia (TCT), de Cabré (1999); dos (ii) princípios da Terminologia e Terminografia de Finatto (2001a) e de Krieger e Finatto (2004); e do (iii) princípio de vocabulário de Barbosa (1990). A LC será o suporte metodológico para a abordagem dos *corpora* (*corpus*, no singular) e para a análise quantitativa e qualitativa dos dados linguísticos (BERBER SARDINHA, 2004; 2009; MCENERY; HARDIE, 2012; TAGNIN; BEVILACQUA, 2013, p. 11-28; FROMM; YAMAMOTO, 2013, p. 129-152).

Analisaremos os dados provenientes de *corpora* da área de Linguística (em sua maioria), cujo tipo é o dissertativo² e o gênero é acadêmico (MOTTA-ROTH, 2009, p. 317-336) consistindo de teses, dissertações, artigos científicos, manuais e material instrucional³ (PARÉ; STARKE-MEYERRING; MCALPINE, 2009, p. 179-193).

A árvore de domínio⁴ foi um dos primeiros passos metodológicos desta pesquisa, baseados nos paradigmas da Terminologia, de sorte que nessa árvore de domínios da Linguística há três subdivisões: (i) a grande área da Linguística, (ii) a subárea da Linguística Descritiva e da Linguística Aplicada (YAMAMOTO, 2018) e, subordinadas a elas, haverá (iii) 47 subáreas no total. A título de exemplo, apresentaremos um possível quarto nível, nas subáreas da Semântica, buscando citar algumas teorias e/ou abordagens, e da Onomástica (com as ramificações da Toponímia e da Antroponímia).

A partir da análise quantitativa e qualitativa⁵ dos *corpora*, os seguintes passos foram cumpridos: (i) descrever, (ii) relacionar e (iii) classificar os conceitos de um dado signo terminológico, doravante ST. Esta tarefa foi realizada a partir dos contextos definitórios e dos contextos explicativos presentes no *corpus*, analisados para a extração de traços distintivos e, posteriormente, utilizados para conceituar termos selecionados da lista de palavras-chave. Essa lista de palavras-chave foi extraída do *WordSmith Tools* (SCOTT, 2016, 2020), doravante WST, com a aplicação de uma *stoplist* em LP e LI, seguindo os princípios da frequência de palavras e da chavicidade. De acordo com a pesquisa conduzida com o público-alvo, a definição terminológica, doravante DT, será a mais adequada e o percurso seguido será o semasiológico, ou seja, quando o ponto de partida é o termo e o ponto de chegada são os conceitos que o definem.

Concluindo esta subseção, nosso foco de trabalho é a construção de um protótipo de vocabulário bilíngue, português-inglês, de Linguística Geral, englobando a Linguística Descritiva (LD) e a Linguística Aplicada (LA) e seus subdomínios nesta pesquisa.

² Para classificar os *corpora* desta pesquisa, nos baseamos em Travaglia (2007) ao afirmar que os gêneros acadêmicos como “Tese, dissertação de mestrado, artigo acadêmico-científico, editorial de jornal, monografia, conferência, artigo de divulgação científica etc.” (TRAVAGLIA, 2007, p. 56) são predominantemente dissertativos.

³ Maior detalhamento desta descrição disponível no capítulo 4: Os *corpora* de estudo.

⁴ 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. Faz parte da tradição wüsteriana para representar hierarquicamente a estrutura ontológica e lógica de uma área de especialidade e para guiar a coleta e descrição da terminologia procurada (MACIEL, 2010, p. 407-408).

⁵ A análise quantitativa baseia-se no paradigma proposto por Fromm (2013, p. 1; FROMM; YAMAMOTO, 2013, p. 134), quando da conclusão de sua tese, na qual o pesquisador constatou que um *corpus* menor que quinhentas mil palavras não proveria traços semânticos suficientes para construção da definição terminológica. A análise qualitativa objetiva a identificação de contextos definitórios e explicativos, conforme proposto por Aubert (2001, p. 69), de onde extraímos os semas para a construção da definição dos termos.

1.1. Objetivos

No intuito de organizar e sistematizar as etapas desta pesquisa, apresentamos adiante os objetivos geral e específicos que nortearão a trajetória do protótipo de Vocabulário na área de Linguística Geral, português-inglês, considerando os princípios da Terminologia, da Terminografia e da LC.

1.1.1. Objetivo geral

O objetivo desta pesquisa foi a **construção** de um protótipo de um vocabulário bilíngue, português-inglês, bidirecional e em contraste, na área de Linguística Geral, com vistas a atender discentes dos anos ou semestres iniciais dos Cursos de Letras, provendo-lhes uma plataforma que conceitue e defina dois termos por subárea da Linguística, e permita-lhes a inserção no universo acadêmico dessa Ciência. Além da definição dos dois termos, elencamos os termos que melhor se encaixem nas subáreas da Linguística, baseado no princípio de maior chavicidade no *corpus* de estudo.

1.1.2. Objetivos específicos

Dentre os objetivos específicos para a confecção do **VoBLing**⁶, teremos:

- (i) propor um paradigma definicional para terminologia da Linguística baseado na Terminologia de *Frames*;
- (ii) descrever o estado da arte dos dicionários de Linguística disponíveis em português;
- (iii) definir 200 (duzentos) termos da Linguística na plataforma do **VoBLing** para consulta dos usuários.

1.2. Justificativa

Assim como outras áreas da Ciência, a Linguística detém um estatuto que garante sua especificidade, o que fica evidenciado em sua linguagem de especialidade que a diferencia das outras Ciências Humanas⁷. A especificidade da Linguística leva os discentes de Letras e os

⁶ Plataforma disponível em: <http://vobling.votec.ileel.ufu.br/>. Acesso em: 20 jul. 2020.

⁷ Em relação à especificidade da Linguística, é relevante considerar o fato de que a CAPES a classifica separadamente das Ciências Humanas. Disponível em http://www.capes.gov.br/images/documentos/documentos_diversos_2017/TabelaAreasConhecimento_072012_a_tualizada_2017_v2.pdf. Acesso em: 5 jun. 2020.

profissionais dessa área à necessidade de dominarem os termos e os conceitos que proporcionam a compreensão do que é a Linguística e suas subáreas. Sem que haja este conhecimento da terminologia e dos conceitos da Linguística, a compreensão da disciplina pode ser comprometida ou até mesmo ser um empecilho para os discentes ingressantes nos cursos de graduação em Letras seguirem adiante nesse curso de formação. A dificuldade do aluno ingressante foi observada na experiência do pesquisador como docente do ensino superior ao ouvir os depoimentos dos discentes relatando suas limitações de compreensão das disciplinas iniciais que tratam da Linguística.

Além dos fatores previamente citados, é relevante considerar que atualmente o inglês é uma *língua franca*. Jenkins (2002) define *língua franca* como a “língua utilizada como meio de comunicação por falantes provenientes de outras línguas que não a LI”⁸. Wiertlewska (2011) faz um breve histórico do inglês como língua estrangeira e seu estabelecimento como *língua franca*. A pesquisadora explica que, após a Segunda Guerra Mundial, a LI superou o ensino do alemão e do francês na Europa e hoje é a língua estrangeira mais ensinada na Comunidade Europeia. Por fim, a autora pontua que a LI atingiu o *status* de *lingua franca* atualmente porque é utilizada amplamente nos setores de comércio, educação e comunicação a nível mundial.

Fadanelli e Monzón (2017) discutem a difusão e a importância da LI concebida como *língua franca* com foco na difusão do conhecimento no Brasil. As autoras indicam a LI como a linguagem da ciência atualmente no mundo globalizado, especificam sua difusão na graduação e na pós-graduação no Brasil, no uso da linguagem técnica, apontam os discentes e docentes como público-alvo e como usuários. Para finalizar, as autoras se valem da LA, da LC e da Terminologia como base teórico-metodológica para o estudo de *datasheets*⁹ e de artigos acadêmicos no contexto de pós-graduação e de ensino do Inglês Instrumental.

1.3. Organização da tese

Nesta seção, informamos ao leitor como esta tese foi organizada a fim de atingir o objetivo de construir o VoBLing:

No capítulo 1, da Introdução, abordamos o histórico da Linguística, listamos nossos objetivos geral e específicos e justificamos a motivação desta tese.

⁸ “[...] it is a means of communication between people who come from different first language backgrounds” (JENKINS, 2012, p. 486).

⁹ Gêneros textuais da Eletrotécnica que trazem especificações sobre o funcionamento de dispositivos elétricos; *Datasheets* são geralmente disponibilizados na internet.

No capítulo 2, da Fundamentação teórica, encontramos cinco subdivisões. Na seção 2.1, justificamos a escolha pelo termo vocabulário em vez de dicionário. Na seção 2.2 explicamos a distinção entre Terminologia e Terminografia, fazemos um breve apanhado teórico sobre o termo e destacamos a Teoria Comunicativa da Terminologia (TCT) de Cabré (1999). Na seção 2.3 descrevemos as novas teorias da Terminologia. Ademais, delineamos a Semântica e a Terminologia de *Frames* de Fillmore (1975, 1976, 2003, 2006), de Faber Benítez et al. (2005) e de Souza (2019) como fundamentos para a estrutura do paradigma definicional baseado em *Frames*. Na seção 2.4, desenvolvemos a discussão da Terminografia, com um olhar mais detalhado sobre temas como a definição terminológica, o paradigma definicional, o percurso onomasiológico e o semasiológico, o termo, o qual denominamos signo terminológico (ST), e versamos sobre a árvore de domínio. Na seção 2.5, a última, descrevemos a metodologia da Linguística de *Corpus* e justificamos sua escolha para a condução desta pesquisa, bem como definimos o conceito de *corpus* e destacamos a taxonomia de Teixeira (2008).

No capítulo 3, fazemos um breve estado da arte de glossários e dicionários de Letras e Linguística, no suporte impresso, eletrônico e on-line. Para atingir este objetivo, fizemos uma análise de obras escritas ou traduzidas para o português disponíveis no Brasil. De igual modo, analisamos obras disponíveis na internet, de caráter bilíngue, englobando obras em inglês, francês, espanhol e japonês com base na proposta de Faulstich (2011). O resultado deste capítulo serviu de parâmetro para que pudéssemos autoavaliar o perfil do VoBLing e suas funcionalidades.

No capítulo 4, da Metodologia, subdividido em 5 seções, versamos sobre os passos desenvolvidos desde o levantamento do perfil do público alvo, até a proposta do paradigma definicional baseado em *Frames*. Na seção 4.1, descrevemos as etapas da pesquisa com o público-alvo, a condução da pesquisa *in loco* e seu resultado. Na seção 4.2, detalhamos a elaboração da árvore de domínio da Linguística, desde a primeira proposta de Fromm (2008) até a proposta atual desenvolvida por Fromm e Yamamoto. Na seção 4.3, abordamos a primeira fase da compilação e tratamento do *corpus*, processo que deve a duração de mais de 10 anos, desde a compilação inicial até o processamento final. Na seção 4.4, tratamos da extração de dados do *corpus* com o auxílio do console do WST (v. 7.0 e v. 8.0). Finalmente, na seção 4.5, descrevemos o paradigma definicional terminológico e o paradigma definicional enciclopédico baseados em *Frames*, contextualizados no VoBLing.

O capítulo 5, abordamos a compilação e tratamento do *corpus* em sua fase final. Na seção 5.1, discorremos sobre a digitalização, a correção e o carregamento do *corpus* na plataforma ToGatherUp (OLIVEIRA, 2019). Em seguida, na seção 5.2, tratamos dos aspectos

como a padronização, o dimensionamento e o redimensionamento do *corpus*. Por fim, nas seções 5.3 e 5.4, explicamos como foram conduzidos a extração e o tratamento dos dados no VoBLing.

No capítulo 6, descrevemos o processo de aplicação do paradigma definicional baseado em *Frames* aplicado ao VoBLing, detalhando o processo de extração dos traços distintivos do *corpus*, o lançamento nas fichas terminológicas do VoTec (FROMM, 2007) e, finalmente, sua disponibilização ao público-alvo como fase final.

No capítulo 7, dos resultados, expomos várias figuras com definições terminológicas e enciclopédicas dos termos definidos da Linguística. Em seguida, fizemos a comparação dessas definições com os paradigmas definicionais propostos e classificamos as definições em análogas ou em anômalas, seguindo o conceito de analogia e anomalia explanado por Robins (1997).

No último capítulo, das Considerações finais, resumimos de onde partimos, onde estamos e aonde queremos chegar, no que se trata da prática terminográfica, concebida a partir dos vieses da Terminografia pedagógica (FROMM, 2020) e da Acessibilidade textual e terminológica (FINATTO, 2019). Registramos o momento em que vivemos e como tem sido desenvolver uma pesquisa terminográfica a partir de um paradigma anterior para o paradigma do “novo normal” que se descortina diante de nós. No contexto do paradigma denominado o “novo normal”, uma plataforma terminológica bilíngue, com recursos multimodais, disponível para acesso gratuito na internet, faz todo o sentido.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, explicitaremos a escolha terminológica de **vocabulário**, em vez de **dicionário** de Linguística para esta pesquisa e trataremos do contraste Terminologia *versus* Terminografia. Apresentaremos também a fundamentação teórica desta pesquisa, envolvendo princípios teórico-metodológicos da Terminologia, em duas de suas subáreas, a Teoria Geral da Terminologia (TGT)¹⁰ e a Teoria Comunicativa da Terminologia (TCT) e da Terminografia. Nosso foco será permitir ao leitor uma visão dos primórdios da Terminologia e explicitar porque adotamos a TCT e seus respectivos princípios. Finalmente, abordamos a Linguística de *Corpus* (LC) baseados em Berber Sardinha (2000) e McEnery e Wilson (2001).

1.4. Vocabulário versus Dicionário

Primeiramente, o item a ser abordado nesta seção será a escolha teórica que embasa o uso do termo **vocabulário**, e não dicionário, para o produto que almejamos produzir. Ao tratar da Tipologia de obras lexicográficas e terminográficas, Barbosa (1990) propõe três termos distintos: **dicionário**, **glossário** e **vocabulário**.

O **dicionário** se aplica ao nível do sistema da língua, cujo objeto é todo o léxico disponível e todo o léxico virtual¹¹. A unidade mínima são os lexemas, com significado abrangente e de frequência regular no sistema. Teoricamente, apresenta todas as acepções de um mesmo verbete, partindo das variações diacrônica¹², diatópica¹³, diafásica¹⁴ e diastrática¹⁵ (FROMM; YAMAMOTO, 2013).

O **glossário** se aplica ao nível da fala, cujo foco é o conjunto de palavras manifestado em um determinado texto. Essas palavras são de aparição limitada na obra e têm um significado específico. Os verbetes de um glossário apresentam uma única acepção a partir de um contexto

¹⁰ Incluímos a TGT nesta fundamentação por ter sido a teoria fundadora da Terminologia.

¹¹ “Léxico virtual é aquele que, através dos processos de derivação, podemos mentalmente formular sua existência e seu significado possíveis. Léxico real é aquele em que, baseado no processo desenvolvido no léxico virtual, ocorre a devida instauração do uso das palavras como realização concreta daquele significado.” (RODRIGUES, 2006, p. 82)

¹² Variação linguística proveniente de períodos de tempo distintos.

¹³ Variação linguística que envolve a diversidade regional ou geográfica do léxico. Ex.: regionalismos e falares locais.

¹⁴ Variação linguística da fala que permite a alternância entre os registros formal e informal da língua, dependendo da situação de comunicação (também chamada de situacional).

¹⁵ Variação linguística que ocorre baseada em aspectos sociais como idade, grupo profissional, identidade étnica e grau de escolarização dos falantes.

determinado. As perspectivas de abordagem deles são: sincrônica ¹⁶, sintópica ¹⁷, sinstrática ¹⁸ e sinfásica ¹⁹ (FROMM; YAMAMOTO, 2013).

O **vocabulário** se aplica ao nível da norma e trabalha com conjuntos manifestados nas áreas de especialidade ou grupos específicos. A unidade-alvo é o vocábulo ou o termo, cujo significado é restrito e de alta frequência, diferentemente das palavras do glossário. No que tange à conceituação, apresenta todas as acepções de um verbete dentro de uma área de especialidade ou grupo. As perspectivas do vocabulário são a sincrônica e a sinfásica.

Além do **vocábulo**, concebido como um conjunto, inserido no léxico de uma língua, também existe o **Conjunto Vocabulário** ou **Vocabulários das Ciências**, ou **vocabulário de especialidades**, conjunto terminológico ou terminologia. Esse conjunto reúne o grupo de termos que representa o universo conceitual de uma área de conhecimento específica. É dentro dessa perspectiva que optamos por denominar o produto desta pesquisa Vocabulário da Linguística e não Dicionário de Linguística (YAMAMOTO, 2015, p. 11, 12).

1.5. Terminologia e Terminografia: princípios teórico-metodológicos

É importante que delimitemos a área de atuação na qual esta pesquisa linguística se insere: a Terminologia. Para tal, apresentaremos a seguir os conceitos de terminologia, de termo e de Terminologia. A Terminologia ou terminologia (grafadas com letra maiúscula ou minúscula) pode ser entendida de duas formas na Linguística: como (i) o vocabulário de especialidade de uma área científica (terminologia) ou como (ii) a subárea da Linguística (Terminologia). Logo, **terminologia** é um signo linguístico polissêmico, não só *per se*, mas também como ele é concebido pela Ciência, pelas correntes teóricas e por especialistas das Ciências do Léxico.

Em primeiro lugar, antes de definirmos termo, unidade lexical especializada e/ou unidade terminológica e associá-los aos seus autores, apresentaremos um breve histórico da gênese da terminologia na perspectiva de Alain Rey, terminólogo francês que propôs o termo terminografia (REY, 1995, p. 23), e de Cabré (1998). O autor afirma que a terminologia se ocupa dos nomes e do processo de **nominalização** nos quais a linguagem e o sentido são imprescindíveis. Diante disso, o autor defende o início da terminologia ocidental com o texto Crátilo (séc. V a.C.) de Platão (REY, 1995, p. 11)

¹⁶ Características linguísticas do mesmo período de tempo.

¹⁷ Características linguísticas do mesmo local.

¹⁸ Características linguísticas de um mesmo grupo sociocultural.

¹⁹ Características linguísticas da mesma situação de comunicação.

Em seguida, Rey (1995) trata da **nomenclatura** presente em obras como glossários, listas de nomes e dicionários em meados de 1500. Em 1735, a nomenclatura de Linneau²⁰ sobre a taxonomia das plantas mostra a **necessidade** científica de uma linguagem separada para a descrição botânica das espécies. Posteriormente, Locke publica a obra *Essay on Human Understanding* em 1690, na qual a relação entre o pensamento, o conhecimento e a linguagem se tornam bastante evidentes. Esta necessidade de expressão terminológica observada pelos naturalistas se estende para a lógica de Leipzig e é contemplada na enciclopédia de Diderot marcada pela relação entre os **conceitos** e as **palavras**.

De acordo com o autor, na segunda metade do séc. XVIII, o termo **terminologia** ocorre pela primeira vez nos escritos do professor alemão Christian Gottfried Schültz (1747-1832) e o adjetivo *terminologisch* é datado em 1788. Em francês, a palavra **terminologia** aparece em 1801 no sentido de jargão (termos incompreensíveis) com conotação negativa. Contudo, esse significado mudaria com a definição de William Whewell, cientista de Cambridge em 1837, na Inglaterra. Segundo este naturalista, a **terminologia** se desenvolve e passa a representar o *sistema de termos utilizados na descrição de objetos da história natural*, contudo limitado às Ciências Naturais. Consequentemente, a mudança do que antes era **nomenclatura** para **terminologia** implicou na variação do foco de **nome** para **termo** (REY, 1995, p. 11-18).

Depois de Francis Bacon, filósofo da ciência e da linguagem considerado o precursor da terminologia moderna, segundo Rey (1995, p. 17), houve muitas outras contribuições. Por exemplo, a criação da Enciclopédia de Diderot e d'Alembert, um passo essencial nesta evolução e que corresponde ao princípio de que uma ciência é uma linguagem bem elaborada. Ademais, o trabalho taxonômico de Linnaeus²¹ levará Michel Foucault a discutir a unidade entre o conhecimento e a língua na obra *As palavras e as coisas* de 1966. O autor presencia o crepúsculo do século XVIII e o amanhecer do século XIX como um período chave na história das ideias, de profunda mudança do conhecimento.

É nesse contexto que nasce a nova ciência da Linguística, da observação das estruturas das línguas e se preocupa com o conceito de denominação que só pode ocorrer em uma língua específica com uma estrutura funcional e formal. Dessa maneira, é descoberta a necessidade de uma ciência das formas e Goethe cria a palavra **morfologia**. O romantismo faz a identificação do uso da língua com a cultura e a Filologia conecta os textos com as questões sociais e históricas.

²⁰ Karl Linnaeus: biólogo sueco, autor do trabalho sobre a classificação das plantas em 1735.

Floresce neste período a Linguística moderna com a contribuição de Rask, filólogo dinamarquês, seguido pelos alemães Grimm, Franz Bopp, Schlegel e Wilhelm von Humboldt. Existe uma predominância das línguas dos grandes impérios, como o português, o espanhol e, depois, o inglês, o alemão e o francês, usadas em contextos administrativos em suas colônias.

O colonialismo abala o equilíbrio comercial, cultural e linguístico pré-existent, mas abre a Ásia e a África para línguas europeias. É desse contexto que surge a necessidade de uma terminologia que atendesse a demanda de expressão do conhecimento especializado. Por exemplo, há o desenvolvimento da metalurgia, da indústria têxtil, da química, da ótica e da fotografia; e essa efervescência tecnológica cria o discurso técnico-científico usado pelas comunidades políticas. Há também a transição de uma sociedade rural para uma sociedade industrial, onde comportamentos culturais se modificam, coexistem ou são substituídos por outros. Novos padrões econômicos passam a existir com a concentração da população nas grandes cidades, com a democratização da educação, das relações interpessoais e dos regimes políticos (REY, 1995, p. 20; CABRÉ, 1998, p. 1- 2).

No século XX, progressos na educação e nas ciências mostram a necessidade de uma terminologia mais especializada para a transmissão do conhecimento na civilização pós-industrial, principalmente na língua escrita. O deslocamento da terminologia das ciências formais e naturais para a linguagem cotidiana, para as atividades profissionais, passa a ter uma visão científica predominantemente positivista (REY, 1995, p. 21, 22).

Cabré (1998, p. 1, 2) afirma que o desenvolvimento da ciência é espontâneo, assim como a criação da tecnologia. A motivação para a existência da terminologia é teórica, pois essas mudanças na sociedade tiveram grandes efeitos na linguagem, criando produtos linguísticos, novos profissionais da terminologia e novas formas de organizar a comunicação. Concluindo, Rey (1995, p. 59) assegura que não é possível nomear sem nomes, assim como é impossível nomear cientificamente sem termos, porém tanto os nomes quanto os termos pertencem a uma língua.

Na próxima seção exporemos a concepção do termo, objeto da terminologia, a partir da perspectiva de teóricos que fundamentam a Terminologia dentro e fora do Brasil.

1.1.3.Termo: concepções

A partir da explanação deste histórico, é possível entender o segundo conceito aplicado ao termo terminologia: o vocabulário especializado utilizado por especialistas e profissionais de diferentes áreas de especialidade. A denominação de conceitos tem como produto os termos

(BARBOSA, 1990; REY, 1995; CABRÉ, 1999; MACIEL, 2001; WÜSTER, 2003), também denominados *unidade lexical terminológica*, *unidade lexical especializada* (KRIEGER, 2001), *signos terminológicos* (BARROS, 2006) e *unidades terminológicas* (UT) (BARBOSA, 1990; REY, 1995; WÜSTER, 1998, p. 21 apud FINATTO, 2014).

Em primeiro lugar, é relevante discutir sobre terminologia, grafada em letra minúscula, aquela presente nos meios de comunicação, em encontros científicos, em contextos comunicativos entre profissionais de diversas áreas do conhecimento, de especialidade, e nos discursos profissionais. A terminologia objetiva a clareza e praticidade na comunicação, por meio de **termos**, unidades que abrigam conceitos pertencentes a um campo especializado (ISO 1087-1, 2000²²; ANDRADE, 2001; PAVEL; NOLET, 2002; KRIEGER; FINATTO, 2004; TALAVÁN, 2016).

Barbosa (1990) conceitua **termos** como os elementos constituintes da metalinguagem científica ou tecnológica, palavras que designam noções úteis a uma ciência, à terminologia ou ao conjunto terminológico, nos quais as relações de significações se fazem a partir da junção de expressão e conteúdo²³ (BARBOSA, 1990, p. 155, 156). Nesse contexto, o tratamento que é dado às unidades é o que distingue o termo da unidade lexical.

Rey (1995, p. 147) afirma que unidades lexicais podem ser termos ou unidades terminológicas, bem como fraseologias. Da mesma forma, Cabré et al. (2007, p. 1-3) trazem uma discussão metaterminológica quanto ao termo e afirmam que a noção de **termo** varia entre a unidade única, as unidades textuais com propriedades morfológicas e sintáticas diversas ou as unidades de compreensão. Estas autoras também discutem o objeto da terminologia e afirmam, com base em estudos empíricos, que a maior parte das terminologias é constituída de termos compostos ou multipalavras. Este fato é ratificado por Flores et al. (2009, p. 28), “Afinal, um termo nem sempre é uma só palavra. Mostra-se, também, como uma cadeia de palavras. E, ao apresentar-se como a cadeia, revela-nos uma unidade de sentido cujos limites nem sempre são fáceis de precisar...”

Maciel (2001, p. 40,41) afirma que **termos** são palavras técnicas inseridas em um texto técnico-científico, reveladores da especialização e da linguagem de especialidade de uma ciência. Logo, o termo constitui-se uma unidade da linguagem de especialidade, como parte da

²² ISO - International Organization for Standardization - do grego igual, representa o objetivo da organização que visa aprovar padrões internacionais em todos os campos técnicos, como classificações nacionais/regionais, padrões técnicos, procedimentos e padrões de processos etc. No Brasil, a ABNT (Associação Brasileira de Normas Técnicas) é a representante da ISO. Disponível em <https://www.iso.org/about-us.html#5;https://www.normastecnicas.com/iso/o-que-e-iso/>).

²³ A expressão são os sons (cadeia sonora) que transmitem a significação e o conteúdo é forma como a língua apresenta a significação (DUCROT; TODOROV, 1972, p. 36).

competência específica do falante. Termos são unidades que denominam conceitos, uma maneira de conhecer e denominar o conhecimento de especialidade.

Para Wüster (2003, p. 21), o **termo** pode ser uma palavra ou um grupo de palavras compostas; ele possui uma forma externa (configuração, forma física) e uma forma interna (conceito), ponto de partida para os terminólogos. Dessa forma, para o autor, o que importa à Terminologia é o léxico das línguas, ou seja, a denominação dos conceitos, em uma perspectiva sincrônica, prevalecendo a forma escrita sobre a fônica (WÜSTER, 2003, p. 25). Assim como Rey (1995), Wüster defende a ideia de que os termos são embasados na ontologia, na lógica e na ciência da informação. Em suma, Wüster conceitua a UT como palavra a qual se atribui um conceito e um significado, em contraste com a definição linguística para palavra, unidade lexical composta por forma e conteúdo (WÜSTER, 1998, p. 21 apud FINATTO, 2014, p. 76). Neste caso, segundo o autor, o conceito ou o conteúdo especializado é o que distingue o termo de uma unidade lexical.

Barros (2006, p. 22) define o signo terminológico a partir de uma perspectiva saussuriana na qual o conceito e o significado, “unidade de pensamento maior que congrega também elementos pragmáticos”, são partes indissociáveis do termo. Esse ponto de vista difere do conceito de termo em Wüster no qual o conceito (significado) e a expressão (significante) são unidades independentes. Outra distinção é que em Wüster o conceito é universal e imutável. Não obstante, como signo terminológico, é necessário considerar a sua interação com os outros elementos textuais constituintes da linguagem de especialidade.

A forma como o **termo** será abordado em relação aos objetivos da pesquisa é o que determina sua denominação desde que seu estatuto seja bem delimitado (CABRÉ, 2003). De acordo com a autora, na Terminologia, a unidade terminológica (UT) é uma unidade lexical com uma estrutura morfossintática correspondente a uma unidade conceitual autônoma de uma área de especialidade, o que não exclui as multipalavras. Rey (1995) define a UT como o conjunto de itens que representa um único conceito, podendo ser marcada por arbitrariedade, no caso de abreviaturas. Por exemplo em BCG (Bacilo Calmette-Guérin), o nome da vacina. Segundo o autor, nas línguas ocidentais, as unidades terminológicas são, em geral, frases substantivas constituídas por um substantivo e o(s) determinante(s). Nesta pesquisa adotamos o conceito de termo segundo Cabré (1999) explicitados na seção da 2.3.3, da TCT.

Em segundo lugar, **Terminologia**²⁴ (grafada com letra maiúscula) designa o estudo técnico-científico dos termos ou disciplinas das áreas de especialidade, os conceitos que os

²⁴ Nesse caso, e para muitos pesquisadores, a Terminologia também engloba a Terminografia, que é a parte prática. Na nossa pesquisa, porém, separamos os termos Terminologia e Terminografia.

subjazem, a denominação, sua utilização e os princípios teóricos que constituem esta área de estudo. Esta área de estudo, que se consolida como subárea da Linguística Descritiva (LD), da Linguística Aplicada (LA) e das Ciências da Linguagem, é onde nossa pesquisa se insere (ISO 1087-1, 2000; KRIEGER, 2001; MACIEL, 2001; PAVEL; NOLET, 2002; ALMEIDA, 2004; KRIEGER; FINATTO, 2004; TALAVÁN, 2016).

Conforme Finatto (2001b), a Terminologia é uma “área de estudos autônomos, de caráter multifacetado, integrada por fundamentos provenientes das Ciências da Linguagem, Ciências da Cognição e Ciências Sociais”. Além dessa definição, a autora cita as diferentes abordagens que a Terminologia pode ter, quais sejam: a linguística, a cognitiva e a sociológica. No aspecto linguístico, haveria a inclusão da “dimensão textual, semiótica (uma extensão da semiologia) e o discursivo” (FINATTO, 2001b, p. 141); a abordagem cognitiva refere-se à terminologia como meio de transmissão e comunicação do conhecimento técnico-científico; e a abordagem sociológica trata dos usuários, do contexto em que estes a usam, a temática que os termos veiculam; e em qual tipo de discurso é mais recorrente (CABRÉ, 1999). Não muito comumente, a Terminologia também pode ser concebida como uma “especificidade” da Lexicologia (ANDRADE, 2001; KRIEGER, 2001).

Como disciplina, a Terminologia se beneficia de outras ciências, como a Linguística, a Documentação, a Lógica, a Ontologia, a Ciência da Informação, as Ciências Exatas e Biológicas, a Teoria da Informação e da Comunicação (SAGER, 1993; WÜSTER, 1974 apud MACIEL, 2001; TALAVÁN, 2016). Esse aspecto faz com que a Terminologia se aplique e se beneficie de outras áreas da Ciência e tenha o caráter inter e transdisciplinar (CABRÉ, 1999). Os objetos de estudo da Terminologia são o termo técnico-científico, as fraseologias especializadas e a definição terminológica, podendo ter dois enfoques: (i) o desenvolvimento teórico e as análises descritivas (Terminologia), bem como (ii) as aplicações terminológicas (Terminografia). Suas bases foram estabelecidas pelo austríaco Eugen Wüster (1898-1977), considerado o fundador da terminologia moderna, na Universidade de Viena, em 1972, objetivando a univocidade comunicacional em nível internacional e o lançamento da TGT, de caráter prescritivo (BARBOSA, 1990; CABRÉ, 1999; ANDRADE, 2001; KRIEGER; FINATTO, 2004).

Na distinção Terminologia *versus* Terminografia, há teóricos que afirmam que a Terminologia abrange aspectos teóricos do estudo dos termos, enquanto a Terminografia aborda os aspectos práticos da produção de obras terminográficas, como vocabulários, (também denominados glossários e dicionários de especialidade) e listas de termos (ISO 1087-1, 3.6.2; ANDRADE, 2001; BARROS, 2004; FINATTO, 2004). Trata-se de uma visão parcial, já que a

Terminografia também responde pelo rigor científico e pelos princípios teórico-metodológicos que moldam a prática adequada da produção de obras terminográficas (BARBOSA, 1990; 1991; BESSÉ; NKWENTI-AZEH; SAGER, 1998). No subitem 2.4, faremos uma abordagem mais detalhada da Terminografia. No próximo item, faremos um breve apanhado da TGT como teoria fundadora da Terminologia como disciplina científica.

1.1.4. Teoria Geral da Terminologia – TGT

Eugen Wüster (1898-1977) foi um professor austríaco responsável pelos fundamentos da TGT em 1972. A TGT se ocupa dos termos e conceitos, a partir de uma perspectiva cognitiva, e busca a normalização/padronização das noções e dos termos (denominações) técnico-científicos. Esta padronização objetiva a univocidade da comunicação especializada no plano internacional (FINATTO; ZILIO, 2015).

Seu surgimento se deu devido à necessidade de técnicos e cientistas normalizarem denominativa e conceitualmente suas disciplinas, objetivando a comunicação profissional e a transferência do saber científico e tecnológico. Wüster defendeu sua tese intitulada *International Sprachnormung in der Technik* (1931), partindo da lógica conceitual clássica para criar os fundamentos conceituais, a metodologia e sua própria teoria com vistas ao estabelecimento da Terminologia (KRIEGER, 2001; WÜSTER, 2003).

Como parte da Escola de Viena, a TGT privilegia a **perspectiva cognitiva** da Terminologia; assim, os termos são concebidos como unidades do conhecimento, o que é distinto da concepção do signo linguístico (signo, conteúdo e expressão). Buscando sistematizar os estudos da Terminologia, que se nutre da Lógica, da Ontologia e da Ciência da Informação, o autor propôs que os conceitos e as denominações dos termos fossem independentes, uma forma de racionalização dos termos. O conceito, constituinte do termo, é estável, paradigmático, universal, revelando uma concepção positivista da Ciência. Esse conceito é dissociado da expressão “unidade lexical/significante”, o que difere da concepção saussuriana do signo linguístico. O percurso adotado por esta corrente é o **onomasiológico**, ou seja, parte-se do conceito para se chegar ao termo (KRIEGER, 2001; WÜSTER, 2003).

Quanto ao fazer terminográfico, Wüster propõe a designação dos conceitos por meio de símbolos ou termos, interligados por relações lógicas (gênero-espécie) e ontológicas (todo-parte). Avaliadas como insuficientes, mais adiante, Arnzt e Picht (apud CABRÉ et al., 1996), propõem duas novas categorias: hierárquicas e não hierárquicas. As hierárquicas abarcam as relações lógicas e ontológicas e as não hierárquicas envolvem relações conceituais mais casuais

(BARROS, 2004, p. 112-115). Dentre as relações não hierárquicas, há as relações sequencial e pragmática.

A **univocidade** ou monossema é outra característica fundadora da TGT, o que automaticamente exclui a possibilidade de ocorrência de sinonímia, homonímia e polissemia. Em Wüster, a variação linguística monolíngue ou interlíngua é concebida como indesejável, o que exclui a análise do termo em seu contexto de ocorrência. Contudo, Faulstich (2006, p. 27) critica essa posição, afirmando ser necessária a padronização de termos em nível internacional e a inclusão de variações locais e nacionais, considerados os critérios sociais e político-linguísticos.

Cabré (1999) pontua algumas considerações sobre a limitação da TGT, dentre as quais se destaca o seu **caráter reducionista**. Esse reducionismo se manifesta nos seguintes aspectos: (1) os conceitos são estáticos devido à perspectiva sincrônica adotada por esta teoria; (2) os termos não têm valores pragmáticos, nem variação semântica, devido ao registro formal profissional; (3) o aspecto prescritivo da Terminologia e sua concepção estão fora do léxico de um sistema linguístico; (4) nega-se que o termo faria parte da variação discursiva da comunidade linguística; (5) padroniza-se a terminologia da engenharia industrial com vistas a eliminar a ambiguidade. A autora conclui que a TGT não pode ser invalidada como teoria, mas a limitação conceitual e funcional e a falta de generalização devem ser consideradas.

Segundo Krieger (2001), a limitação que se pôs à TGT, devido à concepção do termo, era o fato de ser produzida por especialistas para a comunidade de especialistas. Logo, seriam artificiais, à parte dos sistemas linguísticos e de seu funcionamento natural. Consequentemente, a concepção de língua, para Wüster (2003), era de uma instituição idealizada na qual os termos serviriam para evitar a ambiguidade na comunicação especializada. Essa perspectiva concebe o termo dissociado do sistema linguístico e de seu aspecto pragmático de comunicação.

As incompletudes dos pressupostos teóricos da TGT impulsionaram o surgimento de novas teorias, movimento natural da construção do conhecimento, já que a TGT foi a teoria fundadora da Terminologia moderna. As novas teorias vêm propor novos conceitos dentro das relações paradigmáticas das línguas em contextos sociais mais amplos. Dentre as novas teorias, mencionamos a Teoria Socioterminológica da Terminologia fundada por Gaudin em 1993 (GAUDIN, 2005) e a Teoria Comunicativa da Terminologia com Cabré (1999), que surgiram à medida que os estudos terminológicos observavam a necessidade de novos conceitos e perspectivas para abarcar os conjuntos terminológicos. Dessa forma, o termo que era então visto do aspecto conceitual, a partir de uma perspectiva lógica, mais restrita a especialistas, passa a

ser tratado também como uma unidade linguística, pertencente aos níveis científico, técnico e popular.

Adiante fazemos um breve apanhado da Terminologia como subárea da Linguística Descritiva.

1.6. Novas Teorias da Terminologia

A Terminologia, como disciplina científica, também possui várias vertentes, surgidas desde seu estabelecimento com Wüster, em 1972, na Universidade de Viena (KRIEGER; FINATTO, 2004, p. 20). Na seção 2.3.1 abordaremos aspectos relevantes da Teoria Comunicativa da Terminologia (TCT) de Cabré (1999), selecionada para embasar este trabalho.

Apesar desse enfoque, cabe apontar que outras correntes, como a Teoria Socioterminológica da Terminologia²⁵ (TST) e a Teoria Sociocognitiva da Terminologia (TSCT)²⁶, também fazem parte da Terminologia.

1.1.5. Teoria Comunicativa da Terminologia – TCT

Cabré (1999) lança, nos anos 1990, as bases da TCT, concebendo a unidade terminológica como parte da linguagem natural, constituinte de um sistema fonológico, morfológico, semântico, discursivo e pragmático das Línguas Naturais (LN).

²⁵ Teoria Socioterminológica da Terminologia (TST): lançada pelo francês François Gaudin em 1993, concebe a unidade terminológica do ponto de vista descritivo, como unidade das línguas naturais, cujo significado é negociado pela comunidade de falantes (GAUDIN, 2005, p. 88). A TST resgata a dimensão social da Terminologia e a variação terminológica na linguagem da especialidade. Gaudin propõe o termo *normaison*, processo que visa à construção consciente de uma norma unificada, ou seja, um processo de autorregulação dos sistemas terminológicos realizado pelos próprios usuários (CABRÉ, 1991, 1999). Devido ao uso social, a Socioterminologia reconhece a presença da sinonímia, da homonímia e da polissemia, ou seja, a variação linguística nos discursos especializados (BARROS, 2004, p. 69). Em resumo, Barité (2015, p. 145) define a TST como “ramo da terminologia que trata da análise dos termos (surgimento, formação, consolidação e inter-relações), considerando-os a partir de uma perspectiva linguística na interação social” (BARITÉ, 2015, p. 145).

²⁶ Teoria Sociocognitiva da Terminologia: fundada por Rita Temmerman, baseada na Semântica Cognitiva, esta teoria define o termo como a designação de uma unidade complexa de entendimento (*unit of understanding*), concebida e percebida na mente, de estrutura prototípica, sempre em evolução (características ontológicas, lógicas, semânticas e pragmáticas). Esses termos conectam unidades prévias de compreensão às novas unidades, e os módulos de compreensão são eventos estruturados que abarcam informações históricas, intra- e intercategoriais e processuais. A sinonímia, a monosssemia e a polissemia podem ser funcionais, e as expressões metafóricas fazem parte da descrição terminológica. Nesta teoria, o conhecimento linguístico é ligado ao conhecimento de mundo, pois a unidade de entendimento é uma unidade interpretável, dependente da compreensão humana, e as definições dessas unidades de entendimento são condicionadas à especificidade do conhecimento dos especialistas. Como a referida autora questiona a rigidez do formato de definição aristotélica, Gênero Próximo e Diferença Específica (GPDE), pode ser que sua teoria contribua para uma definição mais concisa de termos abstratos, difíceis de serem definidos no padrão GPDE, cuja definição é denominada de enciclopédica pela autora (TEMMERMAN, 2000; MACIEL, 2010).

Os aspectos comunicativos e descritivos das linguagens especializadas são mais valorizados pela TCT que os normativos da TGT, e a unidade terminológica passa a existir dependente do contexto no qual se insere, o da comunicação especializada (BEVILACQUA, 2001; KRIEGER; FINATTO, 2004). É dentro dessa perspectiva teórica que esta pesquisa se insere.

A TCT embasa-se em três teorias: a teoria da língua (1), a do conhecimento (2) e a da comunicação (3). A (1) teoria da língua considera o sistema de uma LN, envolvendo a gramática, a semântica e a pragmática; a (2) teoria do conhecimento sustenta a comunicação do conhecimento especializado por meio das unidades que podem ser termos ou não; a (3) teoria da comunicação leva em conta o contexto comunicativo, no qual concebemos a ideia do emissor, receptor, mensagem, função dos textos especializados etc. (BEVILACQUA, 2001). Na análise do contexto comunicativo, Cabré (1999) salienta a distinção da terminologia em dois pontos de vista: o do produtor e o do consumidor. Nesta pesquisa, nosso foco reside sobre o segundo ponto de vista, já que objetivamos produzir um protótipo de vocabulário, considerando o acadêmico iniciante do Curso de Letras como usuário.

Segundo Dornelles (2015, p. 13), unidade terminológica é “uma unidade lexical da língua natural que adquire valor especializado dentro de um contexto especializado, segundo critérios semânticos, discursivos e pragmáticos”. Em outras palavras, a partir da perspectiva da TCT, a unidade pode ser ora lexical, ora terminológica. O que distingue um *status* do outro é o contexto no qual a unidade se insere, o que Cabré (2011) denomina de “caráter poliédrico do termo” (CABRÉ, 2011, p. 149; YAMAMOTO, 2018, p. 275). O *status* poliédrico significa que o termo pode ser uma unidade conceitual, de conhecimento ou linguística, a partir do contexto de análise.

Essa plasticidade do *status* do termo se dá uma vez que o termo pode ser, ao mesmo tempo, uma unidade linguística, cognitiva e sociocomunicativa. Pode ser (1) 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. Pode ser (2) cognitiva por promover a distinção entre o conhecimento especializado e o conhecimento geral ou não especializado. Ademais, pode ser (3) sociocomunicativa, por considerar os aspectos pragmáticos que identificam a conformação e o uso das unidades em determinadas situações comunicativas.

Neste trabalho, adotaremos a expressão signo terminológico (ANDRADE, 2001), doravante ST, para designar o que foi previamente explicado como uma unidade terminológica. A denominação signo terminológico (ST) se deve à postura que tomamos quanto ao conceito

de termo como proposto por Saussure, no qual o signo linguístico é marcado pela arbitrariedade. Também consideramos que o conceito é indissociável de sua denominação (expressão).

O conceito é único para uma unidade terminológica, da mesma forma como ocorre com o signo terminológico. De igual modo, conforme defendido por Wüster (2003) e Rey (1995), a junção de duas ou mais unidades terminológicas ou termo composto será representado por um único conceito. Para exemplificar a ideia de unidade terminológica nesta pesquisa, temos as subáreas da Linguística, compostas por duas unidades ou mais, como a Linguística Cognitiva, a Linguística Diacrônica, a Linguística de *Corpus*, entre outras. Apesar de serem compostas por unidades distintas, elas são interdependentes e representam um único conceito dentro da rede de conceitos das subáreas da Linguística, como termos subordinados.

1.1.6. Semântica e Terminologia de *Frames*

Nesta seção, explicaremos as teorias adotadas para estruturar o paradigma definicional (PD), provenientes da Semântica de *Frames* e da Terminologia de *Frames*, baseados em Fillmore (1975, 1976, 2003, 2006), Souza (2019) e em Faber Benítez et al. (2005)

Como parte de nosso trabalho é construir os PD dos verbetes do VoBLing, lançamos mão da Semântica de *Frames*, proposta por Fillmore (2006) na busca de teorizar/justificar como os traços distintivos são organizados no PD dos verbetes. Além de Fillmore, baseamo-nos na pesquisa de Souza (2019), que desenvolveu uma proposta teórico-metodológica unindo a onomasiologia, a Lexicografia e a Semântica Lexical Cognitiva, na construção de um mapa conceitual da culinária da imigração. Por fim, trazemos a contribuição da Terminologia de *Frames*, baseados em Faber Benítez et al. (2005), justificando o uso de *corpus* e suas vantagens nas análises dos contextos para extração de dados para definições.

1.1.1.1 Semântica de *Frames*

A Semântica de *Frames*, segundo Fillmore (2006, p. 613), é “uma abordagem para descrever os significados de entidades linguísticas independentes (palavras, frases lexicalizadas e várias construções gramaticais especiais), recorrendo aos tipos de estruturas conceituais subjacentes a seus significados e que motivam seu uso.”

De acordo com Fillmore (1975), os conceitos dos termos **cena** e **frame**²⁷ são definidos da seguinte forma:

Eu gostaria de dizer que as pessoas associam certas **cen**as com certos quadros linguísticos. Uso a palavra **cena** em um sentido amplo, incluindo não apenas cenas visuais, mas também transações interpessoais comuns, cenários padrões definidos pela cultura, estruturas institucionais, experiências representadas, imagem corporal e, em geral, qualquer tipo de fragmento coerente das crenças humanas, ações, experiências ou pensamentos. Uso a palavra **frame** para qualquer sistema de escolhas linguísticas - os casos mais fáceis são conjuntos de palavras, mas também incluem escolhas de regras gramaticais ou categorias linguísticas - que podem ser associadas às instâncias prototípicas de **cen**as. (FILLMORE, 1975, p. 124 – tradução do autor).^{28 29}

Segundo o autor (1975, p. 124), as cenas e os *Frames* são interconectados e ativados na memória, por meio do material linguístico, considerados aspectos de similaridade de suas relações ou de seus contextos de ocorrência. Posteriormente, Fillmore (2003, p. 263) defendeu que os *Frames* devem ser tratados como componente essencial da definição de palavras: “...certos tipos de informações estruturadas de plano de fundo (ou “frames”) devem ser tratadas como componentes essenciais ou acompanhamentos das definições das palavras.”³⁰

Souza (2019) sintetiza esses conceitos, concluindo que, “*Frames* são estruturas onomasiológicas, esquemas abstratos que se manifestam a partir de formas linguísticas” (SOUZA, 2019, p. 18). Para exemplificar o conceito de *Frames*, Fillmore (2003, p. 267) trata da quarta-feira como o quarto dia da semana. Primeiramente, é necessário conceber que a quarta-feira está inserida num quadro semanal de sete dias e que o *frame* é a descrição desse período. Nesse *frame* também identificamos a ideia de calendário e de uma métrica na qual dias e semanas são usados para medir o tempo. Geográfica e culturalmente falando, entendemos que, nas sociedades ocidentais, há um período de cinco dias de trabalho e dois de descanso, e que este ciclo se repete ininterruptamente. Consequentemente, nesse quadro coexistem os

²⁷ O autor afirma que o termo *frame* pode ser encontrado em outras áreas do conhecimento, contudo, a ideia de coerência organizacional é recorrente em todas elas. Ademais, o autor propõe **módulo** como termo alternativo, e evoca a ideia de móveis modulares, já que o produto da montagem é um conjunto de partes menores constituintes do mesmo móvel. Para concluir, o autor esclarece que prefere *frame*, pois evoca a ideia de servir a algo. (FILLMORE, 1975, p. 130).

²⁸ I will content myself here with showing some of the ways in which I would like to use these terms. I would like to say that people associate certain scenes with certain linguistic frames. I use the word scene in a maximally general sense, including not only visual scenes but also familiar kinds of interpersonal transactions, standard scenarios defined by the culture, institutional structures, enactive experiences, body image, and, in general, any kind of coherent segment of human beliefs, actions, experiences or imaginings. I use the word frame for any system of linguistic choices—the easiest cases being collections of words, but also including choices of grammatical rules or linguistic categories—that can get associated with prototypical instances of scenes. (FILLMORE, 1975, p. 124).

²⁹ Informamos aos leitores que todas as traduções contidas nesta tese são do autor.

³⁰ I maintain that certain kinds of structured background information (or “frames”) should be treated as essential components or accompaniments of word definitions. (FILLMORE, 2003, p. 263).

conceitos dos dias da semana, sua ordem sequencial e cíclica, separando o fim de semana dos dias úteis.

Compreender esses conceitos serve ao propósito de aplicarmos a teoria da Semântica de *Frames* à metodologia do VoTec, haja vista que os traços distintivos mais frequentes, escolhidos pelos especialistas, são utilizados para a construção do PD na plataforma.

Na busca de conciliar a Terminologia e a Teoria de *Frames*, temos o processo mental de conceptualização, próprio do signo linguístico e do termo. De acordo com Fillmore (1976), o conjunto de *frames* contido na memória humana serve à estruturação, classificação e interpretação das experiências vividas. Esses esquemas de *frames* podem ser incorporados fisiologicamente (linguagem corporal, interpretação de cores), pela relação de causa e efeito ou pela interpretação de símbolos. Este processo de interpretação das experiências vividas da teoria de *Frames* condiz parcialmente com o que será descrito na seção 2.4.3, os percursos onomasiológico e semasiológico, no quais o conjunto de conceitos (conceptualização) leva ao signo linguístico (lexemização) e vice-versa.

Esta pesquisa segue o perfil semasiológico com a produção de lista de palavras, lista de palavras-chave e na seleção dos traços distintivos (constituintes dos *frames*) selecionados e organizados para a construção do PD. Para entender o processo de organização dos traços distintivos, Souza (2019) explica que é necessário haver “...uma estrutura epistemológica sólida e robusta, que explique o que faz com que unidades específicas do léxico se enquadrem dentro de conceitos, ao mesmo tempo em que oferece subsídios para organizar internamente essa multiplicidade de unidades” (SOUZA, 2019, p. 38).

Organizados os traços distintivos, há a **definição** (ou PD, nesta pesquisa) da palavra ou do ST, o que é de “grande importância para a interpretação de qualquer um dos contextos que uma dada lexia possa apresentar” (SOUZA, 2019, p. 38). O autor baseia-se em Pierre Guiraud (1972; apud SOUZA, 2019, p. 39) e explica que “toda palavra tem uma espécie de ‘nó semântico’ que é mais ou menos denso, [...] cercado por uma auréola de associações secundárias, afetivas ou sociais.” Por conseguinte, interpretamos que a definição se trata da passagem dos dados periféricos (associações secundárias, afetivas ou sociais) para se chegar ao dado central (nó semântico). Este princípio é aplicado neste trabalho, no qual o dado central comporá a definição terminológica, construída com os traços distintivos mais recorrentes (nó semântico), e os dados periféricos (traços distintivos menos recorrentes) comporão a definição enciclopédica.

Fillmore (2003, p. 263) defende que “Devemos conceber a ideia de dicionários em CD ou, preferencialmente, uma versão virtual com *hyperlinks* que facilitem o acesso às informações

básicas necessárias.”³¹ Ou seja, sua proposta objetiva, especificamente, os vocabulários eletrônicos ou virtuais, devido à necessidade de disponibilização de grande quantidade de dados, tornando-a inapropriada para obras impressas. Os tipos de informações básicas a serem disponibilizadas são dicionarísticas e enciclopédicas, com a provisão de exemplos baseados em *corpus*, usos comuns e fraseologias.

“Metadescrições cuidadosas dos próprios textos de amostra, com um *corpus* suficientemente grande, poderiam possibilitar generalizações sobre relações temporais, regionais e sociais do vocabulário da língua, seja por meio de ferramentas computacionais que acompanham o dicionário ou por tipos tradicionais de *corpus* ou pesquisa filológica baseados nos exemplos anotados.” (FILLMORE, 2003, p. 265).³²

Na descrição do *corpus* proposta pelo autor (2003), constatamos que, a partir do *corpus* de pesquisa, podemos construir um PD com dados linguísticos e com informações enciclopédicas do *corpus* por meio da análise tradicional, à exceção de que nosso *corpus* não é anotado (metadescrito). Este dado também serve ao propósito da adequação dessa teoria à construção do VoBLing.

Na próxima subseção, apresentamos a transição da Semântica de *Frames* para a Terminologia de *Frames*, com base em Fillmore (2003) e Faber Benítez et al. (2005).

1.1.1.2 Terminologia de *Frames*

Nesta subseção, buscamos associar a Semântica de *Frames* à Terminologia de *Frames*, justificando seu uso na Terminologia baseada em *corpus*. Embasamos essas teorias em Fillmore (2003) com a análise etnográfica e o padrão de um dicionário baseado em *frames*; e em Faber Benítez et al. (2005) tratando da rede conceitual.

Fillmore (2003) considera a aplicação da Semântica de *Frames* na Terminologia, quando analisa os *frames* dos termos id, ego e superego da Psicanálise e parte da terminologia da religião. Esse autor propõe que os termos sejam analisados e conceituados na perspectiva de *frames* como uma análise etnográfica. Logo, é necessário explicitar as crenças básicas, experiências, práticas, instituições, ou conceptualizações disponíveis pelos falantes da língua como fundamentos do *frame* que subjaz a palavra. Consequentemente, o autor afirma que um

³¹“...we should be thinking in terms of a dictionary on compact disk or preferably a web utility with hyperlinks facilitating easy access to the necessary background information.” (FILLMORE, 2003, p. 263).

³²“Careful metadescriptions of the sample texts themselves, given a large enough corpus, could make it possible to derive generalizations about temporal, regional, and social associations of the vocabulary of the language, either through computational tools that come with the dictionary or through traditional kinds of corpus or philological research based on the body of annotated examples.” (FILLMORE, 2003, p. 265).

bom dicionário deve trazer informações lexicais e enciclopédicas do verbete definido, para que uma abordagem de *Frames* seja aplicada.

Continuando na descrição do perfil de uma obra ideal baseada em *Frames*, Fillmore (2003, p. 286) elicit seis características³³. Dessas, identificamos duas encontrados no VobLing, tendo em mente que nossa proposta é outra: (i) a definição para um verbete buscado é disponibilizada em outra janela, o que no VobLing é disponibilizado por meio de remissivas e janelas *pop-ups*; e (ii) exemplo de uso do verbete baseado em *corpus*.

Especificando os tipos de dicionários, o autor (2003, p. 287) lista os “dicionários históricos, dicionários de alunos, dicionários infantis, **dicionários especializados** etc., diferindo entre si em relação às listas de palavras, estilos de definição, acesso a antecedentes culturais...” Neste caso, encontramos no VobLing um *corpus* especializado, de tamanho considerável, representativo da comunidade acadêmica internacional da área de Linguística, em LP e LI. Essa variedade de dados permite que vários *frames* sejam acessados e vários conceitos sejam extraídos, a partir de perspectivas diferentes das várias subáreas da Linguística.

Faber Benítez et al. (2005) explicam que a conexão entre a Semântica de *Frames* e a Terminologia encontra-se na rede conceitual que evoca o campo semântico de uma palavra ou de um ST. A Semântica de *Frames* e a Terminologia de *Frames* se interseccionam e promovem a organização das redes conceituais secundárias, subjacentes ao conceito principal do *frame* de um ST. A Terminologia de *Frames* leva em consideração eventos das áreas de especialidade, onde encontramos os *frames* como uma estrutura útil para definir conceitos. Isto se justifica pois os *frames* situam os conceitos em ambientes nos quais as categorias estão interrelacionadas, e que um conceito evoca o sistema conceitual como um todo.

Na construção de uma rede de *frames*, estão envolvidos: a classificação, visto que as redes são divididas em domínios, os domínios divididos em quadros, e os quadros englobam vários níveis de especificidade usando a subordinação hierárquica. Devido a esse aspecto, os autores (2005, p. 5) ressaltam a importância da pesquisa terminológica baseada em *corpus*, pois, primeiramente, o *corpus* mostra o contexto no qual o ST está inserido, trazendo à tona o conceito. Em segundo lugar, os contextos permitem recuperar os termos relacionados, os traços distintivos, o(s) sinônimo(s) e as relações entre os conceitos interrelacionados. Em terceiro

³³ As características de um dicionário ideal baseado em frames seriam: 1. identificação do item numa frase; 2. análise da frase por um analisador sintático (parser) e relação com elementos semelhantes à frase buscada; 3. definições e usos do item buscado surgem como *pop-up*; 4. caso o usuário não conheça o *frame* com o item buscado, ele pode ter acesso a itens similares para entender melhor o contexto; 5. exemplos de uso em outros contextos são disponibilizados, a partir de um *corpus*; 6. o dicionário pode disponibilizar dados enciclopédicos do item buscado.

lugar, os contextos possibilitam a verificação conceitual de informações gramaticais e conceituais extraídas de obras de referência, especificam o sentido de abreviações e disponibilizam informações enciclopédicas para serem usadas no PD.

Souza (2019, p. 71-72) e Faber Benítez et al. (2005, p. 3-4) mostram que a estrutura conceitual dos termos é uma estrutura prototípica que emerge de situações convencionais reveladas nos *frames*. Ações prototípicas ativariam *frames* ligados a especialistas, o que envolve situar conceitos em quadros específicos, em contextos de processos dinâmicos que descrevem e definem as áreas de especialidade. Esse percurso faz parte do processamento da terminologia e a frequência maior dos traços distintivos observada no *corpus* revelaria o conceito ou o *frame* útil à definição de um dado ST. O *corpus* provém desses contextos e serve à função de validar os traços distintivos utilizados para a construção dos PD nesta pesquisa.

1.7. Terminografia

A Lexicologia tem a Lexicografia como parte prática, da mesma forma que a Terminologia tem a Terminografia. Apesar de a Lexicografia/Terminografia serem concebidas como a parte prática, essas subdivisões têm princípios teórico-metodológicos que devem ser seguidos e respeitados. Os princípios que regem a prática terminográfica podem ser vistos como a Metaterminografia. Nos parágrafos seguintes vamos apresentar como os teóricos da Terminologia fazem a distinção entre essas duas subáreas das Ciências do Léxico.

A Terminologia, segundo Andrade (2001), tem como objeto a “recompilação, descrição e ordenação dos termos científicos e tecnológicos das linguagens especializadas” (ANDRADE, 2001, p. 194), e o produto são obras terminográficas. O objeto da Terminografia é a “atribuição de denominação aos conceitos” (BIDERMAN, 2001, p. 19).

A Terminografia se caracteriza também como uma disciplina. O termo foi proposto por Rey (1995, p. 23-24) em 1977, comparando a atividade à Lexicografia com embasamento teórico na Terminologia como ciência aplicada, uma intersecção entre o conhecimento e a linguagem. Atualmente, seus métodos e reflexões teóricas têm sido estimulados pela Inteligência Artificial e pela Engenharia do Conhecimento (KRIEGER, 2001, p. 51; MACIEL, 2001). O ponto comum dessas áreas são a “estruturação de conceitos e [o] mapeamento conceitual dos diferentes campos do saber” (MACIEL, 2001, p. 45).

Finatto (2014) define a Terminografia como “a disciplina prática intimamente ligada à Terminologia”, e que se ocupa das

descrições das propriedades linguísticas, conceituais e pragmáticas das unidades terminológicas de uma ou mais línguas, a fim de produzir obras de referência, tais como dicionários, glossários, vocabulários, em formato papel ou eletrônico, bases de dados terminológicas e bases de conhecimento especializado (FINATTO, 2014, p. 439).

Cabré (1999) define a terminografia como atividade que “envolve a compilação, sistematização e apresentação de termos de uma área específica do conhecimento ou atividade humana [...] regida por um conjunto de recomendações procedimentais, formais e técnicas acordadas internacionalmente”³⁴ (CABRÉ, 1999, p. 115.).

A autora enumera alguns princípios que fazem parte da terminografia: (1) os termos são unidades de forma e conteúdo inseparáveis; (2) a forma e conteúdo tendem a evitar a ambiguidade; (3) os termos têm um lugar definido em um campo conceitual específico, sem o qual o termo não terá significado específico; (4) conceitos estão interligados e inter-relacionados na área específica na qual se constituem juntos (CABRÉ, 1999, p. 116).

Como aspectos práticos da Terminologia, vários autores pontuam alguns passos que fazem parte do fazer terminográfico: (a) construção do mapa conceitual ou árvore de domínio; (b) extração de termos; (c) inserção dos termos na ontologia; (d) preenchimento das fichas terminológicas; (e) elaboração da base definicional; (f) elaboração das definições; e (g) edição dos verbetes (ALMEIDA; ALUISIO; OLIVEIRA, 2007, p. 410).

Dentre esses princípios, buscaremos descrever os ST da área da Linguística, extraídos de *corpora* que trazem contextos discursivos da linguagem especializada para a construção do VoBLing. É um trabalho baseado em princípios terminológicos ancorados na TCT e na Terminografia. Na próxima seção, trabalharemos aspectos teórico-metodológicos que farão parte da construção das definições dos ST desta pesquisa.

1.1.7. Definições

A definição, vista da perspectiva da Terminologia/Terminografia, é a parte textual que fará a descrição do conceito de um termo e uni-lo-á ao seu ST. Visando a construção das definições, é necessário que haja a sistematização conceitual para organização dos traços distintivos constituintes do ST. Esta prática pode ser regulada por organismos internacionais como a ISO.

³⁴ Terminography involves gathering, systematizing, and presenting terms from a specific branch of knowledge or human activity [...] governed by a series of technical, formal, and procedural recommendations that have been internationally agreed. (CABRÉ, 1999, p. 115)

Antes de iniciar esta reflexão, é necessário definir conceito na Terminologia. Em Wüster, o conceito é um constituinte do pensamento, que corresponde às características por meio das quais os seres humanos percebem os objetos do mundo real como instrumento da classificação mental. Essas características individuais são chamadas de intenção conceitual, correspondente ao conceito (WÜSTER, 2003 p. 39).

Definir é “estabelecer um vínculo entre um termo, um conceito e um significado” (FINATTO, 2001b, p. 118). Neste passo,

são mobilizados, constituídos e atualizados, em distintos níveis, diferentes valores e potencialidades de conhecimento e significação, descritos com relativa dificuldade por aqueles aparatos analíticos que a tradição dos estudos linguísticos qualifica como estritamente linguísticos” (FINATTO, 2001b, p.118).

Especificamente para esta pesquisa, entendemos que seria um desafio definir os termos da Linguística, pois fazem parte das ciências humanas e sociais (TEBÉ, 1996, p. 110 apud FINATTO, 2001b, p. 146). Ademais, esta dificuldade já foi atestada por especialistas da área da Linguística e Terminologia no Brasil, “se esse ‘perigo’ de limites imprecisos entre as construções e os termos já foi apontado por quem percorreu os caminho das ferramentas e das máquinas, o que mais se poderia esperar no campo das linguísticas? (FLORES *et al.*, 2009, p. 28). Contudo, é necessário romper limites nesse processo de **conceptualização** e nominalização, já que a Linguística se coloca como um campo científico no século XXI.

O conceito é uma unidade pré-linguística de conhecimento, a parte mais abstrata do ST, nomeado no percurso onomasiológico como **mundo referencial** (ver subitem 2.4.3). Ele faz parte do mundo das ideias, é abstrato e, após a conceptualização (BARBOSA, 2014b), será denominado, pelo ST, integrante de um sistema linguístico de uma língua natural (LN) (POTTIER, 1992). A este processo, Barbosa (2014b) denomina **definição**, isto é, “resultado de uma análise e descrição de grandezas sgnicas, situando-se, pois, no nível semiótico, pelo menos no que diz respeito ao seu ponto de partida” (BARBOSA, 2014b, p. 417).

Vale a pena retomar a discussão entre significado e conceito aqui, lembrando que o signo linguístico é composto de significado e significante, sendo que o significado de uma palavra pode ser polissêmico (denotação e conotação), na perspectiva lexicológica. Em contrapartida, o conceito é uma “unidade de conhecimento e contém, [...] apenas conhecimento factual ou técnico e não deveria abrigar elementos emotivos ou conotados” (FINATTO, 2001a, p. 309). Consequentemente, o ST seria monossêmico, gerando assim a **univocidade** terminológica.

Nessa mesma linha de interpretação sobre a **univocidade**, Krieger (2001) afirma que os termos técnico-científicos, ou ST, são considerados como “ideais de expressão de monoreferencialidade, de monossemia e de exclusividade denominativa.” (KRIEGER, 2001, p. 47). Esse é um princípio que levamos em consideração, pois acreditamos que a busca pela definição de ST de subáreas diferentes, mas homônimas, ou aquelas homônimas e plurissêmicas, será um desafio. Consideremos a título de exemplo os ST língua e *language*, presentes na maioria das subáreas da Linguística Descritiva e Aplicada, em português e inglês. Explicamos que, nesta pesquisa, usaremos a expressão Linguística Descritiva em contraste com a Linguística Aplicada, ambas subáreas da grande área da Linguística.

1.1.8.Paradigma definicional

Quando se trata de obras terminológicas, quer sejam vocabulários ou dicionários, a maneira como as definições são construídas nos levam a classificá-las em subtipos diferentes. A forma como um terminólogo organiza as informações usadas para se construir uma definição denomina-se **paradigma definicional** (PD) que, segundo Finatto (2001b), podem ser (i) enciclopédicos (PE), (ii) lexicográficos ou (iii) terminológicos. O (i) PE é mais detalhado, trazendo em si a definição lexicográfica, o histórico do termo, dados geográficos, culturais, históricos, sociais, entre outros. O (ii) PD de cunho lexicográfico pode trazer dados gramaticais, morfológicos e fonológicos do termo, sua definição e exemplos, dependendo da abordagem do lexicógrafo. Finalmente, o (iii) PD de cunho terminológico valoriza os traços distintivos, 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).

Tendo em vista que a Terminologia tem como o objeto o conceito do ST, é relevante considerarmos a intersecção da Terminologia com a Semântica. Consequentemente, apresentaremos algumas considerações sobre o aspecto conceitual do ST no processo de terminologização.

1.1.9.Percurso onomasiológico e semasiológico

As explicações sobre os **percursos onomasiológico e semasiológico** a seguir são propostas por Pottier (1992) no que tange à comunicação e como ela se realiza. Além de Pottier (1992), traremos a contribuição de Fiorin (2002) e de Barbosa (2004) na Linguística. Na

Terminologia, abordamos as diferenças desses percursos, e é ponto comum que o **percurso semasiológico** faça parte da metodologia da Lexicologia/Lexicografia, e o **onomasiológico** da Terminologia/Terminografia (BARBOSA, 1990; ANDRADE, 2001; KRIEGER, 2001; BARROS, 2004; KRIEGER; FINATTO, 2004). Ambos os percursos lidam com dois aspectos constitutivos dos signos terminológicos e das **lexias** ou **unidades lexicais**: os conceitos ou significados e a denominação dos ST e as lexias.

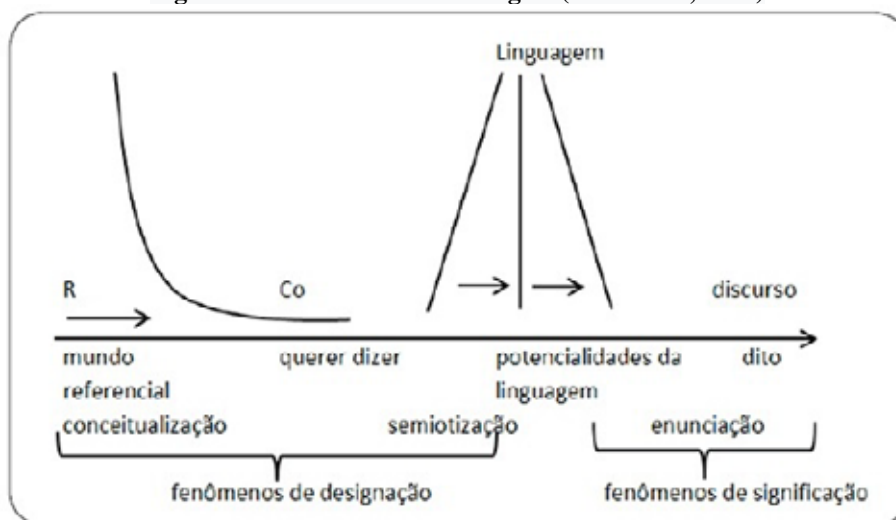
Sobre a produção de conceitos, Fiorin (2002) mostra a importância do léxico ao abordar o aspecto simbólico e virtual da organização do pensamento humano no processo de designação e afirma que,

A atividade linguística é uma atividade simbólica, o que significa que as palavras criam conceitos e esses conceitos ordenam a realidade, categorizam o mundo. Por exemplo, criamos o conceito de pôr-do-sol. Sabemos que, do ponto de vista científico, não existe pôr-do-sol, uma vez que é a Terra que gira em torno do Sol. No entanto, esse conceito criado pela língua determina uma realidade que encanta a todos nós. Uma nova realidade, uma nova invenção, uma nova ideia exige novas palavras, mas é sua denominação que lhes confere existência. Apagar uma coisa no computador é uma atividade diferente de apagar o que foi escrito a lápis, à máquina ou à caneta. Por isso, surge uma nova palavra para designar essa nova realidade, deletar. No entanto, se essa nova palavra não existisse, não se perceberia a atividade de apagar no computador como uma coisa diferente. (FIORIN, 2002, p. 56)

Assim como o autor pontua os aspectos da neologia, é importante considerar que, na Terminologia, a necessidade de haver novos termos que denominem novos conceitos é marcada por uma dinamicidade maior que na comunicação diária, pois o conhecimento científico tem crescido exponencialmente (KRIEGER; FINATTO, 2004; ALVES; TAGNIN, 2012).

Explicitaremos adiante o processo de conceituação na Terminologia, rumo à terminologização, quer seja por meio de um **signo terminológico**, como **deletar**, mencionado por Fiorin (2002), quer seja por meio da **lexemização** (*la mise en lexème*), partindo de um **signo linguístico**, ou **lexema** da linguagem geral ou ainda língua comum (RONDEAU, 1984) para a linguagem de especialidade (BARBOSA, 2014a). Logo, um questionamento que se levanta é: como o conceito se constrói na mente dos falantes? Observe parte deste percurso na **Figura 1**.

Figura 1 - Percurso onomasiológico (POTTIER, 1992)



Fonte: Martins e Zavaglia, 2014, p. 446.

De acordo com a Figura 1, Pottier (1992) explica que os falantes trazem consigo a introspecção, a intuição, a intenção do **dizer**, um ponto de partida referencial (R), podendo ser de qualquer natureza (odor, som, visão, lembrança, o texto do outro). Os falantes têm consciência da intenção do **dizer** à medida que **conceptualizam** (Co) seu desejo de significar, ou intenção de significação, conforme Barbosa (1990). Essa organização mental deve então ser expressa por meio de signos, semiotizados, em uma língua natural (LN).

Barbosa (2004, p. 56), falando a partir de uma perspectiva da Semântica cognitiva e da Semântica linguística, vai mais fundo nas análises e menciona os constructos do primeiro nível precedente à codificação do **conceito**: o arquiconceito³⁵, o metaconceito³⁶ e o metametaconceito³⁷. Antes de definir conceito, a autora salienta a importância dos discursos ou contextos, berços da cognição e da semiose, a partir dos quais os conceitos se formam pela codificação dos traços semântico-conceptuais. É a partir desse contexto cognitivo que acontece a conceptualização: o engendramento de um conceito e sua manifestação linguística. A **conceptualização** parte da percepção de fatos naturais, informação potencial convertida em substância estruturada pelos falantes de uma língua, provenientes de uma sociedade, detentores de uma cultura específica e mantenedores de um núcleo de percepção biológica universal.

³⁵ O **arquiconceito** ou conceito *stricto sensu* é o conjunto de traços semânticos-conceptuais comuns a todas as culturas, ou universais, utilizado para designar um termo (BARBOSA, 2004, p. 62-63). Por exemplo, o arquiconceito de que mãe é aquela que dá a vida = campo conceptual como conjunto unitário.

³⁶ O **metaconceito** é o conjunto de traços semânticos-conceptuais característicos de cada cultura, povo, língua ou nação (noemas culturais) = campo conceptual como conjunto vazio.

³⁷ O **metametaconceito** é o metaconceito modalizado marcado pela subjetividade do falante como sujeito político, econômico e social, isto é, o metaconceito agregado aos semas conceptuais intencionais, modalizadores e/ou manipulatórios (disforia ou euforia) = campo conceptual como conjunto múltiplo.

A língua geral é composta de seu sistema linguístico, em virtualidade, e de mecanismos de enunciação que permitem as realizações discursivas. Na Terminologia, observamos que o conceito não é o ponto final do processo de terminologização. Antes que haja a associação do termo ao conceito, é necessário observar atentamente a rede de conceitos que subjaz ao conceito *per se*, já que esse **um conceito** é extraído de uma rede de conceitos ou microssistema conceptual que compartilha características similares entre si (MACIEL, 2001; BARBOSA, 2014b).

Exemplificando esse princípio, vejamos como conceituar o termo *Terminologia* em detrimento da *terminologia*. Sabemos que a *Terminologia*, transcrita com letra maiúscula, é concebida como ramo da Ciência ou como disciplina. Já a *terminologia* engloba o vocabulário específico de uma área de especialidade ou a atividade desenvolvida pelos terminólogos (BARBOSA, 1990; KRIEGER, 2001; MACIEL, 2001; ANDRADE, 2001; TALAVÁN, 2016). Em suma, entendemos que as terminologias estão inter-relacionadas, mas cada uma traz traços específicos que as diferem: a Ciência, o fazer e o vocabulário (MACIEL, 2001). Há outros mais, mas não é nosso objetivo discutir os traços distintivos dos *signos terminológicos* exaustivamente, mas sim mostrar que há uma rede de conceitos à qual eles pertencem.

Da mesma forma como os *conceitos* e os *ST* se inter-relacionam, assim os *signos linguísticos* e os *significados* também se inter-relacionam para produzir os sentidos e os discursos. Essa relação, presente no sistema linguístico da linguagem geral, também se repete na linguagem de especialidade (ANDRADE, 2001, p. 198). Observemos a explicação de Fiorin (2002):

O valor de um signo é dado por outro signo. Além disso, um signo é sempre interpretável por outro signo: no interior do mesmo sistema pelos sinônimos, pelas paráfrases, pelas definições; em outro sistema, em outra língua, por exemplo, pela tradução. A dificuldade de traduzir indica que não há univocidade na relação entre os nomes e as coisas. (FIORIN, 2002, p. 56)

A reflexão levantada por Fiorin (2002) persiste na Terminologia, quando vemos que há *ST* que se confundem com outros dentro dos sistemas conceituais. Um exemplo deste fato é a recorrência do *ST língua*, analisado dentro da área da Linguística, na qual observamos que o *significante* é o mesmo, mas o *conceito* é distinto entre as subáreas. A este processo de denominação dos conceitos, temos o **percurso semasiológico**, corroborado por Rey (1995). “No plano nocional, para que um nome tenha direito ao título de termo, é necessário que ele possa, enquanto elemento de um conjunto (uma terminologia) ser distinguido de outro nome” (REY, 1995, p. 40-41).

O **percurso semasiológico**, que temos na Lexicologia/Lexicografia, faz parte da Terminologia/Terminografia atual devido aos avanços das análises de *corpus* computadorizados. Graças à Linguística de *Corpus* e ao uso de ferramentas computacionais para análises lexicais, tornou-se possível mudar esse percurso na metodologia da Terminologia, de forma que identificamos as palavras-chave nos *corpora* de especialidade e os respectivos contextos nos levam aos conceitos que a definem. Essa abordagem é adotada neste trabalho e no que o antecedeu, na construção do protótipo de Vocabulário Bilíngue de Linguística Histórica (LH) (YAMAMOTO, 2015).

1.1.10.Contextos definitório, explicativo e associativo

Expostos os conceitos dos **percursos semasiológico** e **onomasiológico**, explicamos que há teóricos que defendem a ideia de que o **percurso semasiológico** pertence à prática lexicográfica e o **onomasiológico** pertence à prática terminográfica (BARROS, 2004; KRIEGER, 2001; KRIEGER; FINATTO, 2004).

Como explicado no subitem anterior, 2.4.3, a partir do advento da Linguística de *Corpus* e do uso de sua metodologia no fazer terminográfico, o **percurso semasiológico** passou a fazer parte da Terminografia. Isto se deve ao fato de que as ferramentas computacionais fornecem ao pesquisador dados textuais mais específicos, como no caso das listas de palavras-chave provenientes da análise quantitativa e probabilística dos *corpora*.

Nesta pesquisa, por exemplo, a partir das palavras-chave, nós podemos fazer uma análise qualitativa e semântica do contexto no qual o candidato a termo se insere. Dependendo da carga semântico-discursiva desse contexto, os traços distintivos, componentes do possível termo, podem ser extraídos para a redação da definição. Esses contextos são classificados em três tipos: **definitórios**, **explicativos** e **associativos** (AUBERT, 2001; DUBUC, 1992; PAVEL; NOLET, 2002; BARROS, 2007).

A definição desses conceitos, por Pavel e Nolet (2002), se dá da seguinte forma:

Os **contextos definitórios** apresentam características essenciais do conceito em estudo, enquanto que os **contextos explicativos** fornecem informação sobre algumas das características. Os **contextos associativos** demonstram o uso do termo na área em estudo, mas não auxiliam a ilustrar a equivalência textual por meio da correspondência de traços semânticos. (PAVEL; NOLET, 2002, p.48, grifo nosso)

Em outras palavras, os **contextos definitórios** são aqueles usados pelos autores para definir algo ou alguém, usando para tal uma rede de traços distintivos para delimitar o objeto,

de forma que o sentido fique claro para seu interlocutor. Neste trabalho, algumas marcas que usamos para identificação dos **contextos definitórios** são o verbo **ser**, a **vírgula** e o **hífen**, identificados como colocados na ferramenta do *Concord*, ou por meio da ferramenta de busca do WST.

O **contexto explicativo**, conforme Barros (2007, p. 39), “apresenta, sucintamente, dados a respeito da natureza e de certos aspectos do termo, sem defini-lo claramente”. Em geral, neste tipo de contexto, encontramos alguns dados sobre o objeto, como: materialidade, finalidade, funcionamento e constituição. É semelhante ao que Wüster (2003) propõe sobre **características extrínsecas** do objeto como propriedades de uso e de origem.

Por fim, o **contexto associativo** mostra o termo em uso, sem que traga em seu contexto elementos que possam ser usados para sua definição. De acordo com Pavel e Nolet (2002), o **contexto associativo** seria equivalente ao **exemplo de uso**. Os referidos autores o definem da seguinte forma: “Breve citação que ilustra, em uma ficha terminológica, o uso de um termo em uma área temática, sem fazer referência aos traços distintivos do conceito designado” (PAVEL; NOLET, 2002, p. 121).

Para o fazer terminográfico deste trabalho, são usados os **contextos definitórios e explicativos**, encontrados nos “gêneros científicos de divulgação e instrucional, por exemplo: apostila, livro-texto, manual” (ALMEIDA; CORREIA, 2008, p. 81).

Na subseção seguinte, trataremos dos conceitos do ST e como eles se aplicam a esta pesquisa.

1.1.11.O signo terminológico (ST)

O ST é formado pelo **conceito** e pelo **termo** (denominação dada ao conceito, de *terminus* “borda, limite”). Neste trabalho, adotaremos a expressão **signo terminológico (ST)** para designar **termo** ou **unidade terminológica**. Compartilhando a concepção de Cabré sobre a constitutividade do termo, a saber: “Os termos, como as palavras do léxico geral, são **unidades sígnicas distintivas e significativas**, ao mesmo tempo que se apresentam de forma natural no discurso especializado” (CABRÉ, 1993, p. 169; CABRÉ, 1999, p.196).

Para além do ST, nossa pesquisa contemplará termos compostos, isto é, aqueles formados por duas **unidades terminológicas**. Para nomear esses termos compostos, adotaremos a expressão *signo terminológico fraseológico*, doravante STF. Esta opção assemelha-se, dentro da Lexicologia, ao uso do termo *unidade lexical especializada*, correspondente ao que na Terminologia denominamos *unidade terminológica* (UT).

Para explicar os termos compostos da Linguística, partiremos do conceito de *fraseologia* de Barbosa (2012, p. 249), para quem é importante considerar “A classe de equivalência sintática e semântica; classes de elementos não idênticos, mas que podem ser agrupados de acordo com alguns critérios”. Isto é: o STF *mudança linguística* une as UL *mudança* (subst.) e *linguística* (adj.) para formarem, além dos sentidos de mudança e linguística, um terceiro sentido: *estudo das alterações graduais das línguas no tempo*, específico da subárea da LH. Os critérios que os classificam dentro dessa subárea como fraseologia são a recorrência, o uso, a fixidez e a estrutura (duas ou mais UL) (YAMAMOTO; LISBOA, 2019). Além desses conceitos, abarcamos a proposta de Fromm (2020) que conceitua os fraseologismos como “qualquer agrupamento de duas ou mais palavras cuja frequência combinada seja destacada pelos programas de análise lexical.” (FROMM, 2020, p. 769).

Finatto (2001b) cita Cabré (1994) e Rey (1995) ao tratar da **conceituação** e da **definição** na Terminologia, explicando que a **definição terminológica** não busca representar ou reproduzir um termo, mas evocar um conceito; assim, busca-se ressaltar uma **noção**, dentro de um conjunto de outras noções (FINATTO, 2001b). Se considerarmos o princípio da Lexicologia, em que uma unidade lexical, diferente de outra, produz sentidos diferentes, compreenderemos que há fronteiras distintas entre o **conceito**, o **significado** e a **definição**.

Diferentemente da constituição tradicional do signo linguístico proposta por Saussure (2012), o **significado** não é sinônimo de conceito no que tange ao ST. O **significado** de uma unidade lexical deve compreender todas as suas acepções, envolvendo tanto o sentido denotativo quanto o conotativo, gerando vários significados. Essa polissemia não acontece com o ST, já que seu conceito seria único (monossêmico), estável e universal (BARBOSA, 1999; ANDRADE, 2001; FINATTO, 2001b; KRIEGER, 2001).

Para retratar um **conceito**, pensemos em uma analogia com os gomos de uma poncã, que representariam os **traços distintivos** do ST, e com o fruto da poncã, que seria o ST. Consideremos então que as frutas cítricas são a **rede conceitual** de um ST. Apesar de termos vários representantes da rede dos frutos cítricos, como a tangerina, o limão, a laranja e a murcote, cada um tem sua especificidade. Os citros têm características físicas comuns: árvores semelhantes, de porte médio a alto; folhas de tamanho médio, pontiagudas, de base arredondada; frutos do tipo baga, gomos que podem se soltar da casca do fruto, com maior ou menor facilidade. Por esta razão, os citros são classificados em três famílias: Meliáceas, Simaruláceas e Rutáceas (GOMES, 1989).

Assim como um fruto cítrico pertence a uma árvore e essa árvore a um conjunto de outras semelhantes entre si, assim também os **traços distintivos** fazem parte de uma rede maior,

pertencentes ao mundo conceitual das ideias (mental). Quando identificamos estes **traços** e os diferenciamos dos outros, chegamos à definição de um ST. Maciel (2001) define os **traços distintivos** da seguinte forma: “são traços nítidos e constantes, que, examinados semanticamente, revelam-se exclusivos em seu conjunto. Esses traços particulares caracterizam a **individualidade do domínio**, constituindo a pertinência temática” (MACIEL, 2001, p. 278). Este é o trabalho que buscamos desenvolver nesta pesquisa: identificar os **traços distintivos** específicos que caracterizam cada ST/ “poncã”, partindo do princípio de que os **traços distintivos**/“gomos” mais recorrentes nos *corpora* servirão à construção da **definição terminológica** (DT).

Do ponto de vista terminológico, esses **conceitos**, compostos de **traços semânticos/distintivos** ou de sentido, servem como unidades de pensamento (conceitual/mental), de conhecimento (especificidade da área) e de comunicação (pragmática/uso dos especialistas).

Nesta pesquisa, prevemos um desafio que se manifestou na elaboração do vocabulário de LH: como organizar os dados na DT de ST abstratos? Esta dificuldade foi pontuada por Cabré (1991, p. 61) ao tratar da normatização da terminologia de uma área de especialidade. A autora expôs que várias disciplinas das Ciências Humanas e Sociais apresentam dificuldade quanto à organização conceitual de certos termos, ao afirmar que em várias disciplinas dessas áreas há a impossibilidade de isolar seus conceitos, ocasionando uma variedade de pontos de vista e, conseqüentemente, uma diversidade conceitual do mesmo objeto³⁸.

Na próxima seção, abordaremos a **polissemia** e a **homonímia** na Terminologia e como trabalharemos essas especificidades nesta pesquisa.

1.1.12. Homonímia e polissemia

O papel da Terminologia é analisar o termo em relação aos aspectos semânticos, funcionais, formais e de seu *status* a partir de um *corpus* (CABRÉ, 1991, p. 61). Na Terminologia, a relação entre o **conceito** e a **forma** de um termo não devem gerar ambiguidade, logo, a **homonímia** é o fenômeno que designa palavras diferentes, com possíveis origens diferentes, porém que apresentam a mesma forma. Por exemplo, há a UL **manga**, que pode ocorrer em contextos diversos com sentidos diferentes: (1) “A manga da camisa de Pedro é muito curta”; (2) “Manga é a fruta favorita de Tiago”. Segundo a etimologia proposta por

³⁸ «...plusieurs disciplines sociales et humaines où l'impossibilité d'isoler leurs concepts provoque une multiplication des points de vue et donc une diversification conceptuelle du même objet».

Nascentes (1955), o exemplo (1) é um substantivo que descreve uma parte das vestimentas, proveniente do latim *manica*³⁹; no exemplo (2), é o nome da fruta tropical de origem malaiala, língua de uma região da Índia. Em outras palavras, temos a mesma **forma lexical**, cujo **conteúdo semântico** e etimologias são diferentes.

Embora a relação única entre **conceito** e **termo** seja desejável para a Terminologia, a **polissemia** e as mudanças na relação de sentido e forma têm sido encontradas. Em contrapartida, Kageura (2015) pontua a importância da **monossemia** instituída por meio do controle social e prescritivo do ST, principalmente em contexto empresarial e industrial, objetivando uma comunicação efetiva. O autor explica que, apesar das iniciativas de **padronização**, a **polissemia** se justifica e se faz presente na comunicação especializada, pois os ST são itens lexicais subordinados ao dinamismo do sistema linguístico.

Schmitz (2015) aborda a **homonímia** da seguinte forma: um ST deveria corresponder a um conceito, e **homônimos** e sinônimos não deveriam existir na mesma subárea de uma Ciência. Esse objetivo nem sempre pode ser atingido, já que inventores, responsáveis por produtos e empresas, criam termos em espaços físicos e temporais distintos (SCHMITZ, 2015, p. 454, 455). A distinção na **temporalidade** difere da proposta de Wüster, pois a univocidade seria aplicada a dado ponto no tempo, geralmente no presente. Por essa razão, concebe-se a inexistência de **sinônimos**, de **homônimos** e de **polissemia**, de acordo com sua teoria (WÜSTER, 2003).

Contrária à abordagem da **homonímia** adotada pela Terminologia, a Lexicografia engloba a língua geral, conseqüentemente, a quantidade de **homônimos** na língua geral será menor que na Terminologia, pois estes são concebidos como **polissêmicos** e **não homônimos** (CABRÉ, 1999). Isto acontece porque na **Terminologia** os termos são classificados a partir dos sistemas conceituais de áreas de especialidade. Ou seja, os termos *pé* serão: a parte de um móvel na marcenaria; uma parte do corpo humano ou de um animal na Medicina e na Veterinária; representará um corpo inteiro como pé de manga ou mangueira na Agronomia. Em contrapartida, na **Lexicologia**, *pé* será uma UL polissêmica e não um homônimo.

Quanto à **polissemia**, Cabré (1999) explica que a Terminologia se baseia no princípio de que uma **designação** corresponde a um único **conceito**, apesar de essa relação unívoca nem sempre acontecer na prática. Isso ocorre pois o **valor semântico** do ST é estabelecido com base na sua relação com um **sistema conceitual** específico em contraposição a outros **sistemas**

³⁹ Antenor Nascentes afirma que o termo manga (tubo) é proveniente do latim *manica*, *manica* em italiano e *manche* em francês. Quanto ao nome da fruta, é de origem indiana, “próprio do fruto verde”, introduzido pelos portugueses (NASCENTES, 1955, p. 314).

conceituais. Consequentemente, o que em Lexicografia se denomina **polissemia**, em Terminologia é concebido como **homonímia**. Apesar da existência desse princípio, Cabré (1999) pontua que a **polissemia** pode ocorrer na Terminologia, quando ST idênticos compartilham diferentes conceitos. Isso ocorre quando um ST possui um *status* conceitual diferente, dependendo da subárea a que pertence, dentro da mesma grande área (CABRÉ, 1999). Um exemplo é o conceito do ST *língua* nas diferentes subáreas da Linguística, considerando o trabalho terminológico/terminográfico que temos desenvolvido.

Pavet e Nolet (2002) definem **polissemia** como a “relação de dois ou mais conceitos em uma mesma designação”. As autoras pontuam que a **polissemia** faz parte do fazer lexicográfico, relacionado aos dicionários de língua geral, em oposição ao princípio uninocional, característico do fazer terminográfico, no qual cada conceito deve ser tratado em separado. Essa linha de interpretação concorda com as propostas de Wüster e da ISO, nas quais cada conceito só pode ser relacionado a um ST, o que seria factível caso a terminologia científica fosse um sistema nominativo à parte do sistema de uma língua natural (KOCKAERT; STEURS, 2015; WÜSTER, 2003). Em contrapartida, Rey (1995) define **polissemia** a partir da perspectiva lexicológica e terminológica: “um signo para vários significados/sentidos; um termo para vários conceitos” ⁴⁰ (REY, 1995, p. 70.).

Ainda sobre a discussão conceitual de **homonímia** e **polissemia**, é relevante expor a teoria de Wüster quanto à relação conceito x nominalização do termo. O autor considera monossêmico o termo que, em todos os contextos de uso, mantém a relação termo/UL e conceito/significado. Já quanto à univocidade, o autor explica que, em sentido restrito, um **termo unívoco** ou **monovalente** é **aquele que, em dado contexto discursivo**, determinado pelas circunstâncias, **só tem um significado**, mesmo que possa ser polissêmico. Portanto, o contexto discursivo gera, produz e propicia a **monovalência conceitual** (WÜSTER, 2003). Advém que Steurs, De Watcher e De Malsche (2015, p. 223, 224) afirmam, baseados em Bertels (2011, p. 95, 96), que o **caráter monossêmico** do ST tem sido questionado pelos terminólogos na última década, o que já era afirmado por Rey (1995) anos atrás: “a **polissemia** e a **sinonímia**, as quais, por uma questão de princípio, não deveriam existir na terminologia, são, na verdade, bastante comuns” ⁴¹ (REY, 1995, p. 56, tradução do autor).

Ademais, Cabré (1999) explica que a **polissemia** é uma das formas mais profícuas de se estender o léxico de uma língua por meio da analogia de conceitos, o que permite que a

⁴⁰ “one sign for several meanings; one term for several concepts” (REY, 1995, p. 70).

⁴¹ “polysemy [...] and synonymy [...], which as a matter of principle should not exist in terminology, are, in fact, very common.” (REY, 1995, p. 56).

designação de um **conceito** seja usada por outro signo. Dessa forma, um novo termo é criado por uma sobreposição semântica parcial (*partial semantic overlap*) o que ocorre no caso de vírus (Medicina e Computação); descongelamento (Hidrologia e Gestão, Economia e Políticas Públicas); e reação (Química, Física e Medicina).

Como possível solução, para que exista uma melhor compreensão dos princípios que governam a comunicação especializada, é necessário que haja uma abordagem que envolva elementos descritivos, linguísticos e semasiológicos, baseada em uma análise dos *corpora* de língua especializados, no intuito de se tratar o aspecto polissêmico de alguns signos terminológicos. A proposta feita pelos cientistas é a de acompanhar a mudança de sentido que se faz no sistema da linguagem de especialidade. Esse acompanhamento seria possível pela coleta e análise de *corpora*, envolvendo a compilação, o processamento, a atualização e a preparação dos dados terminológicos (STEURS; DE WATCHER; DE MALSCHE, 2015).

1.1.13. Árvore de domínio: princípios teóricos

Nesta seção são descritos os princípios propostos por Aubert (2001), Krieger e Finatto (2004) e Barros (2004) que orientam o passo a passo para a elaboração da árvore de domínio. A **árvore de domínio** ou grade conceitual é um passo metodológico imprescindível quando se realiza um trabalho terminográfico (AUBERT, 2001, p. 65; BARROS, 2004, p. 111-112; KRIEGER; FINATTO, 2004, p. 134) que precede à escolha dos termos e à construção das definições.

O sistema conceitual é definido como um “diagrama hierárquico composto por termos-chave de uma especialidade, semelhante a um organograma” (KRIEGER; FINATTO, 2004, p. 134) com coerência interna capaz de prever e de se adaptar a novos fenômenos; serve à identificação das relações existentes entre os termos de uma área ou subárea. Esse sistema também contribui para a elaboração de um sistema coerente de remissivas e pode ser representado pelo diagrama de uma árvore de domínios, como no caso de nossa pesquisa (BARROS, 2004, p. 129).

Barros (2004) explica que a árvore de domínio parte de um sistema conceitual já existente, em nosso caso, o da Linguística. Esse sistema conceitual não deve ser definitivo nem único, mas “deve ser flexível para comportar novas relações e novos termos” (BARROS, 2004, p. 127). Na relação hierárquica são organizados os termos baseados na relação conceitual partitiva ou genérica existente entre eles, por exemplo, a relação de hiperonímia e hiponímia. Há os conceitos superordenados e os subordinados: o superordenado é o conceito genérico ou

um conceito integrante; o subordinado é o conceito mais específico ou partitivo (ISO 1087: 2000, p. 4). Por exemplo, na árvore de domínio da Linguística, a Linguística Descritiva é um termo subordinado à grande área da Linguística, mas é superordenado em relação à Semântica.

De acordo com Aubert (2001, p. 63, 64), a árvore de domínio serve à delimitação do campo semântico a ser abordado, à identificação da ocorrência das noções e às designações em situação efetiva de uso. Ademais, auxilia na escolha do que é relevante para o trabalho, prevenindo o hibridismo temático, pois textos de áreas de especialidade podem tangenciar outras áreas. A construção da área de domínio evita e previne o ruído e o silêncio. Aubert (2001, p. 63) define o ruído como a mescla de termos de outras áreas em uma área específica e o silêncio é a omissão de termos básicos de uma área na árvore de domínio e, consequentemente, na obra terminográfica. Em nosso trabalho, a inclusão do termo em dada subárea é baseada em dados estatísticos e probabilísticos da LC, mais especificamente da chavicidade dos termos.

Krieger e Finatto (2004) enfatizam a importância de a árvore de domínio preceder à composição de uma obra lexicográfica ou terminográfica, servindo à organização e compreensão prévia das hierarquias básicas e à delimitação ou recorte do domínio analisado e sua relação com outras áreas. A árvore de domínio que adotamos nesta pesquisa é fruto da pesquisa terminológica iniciada por Fromm (2013; 2018). Em 2015, finalizamos a pesquisa para compor a árvore da LH (YAMAMOTO, 2015) e atualmente conduzimos os três primeiros níveis neste protótipo de Vocabulário de Linguística: (i) a grande área da Linguística, (ii) as subáreas da Linguística Descritiva (LD) e da Linguística Aplicada (LA); e (iii) um termo de cada subárea das LD e da LA. No produto desta pesquisa, o VobLing, identificamos os arquiconceitos e os metaconceitos dos termos nos *corpora* de Linguística em LP e em LI, de forma que os arquiconceitos sirvam de padrão para as definições terminológicas e os metaconceitos sejam utilizados nas definições enciclopédicas. Nesta pesquisa, excluimos os metametaconceitos, pois não analisamos um autor em específico. Além da teoria proposta por Barbosa (2004), nos baseamos em Fillmore (1975; 1976; 2003; 2006) e Faber Benítez et al. (2005) fundamentando a Semântica de *Frames* e a Terminologia de *Frames*.

1.8. Linguística de Corpus e corpus

Nesta seção, abordamos a LC como fundamentação teórica desta pesquisa. Além da explicação do que é a LC, trazemos o conceito de *corpus*. O caráter metodológico da LC será abordado na seção 3, de Metodologia.

1.1.14. Linguística de *Corpus*: histórico

Nesta seção, objetivamos trazer uma breve reflexão sobre o histórico da LC, no sentido de mostrar como ela se constituiu como disciplina, metodologia e abordagem no cenário da Linguística atual, baseados em McEnery e Wilson (2001) e Berber Sardinha (2000).

McEnery e Wilson (2001) fazem um apanhado do início da LC como metodologia, com abordagem baseada em *corpus*. Eles a definem como estudo da língua fundamentado em exemplos reais de uso, mas não a concebem como uma subárea da Linguística como Sintaxe, Semântica e Morfologia, apesar de compreenderem a existência de afirmações como a Sintaxe baseada em LC e a Sintaxe não baseada em LC.

Berber Sardinha (2000) assim define a LC de forma mais pragmática:

ocupa-se da coleta e exploração de *corpora*, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador. (BERBER SARDINHA, 2000, p. 325)

O autor ressalta o caráter empírico da LC, o fato de ela usar dados que sejam legíveis por computador, e sua natureza metodológica numa pesquisa científica da língua. Antes de descrever a cronologia da LC, tanto Berber Sardinha (2000), quanto McEnery e Wilson (2001) fazem questão de contrastar a relação da LC e as teorias de Noam Chomsky (1957) com o lançamento de sua obra *Syntactic Structures*, cujo **caráter mentalista** se opunha ao aspecto descritivo da língua por meio de *corpora*.

Em seguida, McEnery e Wilson (2001) explicam que a impopularidade da LC nos anos 1960 e 1970 se deveu à adesão de grande parte dos cientistas à teoria racionalista⁴² de Chomsky, em oposição ao **caráter empirista**⁴³ da LC. Nesse sentido, Chomsky (1957) defendia que os objetivos da Linguística eram a introspecção e a explicação da competência linguística, e não a enumeração e descrição dos fenômenos do desempenho linguístico.

No capítulo um, McEnery e Wilson (2001) mostram que, em alguns aspectos, os argumentos de Chomsky foram irrefutáveis, como, por exemplo, o fato de que a língua não é quantificável; logo, nenhum *corpus* finito pode representar a língua, infinita em sua integridade, quanto ao desempenho. Outra crítica foi a falta de confiabilidade nas análises linguística feitas

⁴² Teoria mentalista: busca explicar como a língua é mentalmente processada; seu objetivo é a plausibilidade cognitiva.

⁴³ Teoria empirista: busca evidências gramaticalizadas das estruturas linguísticas, por meio da observação de *corpora*.

a partir de *corpora*, pois naquela época não se contava com a análise de dados por meio de computadores como hoje (MCENERY; WILSON, 2001; BERBER SARDINHA, 2000).

Em contrapartida, em resposta à artificialidade dos exemplos propostos pela teoria de Chomsky, linguistas de *Corpus* mencionaram a questão da originalidade do uso da língua retratada nos *corpora*, respeitados os princípios da **representatividade** e do **balanceamento**. Numa tentativa de parcialmente concluir esse debate, os autores trazem a seguinte reflexão: “Se o linguista de *corpus* pode muitas vezes parecer o escravo dos dados disponíveis, então o não linguista de *corpus* pode interpretar os dados ao bel-prazer de sua imaginação” ⁴⁴ (MCENERY; WILSON, 2001, p. 15.).

Para a conclusão desse debate, McEnery e Wilson (2001) citam Fillmore (1992) com a declaração de que as duas formas de investigação da língua se complementam.

Não acredito que possa haver *corpora*, apesar do tamanho, que contenha informações sobre todas as áreas do léxico e da gramática do inglês que eu queira explorar ... [mas] cada *corpus* que pude examinar, apesar de pequeno, tem me ensinado fatos que não poderia imaginar encontrar de outra forma. Minha conclusão é que os dois tipos de linguistas precisam um do outro (FILLMORE, 1992, p. 35, apud MCENERY; WILSON, 2001, p. 25.) ⁴⁵.

A manutenção dessa lucidez para os estudos atuais em LC seria algo desejável e, certamente, traria mais ganhos que perdas para os estudos linguísticos.

Berber Sardinha (2000) menciona que a LC é estritamente ligada aos grandes *corpora* de línguas. Nos anos 1960, os computadores disponíveis eram os do tipo *mainframe*, e nos anos 1980, com o advento dos microcomputadores pessoais, a LC deu um salto rumo à análise de dados mais confiáveis e à produção de banco de dados maiores. O autor encerra o apanhado histórico citando países onde se desenvolvem pesquisas em LC, como a Grã-Bretanha (Birmingham, Brighton, Lancaster, Liverpool, Londres etc.); os países escandinavos (Noruega, Suécia e Dinamarca); e os Estados Unidos (Douglas Biber e o Processamento de Linguagem Natural – PLN).

Na próxima seção, apresentaremos o conceito de *corpus* na LC e como os *corpora* podem ser classificados.

⁴⁴ If the corpus linguist can often “seen” the slave of the available data, so the non-corpus linguist can be seen to be at the whim of his or her imagination. (MCENERY e WILSON, 2001, p. 15).

⁴⁵ I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore...[but] every corpus I had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way. My conclusion is that the two types of linguists need one another. (FILLMORE, 1992, p. 35, apud MCENERY e WILSON, 2001, p. 25).

1.1.15. *Corpus*: definição e taxonomia

A perspectiva de uma obra baseada em *corpus* advém da Linguística de *Corpus*, de sua **concepção de língua** e da **abordagem** que ela propõe para a construção de obras lexicográficas ou terminográficas. Devido à sua **concepção da língua** como um sistema probabilístico de escolhas (HALLIDAY, 1991), a LC propõe que as análises sejam feitas a partir de um *corpus* representativo da língua, com textos de ocorrência natural, autênticos, que permitam seu processamento por computador e que sejam criteriosamente selecionados para que o objetivo da pesquisa seja atingido (BERBER SARDINHA, 2004; VIANA; TAGNIN, 2010).

O que tratamos sobre **possibilidade** e **probabilidade** nos estudos linguísticos a partir da LC deve-se ao advento dos computadores propiciarem a análise de *corpora* gigantescos, o que revela a natureza probabilística da recorrência das estruturas linguísticas. Isto implica no fato de que mesmo que haja estruturas possíveis num sistema linguístico, a **análise quantitativa** mostra que há estruturas que são mais recorrentes que outras, dependendo da escolha dos falantes. Como exemplo, temos o fato de que os livros de inglês como língua estrangeira ensinam a estrutura de *can* + verbo como um dos primeiros modais, contudo, a estrutura *will* + verbo é mais recorrente na LI que aquela.

Detalhando o conceito de *corpus* dado por Viana (2010), temos: uma compilação de textos de ocorrência natural que representa uma certa língua ou seus aspectos específicos, possibilitando uma análise linguística preestabelecida. Tagnin (2015) define *corpora*, da seguinte forma: “bancos de textos de linguagem autêntica, criteriosamente construídos, destinados à pesquisa e legíveis por computador” (TAGNIN, 2015, p. 20).

Além de conceituar *corpus*, é necessário classificá-lo dentro da LC. Neste trabalho, adotaremos a proposta de Teixeira (2008), cuja taxonomia sintetiza os princípios estabelecidos até então por vários autores de referência da LC.

De acordo com Teixeira (2008), a **taxonomia** dos *corpora* seria aquela disposta no Quadro 1.

Quadro 1 -Taxonomia de *corpora* segundo Teixeira (2008).

1.	Língua	Monolíngues, bilíngues/multilíngues, multivarietais
2.	Data de publicação	Sincrônico/diacrônico, histórico/contemporâneo, fechado/aberto (<i>corpus</i> monitor ⁴⁶)
3.	Modo	Textos escritos; transcrições de textos orais; ambos
4.	Conteúdo	Representativos da língua geral; qualquer parte da língua ⁴⁷
5.	Uso na pesquisa	Estudo ou referência

⁴⁶ *Corpus* aberto ou monitor é aquele que está em constante expansão.

⁴⁷ Especializados, regionais etc.

6.	Autoria	Falantes nativos ou não nativos; individual, coletiva ou institucional
7.	Tamanho	Pequeno (até 80 mil palavras); pequeno-médio (80-250 mil); médio (250-1 mi).; médio-grande (1-10 mi); grande (acima 10 mi) ⁴⁸
8.	Nível de codificação	Cabeçalhos e etiquetas
CORPORA MULTILÍNGUES		
9.	Comparáveis	Dois ou mais <i>subcorpora</i> escritos na mesma língua e semelhantes entre si ⁴⁹
10.	Paralelos	Escritos em dada língua, acompanhados de suas traduções. Subtipo: alinhado (tradução lado a lado)

Fonte: Teixeira (2008, p. 209)

Na seção 4, da **Metodologia**, faremos a explicação pormenorizada dessa taxonomia e de como ela se aplica a este trabalho, quando da descrição dos *corpora* que compõem esta pesquisa.

⁴⁸ A autora recomenda observar o tamanho e a relação tamanho-autoria, para que a amostra não reflita idiossincrasias de autores ou predominância temática advindas de textos longos ou de mesma autoria.

⁴⁹ Semelhanças tais como: tipologia textual, gênero, tamanho, data de publicação, área e subárea temática etc.

3. OBRAS DE REFERÊNCIA: BREVE ESTADO DA ARTE

Neste capítulo, tratamos das obras de referência na área de Linguística e, dentre elas, encontram-se os vocabulários, os dicionários e os glossários. O estado da arte dos dicionários de Linguística objetiva mostrar o que já foi consolidado quanto à elaboração de obras terminográficas. Este passo metodológico faz a descrição e a análise das obras para evidenciar a especificidade dos dicionários. Para realizar esta análise, nos baseamos em Cardoso (2017), com a análise de dicionários de Teologia, e Faulstich (2011), com os parâmetros propostos para análise de dicionários.

Na proposta de Faulstich (2011), avaliamos os itens sobre o autor, a apresentação da obra pelo autor, a apresentação material da obra, o conteúdo, a edição e a publicação. Cardoso (2017) analisa os tipos de padrões definicionais presentes nas obras de Teologia. De igual modo, fizemos análise do padrão definicional no tocante às obras, considerando o padrão definicional terminológico e enciclopédico (FINATTO, 1994. p. 56- 57; CARDOSO, 2017, p. 192-193).

Neste estudo analisamos 18 dicionários na área de Linguística, impressos ou virtuais, monolíngues ou bilíngues, em português, inglês, francês, espanhol e japonês. Há onze dicionários de Linguística em português do Brasil, impressos e eletrônicos. Além desses, analisamos sete obras monolíngues ou bilíngues disponíveis on-line na área de Linguística e suas subáreas. Quanto à nomenclatura das obras impressas, há dez dicionários, e nas obras digitais, há nomenclaturas diversas: *Glossary/Glossaire*, *Glossary Terms*, *Términos*, *Dictionary* e *Lexicon*.

Quanto ao acesso aos dicionários, os impressos foram pesquisados nas universidades Federais de Uberlândia e de Jataí; os outros foram adquiridos por meio físico ou digital; e os virtuais foram acessados na internet.

O Quadro 2 traz a lista dos dicionários analisados: os impressos, os eletrônicos e os virtuais. Os itens iniciais da descrição das obras foram: o nome e o(s) autor(es). Em seguida, o item **A** representa o ano de publicação da obra; o item **B** indica se a obra é **tradução** (+) ou **original** (-); o item **C**, se há a versão impressa; o item **D**, se há a versão eletrônica ou on-line; e o item **E** representa o tipo de PD. Os tipos de PD observados foram o PD enciclopédico (E) e o terminológico (T). Com exceção do item B, as respostas afirmativas foram marcadas pelo sinal de positivo (+) e as respostas negativas pelo sinal de menos (-).

Quadro 2 - Dicionários de Linguística analisados (físicos, eletrônicos ou virtuais) em ordem de publicação.

Título	Autor	A.	B.	C.	D.	E.
1. Dicionário enciclopédico das ciências da linguagem	Ducrot e Todorov	1977	+	+	-	T/E
2. Dicionário de Linguística e Gramática: referente à língua portuguesa	Mattoso Câmara Jr.	1986	-	+	-	T/E
3. Dicionário de Termos Literários	Massaud Moisés	1995	-	+	-	T/E
4. <i>Glossary of Grammar Terms</i>	Johanson	1996	-	+	-	T
5. <i>Lexicon of linguistics</i>	Jan Don (Ed.)	1996-2001	-	-	+	T/E
6. <i>Bilingual Grammar Glossary for Japan</i>	David V. Appleyard	1998-2020	-	+	+	T
7. Dicionário de Linguística e Fonética	Crystal	2000	+(adap.)	+	-	T/E
8. <i>Systematic Dictionary of Corpus Linguistics</i>	<i>Lithuanian Terminological System of CL</i>	2000	-	-	+	T
9. <i>SIL Glossary of Linguistic Terms</i>	Loos	2003	-	-	+	
10. Dicionário de Linguística	Jean Dubois et al.	2004	+	+	-	T/E
11. Dicionário de semiótica	Greimas e Courtes	2008	+	+	-	T/E
12. Dicionário de linguagem e linguística	Trask	2015	+(adap.)	+	-	T/E
13. Dicionário de Linguística da Enunciação	Flores et al.	2009	-	+	+(K)	T
14. 'Letras' entre aspas	Silva	2015	-	+	+	T
15. <i>Index of French Grammar and Pronunciation Glossary Terms</i>	Meiners	2016	-	-	+	T
16. <i>Lingüística-Términos</i>	<i>Grupo de Ciencias del Language</i>	2016	-	-	+	T
17. Dicionário de gêneros textuais	Costa	2018	-	+	+	T/E
18. Bases para um dicionário linguístico-gramatical ⁵⁰	Silva	2019	+ / -	-	+ / -	T/E

Fonte: Elaborado pelo autor.

Os dicionários eletrônicos são aqueles em formato PDF, em CD ou em dispositivos móveis (celulares, i-pads), acessíveis on-line ou off-line. Os dicionários on-line são os tipos que só podem ser lidos com acesso à internet, cujos endereços estão disponíveis nas referências. As obras abrangem um período de publicação de 1977 a 2020, conforme mostra a coluna A do Quadro 2. No próximo parágrafo, fazemos uma breve descrição das obras impressas no Brasil e, posteriormente, descrevemos as obras como um bloco. A análise das obras no Brasil busca

⁵⁰ Disponível em http://www.filologia.org.br/xi_sinefil/completos/bases_PEREIRA.pdf. Acesso em 15 maio 2020.

levantar o perfil das publicações da Linguística nesse país. A segunda parte da análise, das obras em conjunto, foi adotada porque nossa proposta é que o VoBLing seja um vocabulário de Linguística bilíngue, de acesso gratuito na internet. Devido a este perfil, a análise das obras como um todo é mais adequada, já que o VoBLing compartilha das características de todas estas obras, impressas, eletrônicas e on-line.

Em relação às obras impressas de Linguística no Brasil, inicialmente, temos a publicação fundadora de Mattoso Câmara (1986) e depois a obra de Moisés (1995), de acordo com o recorte que fizemos. Dentre os 11 dicionários, 6 são traduções, com ou sem adaptação, o que corresponde a 54% das obras do Brasil analisadas. A peculiaridade reside na obra de Silva (2019) pois é uma compilação, cujo *corpus* engloba os dicionários de gramática, de literatura e de linguística, desde o Dicionário Gramatical de João Ribeiro (1906) até as publicações mais recentes de 2018, duas das quais constituem esta análise: Flores et al. (2018) e Costa (2018). O *corpus* total da obra de Silva (2019) envolve aproximadamente 5000 títulos na sua bibliografia; a obra contém 6869 páginas e objetiva servir de *corpus* para a produção de dicionários de outros pesquisadores. Uma versão preliminar e parcial já foi publicada, intitulada Vocabulário Terminológico Geral de Linguística e Letras, pela Editora Autografia⁵¹. Infelizmente, talvez, o trabalho seja considerado como encerrado, pois o pesquisador nos deixou em junho deste ano, vítima da pandemia da COVID-19.

Em relação às obras totais analisadas, a coluna B registra que 33,3% dos dicionários são traduções para o português, provenientes de obras de língua francesa e inglesa. Quanto ao item C, da impressão, 67% das obras são impressas. Já em relação ao item D, relativo às obras disponibilizadas em formato eletrônico e on-line, temos uma porcentagem de 55,5%. Com relação à coluna E, dos paradigmas definicionais, todos os dicionários apresentam a definição terminológica e 55,5% delas disponibilizam a definição enciclopédica no mesmo PD.

Na análise das obras baseada em Faulstich (2011), o primeiro item é sobre o autor e o reconhecimento devido à publicação de trabalhos anteriores ou à publicação de várias edições. Nossa observação mostra que 50% das obras têm como autores especialistas da Linguística. Em relação à titulação, observamos que 72% dos autores possuem nível de doutorado em Linguística ou áreas correlatas como a literatura, a filologia e o ensino de línguas.

O segundo item de avaliação trata da apresentação da obra pelo autor. Em relação ao primeiro subitem, se o objetivo da obra ficou claro para o consulente, constatamos um resultado positivo em mais de 80% das obras, presentes na apresentação dos dicionários ou em páginas

⁵¹ Obra disponível em <https://www.autografia.com.br/produto/vocabulario-terminologico-geral-de-linguistica-e-letras/>. Acesso em 02 ago. 2020.

virtuais que informavam a motivação do glossário ou vocabulário. O perfil traçado para o público-alvo é: alunos e professores da área de Linguística, leitores da área de Humanas e pesquisadores; dois autores concebem a obra como fonte de pesquisa ou obra de referência. A obra de Silva (2019) e o *site* SIL, com o glossário de termos da linguística (português e francês), explicitam que a obra tem como objetivo ser útil a pesquisadores. Um dado bastante recorrente na introdução das obras é a motivação que leva os autores a produzirem o dicionário: a busca de padronização ou harmonização dos termos e conceitos da linguística e suas subáreas, desafio esse que perdura até os dias de hoje. No que tange à referência e às fontes que os autores utilizaram para produzir as obras, constatamos que a obra de Silva (2015), “Letras” entre aspas, não se adequa à legislação brasileira de direitos autorais e da lei antiplágio. Surpreendentemente, é uma obra recente, publicada no suporte físico e eletrônico, disponibilizada à venda no Brasil e no exterior, em *sites* de renome internacional, e publicada por editora com corpo editorial.

O terceiro item da análise trata da apresentação material da obra. Observamos que nas obras impressas, a maioria dos autores apresenta o dicionário ou ele é apresentado por especialistas de renome da área. Nas obras virtuais esta apresentação quase não acontece, a não ser em projetos mais bem estruturados que justificam o porquê da obra, como foi feita a pesquisa e para qual objetivo deve ser utilizada. Em relação à ordem de apresentação dos verbetes, a maioria dos autores prefere a ordem alfabética à ordem sistemática final. Contudo, o dicionário de Ducrot e Todorov é o exemplo de uma obra impressa, com verbetes organizados sistematicamente, de fácil manuseio, que traz as remissivas, as páginas onde os termos ocorrem e a lista de verbetes. Esta mesma praticidade está disponível em obras on-line, principalmente por meio de *hyperlinks*. Contudo, tanto as obras impressas como as virtuais precisam melhorar neste aspecto, já que nem todas disponibilizam *hyperlinks* para remissivas, como por exemplo a obra *Linguística-Términos*.

A presença de ilustrações foi analisada no item 3.3. Devido ao fato de que as obras são da terminologia da linguística, cujos conceitos são mais abstratos, a presença de imagens foi raríssima. Apesar disso, houve um uso interessante presente no dicionário de Silva (2019), recurso que o pesquisador observara em um dicionário de LP da Civilização Brasileira, quando de sua leitura da infância. O fato é a existência da imagem do alfabeto fenício e do grego que originou o nosso alfabeto Latino, encabeçando a lista de verbetes. O uso de imagens ou vídeos como recurso didático no ensino de línguas foi observado sob este aspecto. A plataforma *Index of French Grammar and Pronunciation Glossary Terms* disponibiliza vídeos como uma das

ferramentas para o ensino aprendizagem de língua francesa, além de textos de leitura e outros recursos utilizados.

No item 3.8 **a obra está editada em suporte informatizado**, observamos a tendência de mercado e a mudança para o paradigma do “novo normal”. Dentre as 18 obras analisadas, seis estão disponíveis em suporte físico e informatizado em formato PDF, para o aplicativo Kindle ou no aplicativo da editora, o que equivale a 33% das obras. Elas são: o *SIL Glossary of Linguistic Terms*⁵² de 2003, o “Letras” entre aspas de 2015, o dicionário da Linguística da Enunciação e o dicionário de Gêneros textuais, ambos de 2018, e o Vocabulário Terminológico Geral de Linguística e Letras, proveniente do projeto de dicionário de Silva (2019). A obra de Silva foi concebida para ser disponibilizada em formato eletrônico, enquanto a versão impressa, com mais de 6000 páginas, seria produzida sob encomenda.

Considerar os dados do item 3.11, sobre a **ampla divulgação** da obra foi desafiador. Devido ao contexto de pandemia da COVID-19 pela qual passamos atualmente, as lojas físicas estão parcialmente fechadas e nem todas elas oferecem o serviço de vendas pela internet ou o serviço de entrega domiciliar. Por conseguinte, utilizamos o parâmetro da venda on-line nas livrarias virtuais. Dessa forma, pudemos identificar as obras comercializadas no Brasil e no exterior pelo *site* das livrarias virtuais. É surpreendente observar a abrangência que as obras podem atingir nessa modalidade de vendas virtuais, como por exemplo, pela Amazon.

O último item 4 é relativo à microestrutura da obra, como são tratados os verbetes e como é construído o PD. A maioria das obras, principalmente as mais recentes, não disponibiliza a categoria gramatical nem o gênero. Os dicionários de cunho mais científico disponibilizam as subáreas da linguística às quais os termos pertencem, como o *Lexicon of linguistics* (1996) e o *SIL Glossary of Linguistic Terms* (2003). A diferença entre ambos é que o *Lexicon of linguistics* classifica a subárea do termo antes da sua definição, enquanto o *SIL Glossary of Linguistic Terms* disponibiliza o recorte de uma árvore de domínio indicando a sua classificação. Os dicionários bilíngues disponibilizam os equivalentes por meio de *hyperlinks* e a pronúncia não é comum constar com as definições. O *Glossary of Grammar Terms* e o *Index of French Grammar and Pronunciation Glossary Terms* disponibilizam a pronúncia em vídeos paralelos ou em outra página direcionada ao ensino e à prática da pronúncia, não especificamente na página do verbete.

Para finalizar essa avaliação e esse item, observamos que a maioria dos dicionários trabalha com o PD terminológico e outros trabalham com o paradigma terminológico iniciando

⁵² Disponível para aquisição em <https://www.sil.org/resources/publications/entry/67412>. Acesso em 03 ago. 2020.

a microestrutura, mas completando-a com o paradigma enciclopédico. Em geral, os glossários disponibilizam somente o termo equivalente, mas, nem sempre, disponibilizam a definição terminológica.

1.9. Algumas considerações.

Diante do resultado das análises apresentadas no item anterior, conseguimos ter uma visão panorâmica de como tem sido traçado o percurso das obras terminológicas na área da Linguística a nível nacional e internacional. Constatamos que o que define o perfil da obra são os objetivos a partir dos quais ela foi concebida ou idealizada, isto é, se as obras apresentam estrutura sucinta, bem elaborada, de perfil acadêmico que referencia o *corpus* e os exemplos que elas disponibilizam. Há também aquelas obras concebidas como uma ferramenta de ensino-aprendizagem alocadas em uma estrutura maior, como a plataforma bilíngue de francês-inglês citada anteriormente.

Estas constatações nos levam a repensar a estrutura do VoBLing, de forma que ela esteja alinhada com o perfil do seu público-alvo, seja de fácil acesso e que disponibilize ao leitor vários recursos em uma mesma plataforma. Isto significa disponibilizar o termo em português, seu equivalente em inglês, as pronúncias, o material pedagógico como vídeo e imagem, e os exemplos de uso baseado em *corpus*.

Enxergar a possibilidade de o VoBLing ser uma plataforma terminológica e pedagógica coloca diante de nós um desafio de trabalho persistente, árduo e contínuo no intuito de atender a um público-alvo local e internacional. Isto implica em considerar que o VoBLing pode ser usado não só para os alunos iniciantes do Curso de Letras, falantes de português que aprendem inglês, mas também o oposto pode ser real, na medida em que a plataforma pode ser usada por um público falante de LI adquirindo a LP como língua estrangeira.

Finalmente, compartilhamos os resultados obtidos com a análise de alguns dicionários de Linguística impressos realizada por Fromm e Yamamoto (2020), na qual constatamos que os dicionários de Linguística disponíveis aos alunos iniciantes do Curso de Letras não atendem à proposta de um vocabulário que comunique os conceitos básicos, de fácil entendimento dos conceitos-chave da Linguística para alunos iniciantes. Neste viés, adequamos a simplificação das definições, a partir de dados dos *corpora*, para aumentar a compreensão e a acessibilidade desse público-alvo aos conceitos do VobLing. Este procedimento será detalhado no capítulo 5.

4. METODOLOGIA

Nesta seção, explicamos que a Linguística de *Corpus* foi selecionada para atender ao rigor teórico-metodológico da Terminologia/Terminografia, observando-se os princípios de **frequência** e **representatividade** terminológica, de natureza probabilística (BERBER SARDINHA, 2004; VIANA; TAGNIN, 2010). Além da LC, adotamos a plataforma *ToGatherUp* (OLIVEIRA, 2019), onde são abrigados os *corpora* de estudo, na nuvem, e a plataforma do VoTec (FROMM, 2007), onde serão preenchidas as fichas terminológicas e os dados disponibilizados para o público, via internet.

A primeira fase da pesquisa foi realizada para quantificar os textos na LP e na LI existentes na internet. Nessa busca pelos textos, via Google, usamos recursos de codificação para especificar: (i) a área da Linguística marcada por asterisco; (ii) os gêneros textuais: teses, dissertações e artigos científicos; e (iii) o formato do texto: PDF. O formato pdf foi escolhido pelo fato de ser o mais usado para publicações acadêmicas (YAMAMOTO, 2015). O resultado pode ser visualizado no Quadro 3.

Quadro 3 - Quantificação de dados na área da Linguística disponíveis na internet (2015).

Gênero	Palavras-chave (PT)	Palavras-chave (inglês)	Resultado (PT) 2015	Resultado (EN) 2015
Teses	Linguística*teses:pdf	<i>Linguistics*theses:pdf</i>	692 mil	1.080 milhões
Dissertações	Linguística*dissertações:pdf	<i>Linguistics*dissertations:pdf</i>	545 mil	1.210 milhões
Artigos científicos	Linguística*artigos científicos:pdf	<i>Linguistics*scientificarticles:pdf</i>	443 mil	2.040 milhões
Artigos acadêmicos	Linguística*artigos acadêmicos:pdf	<i>Linguistics*academicarticles:pdf</i>	173mil	2.870 milhões

Fonte: Elaborado pelo autor (2015).

Quadro 4 - Quantificação de dados na área da Linguística disponíveis na internet: artigos científicos/scholarly articles (2019).

Gênero	Palavras-chave (PT)	Palavras-chave (EN)	Resultado (PT) 2015 (milhão)	Resultado (EN) 2015 (milhão)
Artigos científicos	Linguística*artigos científicos:pdf	<i>Linguistics*scholarlyarticles:pdf</i>	20,3 milhões	50,5 milhões

Fonte: Elaborado pelo autor (2019).

Nos Quadros 3 e 4, os dados foram organizados da seguinte forma:

- (i) Coluna 1 – gêneros textuais: teses, dissertações, artigos científicos e artigos acadêmicos. Foram consideradas as duas expressões devido à existência das

duas terminologias: artigos científicos e acadêmicos. Essa classificação de gênero serviu tanto para a pesquisa em português quanto para a em inglês;

- (ii) Coluna 2 – as palavras-chave usadas para busca na internet em português;
- (iii) Coluna 3 – as palavras-chave usadas para busca na internet em inglês;
- (iv) resultados da busca em português; e
- (v) resultados da busca em inglês.

Analisando os resultados, é possível observar que, na busca por teses e dissertações, o número desses gêneros em LI era, em geral, o dobro da apuração feita em português (1:2) em 2015.

Quanto ao resultado de artigos acadêmicos, o número encontrado é semelhante ao de teses e dissertações, já que a quantidade de obras em LI é maior do que em LP. Em 2015, a proporção de número de artigos disponibilizados pela busca foi de 20% dos artigos em português, em relação às publicações em inglês, sob os termos **artigos científicos** e *scientific articles*. Sob os termos **artigos acadêmicos** e *academic articles*, a proporção foi de 6% de artigos em português em comparação com as publicações em inglês.

Para especificar melhor essas relações, em outubro de 2019, o resultado da pesquisa da busca por **artigos científicos**/*scholarly articles*, em português *versus* em inglês nos trouxe um total de 65,3 milhões de páginas em inglês contra 20,3 milhões de páginas em português. A correspondência seria de 40% de artigos em português em relação aos artigos em inglês (Quadro 4, p. 57).

1.10.Etapas da pesquisa

Este trabalho tem como objeto de estudo as terminologias (ST e STF) presentes nos *corpora* de Linguística. As etapas metodológicas concluídas do trabalho foram: (i) entrevista com o público-alvo para definição do PD mais adequado para atendê-lo; (ii) compilação, (iii) seleção e (iv) estudo/análise de *corpora*; (v) alimentação de dados na plataforma do *ToGatherUp* e VoBLing; e a (vi) escritura da tese.

Nosso ponto de partida foi realizar uma pesquisa que buscou definir o padrão de PD a ser adotado no VoBLing, baseada no VoTec (FROMM, 2007), disponível em <http://treino.votec.ileel.ufu.br/>.

As inovações que estamos efetivando em relação ao produto final, ou seja, ao VoBLing, são a disponibilização de (i) uma janela *pop-up* com a definição das remissivas; (ii) arquivos

de áudio com a pronúncia dos termos em português e em inglês; (iii) etimologia do termo; (iv) vídeo com explicação da subárea da Linguística.

Nesta pesquisa, disponibilizamos os padrões definicionais (PD) por (1) **implicação**, (2) **denotação**, (3) **compreensão**, (4) **sinonímico** e (5) **enciclopédico**⁵³, via *Google Docs*, para que o público-alvo, discentes do curso de Letras, pudesse escolher o que melhor atendia ao objetivo de conceituar e definir o vocabulário de especialidade da Linguística com clareza e objetividade.

A entrevista foi conduzida por meio de um questionário do tipo estruturado, disponibilizado impresso e on-line aos discentes, e as respostas dadas serviram de subsídio para que pudéssemos definir o melhor PD a ser adotado na pesquisa.

1.1.16.Contextualização

Nesta seção, objetivamos apresentar o perfil das universidades nas quais encontramos nosso público-alvo: discentes dos Cursos de Letras da Universidade Federal de Uberlândia e da Universidade Federal de Jataí, previamente chamada de Regional Jataí da Universidade Federal de Goiás, quando do início desta pesquisa.

1.1.17.Perfil do público-alvo

Considerando a especificidade da Linguística como área da Ciência, acreditamos que uma obra terminológica direcionada a discentes iniciantes no curso universitário de Letras viria a suprir a necessidade de conceituação do vocabulário de especialidade. Dessa forma, tanto o público acadêmico quanto os profissionais da área teriam maior acesso ao conteúdo produzido na área de Linguística em nível nacional. Isto é, nosso público-alvo abrangeria discentes e professores da área de Letras como usuários principais.

Os discentes de Letras seriam os **usuários primários** ou **diretos** que buscam dominar os conceitos relativos à Linguística. Além dos discentes de Letras, os profissionais da área de Tradução que trabalham com obras em LP e LI, ou mesmo estudantes desse curso, fariam parte do grupo de usuários primários.

Os professores, em especial, pesquisadores e aqueles que desenvolvem trabalhos interinstitucionais com universidades do exterior, seriam **usuários secundários**, os quais

⁵³ PD por implicação: o termo é definido a partir de um contexto explicativo; por denotação: definido por meio de exemplos; analítico: definido a partir de hiperonímia para hiponímia; por sinonímia: o termo é definido por meio de paráfrases; e enciclopédico: definição com dados abrangentes sobre o termo (SAGER, 1993).

poderiam se beneficiar do VoBLing para a tradução e leitura de obra escritas em inglês. Professores que recorrem à bibliografia proveniente do exterior para ministrar seus cursos, desenvolver pesquisas e publicar em eventos nacionais e internacionais também seriam usuários dessa obra. Salientamos que, para profissionais com este perfil e para os discentes de graduação, o uso desse protótipo de vocabulário poderá ser de grande valia, principalmente na redação de resumos/*abstracts* necessários para a participação em eventos nacionais e internacionais.

Finalmente, outro perfil de usuários, **os terciários**, seriam os demais usuários que utilizam a internet. Eles também poderiam usufruir de tal protótipo de vocabulário, pois o acesso é livre a todos os públicos.

1.1.18. Realização da pesquisa

A pesquisa foi realizada no período de 07 a 27 de maio de 2019 com os discentes do curso de Letras da UFU e da UFJ. Obtivemos um total de 110 respondentes, sendo 69 da UFU e 41 da UFJ. As pesquisas foram conduzidas em sala de aula, de forma presencial, o questionário de pesquisa para escolha do paradigma definicional foi disponibilizado por meio de acesso on-line ao formulário e por meio de formulário impresso, apresentado no Apêndice A.

O pesquisador contatou os professores das disciplinas do curso de Letras de ambas instituições para que as visitas e entrevistas fossem permitidas em sala de aula. No dia combinado, o pesquisador dirigiu-se às salas de aula e iniciou a coleta de dados para a pesquisa por meio de um questionário virtual e impresso. O pesquisador explicou o Termo de Compromisso Livre e Esclarecido – TCLE – e, em seguida, disponibilizou o questionário para ser respondido pelos discentes.

Na UFG/REJ foram entrevistados discentes do curso de Letras com habilitação em Inglês e em Português. Os entrevistados totalizaram 41 discentes que responderam ao questionário, via formulário impresso, pois se encontravam em sala de aula sem acesso à internet e sem computadores.

Na UFU foram entrevistados discentes do curso de Letras com habilitação em Inglês, Francês, Espanhol e Português. Do total de 69 discentes, 17 responderam ao questionário on-line no laboratório do curso de graduação; os 52 restantes responderam o questionário via formulário impresso, pois se encontravam em salas de aula sem acesso à internet e sem computadores.

Originalmente, o questionário foi concebido para que os discentes respondessem às perguntas on-line, via formulário do *Google Docs* (*Gdocs*). Contudo, quando não foi possível fazer a pesquisa on-line, utilizamos os formulários impressos. Respondidos os questionários, o pesquisador transpôs os dados para o formulário do *Gdocs* para serem tabulados de forma mais rápida e prática.

1.1.19.Resultado da pesquisa

No questionário de pesquisa, os discentes escolheram o paradigma definicional que, para eles, melhor definia os seguintes termos da Linguística: (A) hápax legomena, (B) traços distintivos, (C) léxico, (D) topônimo, e (E) estilística. Foram disponibilizados os seguintes paradigmas definicionais⁵⁴ baseados em Sager, 1993 (apud CARDOSO, 2017): (1) definição por implicação; (2) definição por denotação; (3) definição por compreensão; (4) definição por sinonímia; e (5) definição enciclopédica. Os termos e suas definições estão disponíveis no Apêndice A. Para exemplificar esses paradigmas, trazemos as definições do termo hápax legomena.

A. HAPAX LEGOMENA

1. () **Hapax legomena** – Sabemos que algumas das palavras correm o risco de terem sido empregadas uma só vez e por uma só pessoa. Contudo, optamos por apresentá-las com sua frequência única - as hápax legomena.
2. () **Hapax legomena** – Somente uma variante foi encontrada na busca: prepina. Tal variante foi encontrada apenas em um contexto, se constituindo, dessa forma, hápax legomena.
3. () **Hapax legomena** – Dentre as palavras de uma obra escrita, literária ou não, é aquela que aparece uma única vez.
4. () **Hapax legomena** – Neologismo genuíno.
5. () **Hapax legomena** – Do grego hápax, "uma única vez", e legomenon, "dito", considerada a ocorrência única de uma palavra em uma obra escrita; concebido como neologismo, a partir da perspectiva da produtividade morfológica, ou formações analógicas como em "enxadachim" (Guimarães Rosa), e bebermar e trêbados, comuns atualmente em contextos informais de comunicação.

Como resultado da pesquisa, os respondentes escolheram a **definição enciclopédica**, com maior média percentual: 41,49%, e a **definição por compreensão**, como segunda opção, com 33,46% (Veja Apêndice B para dados em gráfico tipo pizza).

Os dados numéricos detalhados seguem abaixo, na Tabela 1, com o detalhamento das porcentagens. As colunas trazem as opções de um a cinco, representando os padrões de

⁵⁴ As definições dos paradigmas definicionais são: 1. uso do termo em um contexto explicativo; 2. relaciona exemplos; 3. termo genérico e características específicas; 4. O mesmo significado; e 5. descrição exaustiva do termo nomeado (CARDOSO, 2017 apud SAGER, 1993).

paradigmas definicionais. As linhas representam os termos que foram definidos, de A a E, sendo A: hápax legomena, B: traços distintivos, C: léxico, D: topônimo e E: Estilística.

Tabela 1 - Resultado da pesquisa dos paradigmas definicionais com discentes da UFU e da UFJ

	Implicação	Denotação	Compreensão	Sinonímia	Enciclopédica	TOTAL
A	6,4%	7,3%	20,9%	3,6%	61,8%	100%
B	9,1%	28,2%	16,4%	5,4%	40,9%	100%
C	5,4%	20%	56,4%	0,9%	17,3%	100%
D	6,4%	5,5%	19,1%	4,5%	64,5%	100%
E	10,9%	8,3%	54,5%	3,6%	22,7%	100%
Média	7,64%	13,86%	33,46%	3,6%	41,4%	100%

Fonte: Elaborado pelo autor com base no questionário respondido pelos discentes da UFU e da UFJ.

De acordo com os dados expostos na Tabela 1, podemos observar que para a **definição por implicação** obtivemos a média de 7,64%; para a **definição por denotação**, obtivemos a média de 13,86%; para a **definição por compreensão**, obtivemos a média de 33,46%; para a **definição por sinonímia**, obtivemos a média de 3,6%; e, por fim, para a **definição enciclopédica**, a média foi de 41,4%. Diante desses dados, entendemos que o **PD enciclopédico** foi o mais escolhido pelos discentes e, em segundo lugar, o **PD terminológico por compreensão**.

Para atender aos dois padrões, construiremos a **definição terminológica por compreensão** logo após a entrada do termo, o que é comum para trabalhos terminográficos. Em seguida, disponibilizaremos a **definição enciclopédica** de duas formas: (i) no campo **Notas**, quando o *corpus* disponibilizar dados suficientes para tal e (ii) através da definição da *Wikipedia*, ou de outro *site* disponível na internet, como parte estrutural da plataforma do VoBLing. Flores et al. (2009, p. 28) salientam a importância do campo **Notas** em um trabalho terminográfico: “notas explicativas são muito importantes e devem situar a definição de um conceito em meio a outro conceito...”. Em nossa pesquisa, este campo disponibiliza as informações enciclopédicas.

1.11.Elaboração da árvore de domínio da Linguística

A organização das subáreas da Linguística, elencadas na **árvore de domínio** dessa área, foi outro procedimento metodológico. Ela é fruto do trabalho desenvolvido pelo professor Fromm nos cursos de Letras e Tradução (FROMM; YAMAMOTO, 2013). Nesta tese,

propomos, no segundo nível da árvore, a nomenclatura Linguística Descritiva⁵⁵, que contrasta com a Linguística Aplicada. Além desta proposta, incluímos a **Biolinguística** e a **Linguística Diacrônica** como subáreas da Linguística Descritiva. É importante lembrar que a abordagem adotada nesta pesquisa é direcionada pelo *corpus*, ou seja, a inclusão dessas duas subáreas se deve ao fato de encontrarmos um *corpus* acima de 500 mil itens para cada uma dessas áreas em inglês e em português.

A **Figura 2** mostra a **árvore de domínio** da Linguística a partir da proposta de Fromm (2015; Anexo C) e por nós modificada em 2020.

⁵⁵ A Linguística Descritiva que nomeamos neste trabalho não deve ser confundida com a corrente linguística existente dos anos 1920 aos anos 1950. A expressão Linguística Descritiva é usada por nós para designar o que alguns teóricos chamam de Linguística Teórica, de caráter sincrônico, seguindo o perfil de registro e análise das línguas, precursora da LA (COSTA, 2011).

Figura 2 - Árvore de domínio da Linguística alterada com base em Fromm (2015)



Fonte: Adaptada de Fromm (2015).

De acordo com a **Figura 2**, podemos visualizar as subáreas da Linguística Descritiva que foram analisadas neste trabalho. O que difere a árvore da Linguística atual da árvore de Linguística de Fromm (2018 – Anexo D) é o acréscimo de duas subáreas da Linguística Descritiva, a saber, a **Biolinguística** e a **Linguística Diacrônica**. Quanto à subárea da LA, não houve proposta de acréscimo de subáreas nem de nomenclaturas.

1.12. Compilação e tratamento do corpus

Nesta seção, trabalhamos os aspectos metodológicos de compilação e tratamento do *corpus* de Linguística. Quanto ao dimensionamento dos *corpora*, para cada subárea da Linguística, trabalhamos com o padrão de 500 mil itens em cada língua: português e inglês (FROMM, 2013, p. 1; FROMM; YAMAMOTO, 2013; FROMM, 2018). No total, obtivemos 1 milhão de itens por subárea, em inglês e em português. Este número é proveniente dos estudos desenvolvidos por Fromm no âmbito das pesquisas terminográficas em contexto de ensino superior de graduação e pós-graduação (FROMM, 2013; FROMM; YAMAMOTO, 2013; FROMM, 2018).

Este padrão foi atingido, à exceção da subárea Linguística Matemática (LM), para a qual foi possível compilar somente 220.245 itens em português, apesar dos 507.984 itens compilados em inglês. Apesar das buscas em bancos de dissertações e de teses e em plataformas como *Scielo* e Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), não atingimos os quinhentos mil itens para essa subárea.

Levando em consideração o fato de que há 29 subáreas na LD e 18 subáreas na LA, teríamos um total aproximado de 47 milhões de palavras compondo os *corpora* de teses, dissertações, artigos científicos e material instrucional. Considerando o exposto no parágrafo anterior sobre a LM, reduzimos o *corpus* de LM, em inglês, para o número de 220 mil itens. A redução do tamanho do *corpus* dessa subárea em inglês objetiva o equilíbrio desses *corpora*. Esses dados são baseados no planejamento feito para o dimensionamento e balanceamento dos *corpora*. Por conseguinte, há 46,4 milhões de palavras em português e em inglês. Adiante, mostraremos o tamanho exato dos *corpora* lidos e quantificados pelo WST na Tabela 2.

Já o dimensionamento médio dos manuais é de 1,09 milhão de itens, em cada língua, totalizando 2,18 milhões de itens nas duas línguas. A inclusão do gênero manuais foi proveniente de um melhoramento da pesquisa de mestrado. Naquela pesquisa, constatamos que o uso de artigos científicos, de teses e de dissertações de mestrado não trazia contextos definitórios e explicativos suficientes para definição de candidatos a termos. Portanto, recebemos a sugestão de incluir em nossa base de dados os manuais de linguística, nos quais haveria uma probabilidade maior de recorrência desses contextos definitórios, pois esse é um gênero que visa à comunicação de especialistas para leigos. Nesse contexto, o parâmetro de tamanho foi o maior possível encontrado no português. A partir desse tamanho, coletamos um *corpus* de manuais em inglês com tamanho igual para equilibrar as duas línguas.

Portanto, o tamanho final dos *corpora* é de aproximadamente 48,6 milhões de itens, inclusos os gêneros teses, dissertações, artigos científicos, manuais e material instrucional, de acordo com a projeção feita para esta pesquisa. Esse número representa o tamanho final

aproximado com o *corpus* de LM em LI reduzido, a fim de equilibrá-lo com o *corpus* de 238.082 itens dessa subárea em LP.

A quantificação de itens provenientes dos manuais foi em torno de 1,3 milhões de palavras em português e a mesma quantidade em inglês. Nesse caso, a LP serviu de base para quantificarmos o número de palavras, já que a quantidade de manuais em português é menor que a quantidade de manuais em LI. Finalmente, como parte do processo de dimensionamento final dos *corpora*, nós chegamos a um total de 49,89 milhões de palavras, de acordo com a leitura realizada pelo WST.

Nas Tabelas 2 e 3, são relacionados os dados específicos do dimensionamento dos *corpora* de Linguística. Na Tabela 2, apresentamos os números provenientes do dimensionamento dos manuais de Linguística em LP e LI, aos quais aplicamos as *stoplists*⁵⁶ em LP e LI, antes de fazer as listas de palavras. As *stoplists* utilizadas neste trabalho não incluem substantivos, classe gramatical primordial à nossa pesquisa terminológica. A *stoplist* de LP contém 995 palavras gramaticais, incluindo artigos, numerais, preposições e advérbios, cujo dimensionamento é de 9KB. Em contrapartida, a *stoplist* de LI contém 571 palavras, e além das classes de palavras de da *stoplist* em LP inclui alguns verbos e tem 5KB de dimensionamento.

Tabela 2 - Dimensionamento do <i>corpus</i> de manuais em LP e em LI.		
	Português	Inglês
Número de textos	13	5
Tokens	1.089.748	1.099.776
Types	44.407	35.733
Type/token ratio	4,07%	3,25%
(TTR)		
TOTAL tokens	2.189.524	

Fonte: Elaborada pelo autor.

A Tabela 2 apresenta os dados quantitativos referentes ao *corpus* de manuais de Linguística. Os manuais em LP são em maior quantidade, sendo um deles tradução do manual de Lyons (1979) em LI. Em inglês, a quantidade é menor, pois os manuais são maiores que em LP.

⁵⁶ A *stoplist* ou *stopwords* consiste em uma lista de palavras utilizadas pelo WST ao processar termos. O WST ignora essas palavras ao gerar listas porque elas não estão relacionadas ao objetivo do estudo, logo, evitam o "ruído" ou palavras que não estão relacionadas ao estudo (AUBERT, 2001; SCOTT, 2020). As *stoplists* aplicadas no *corpus* de Linguística são provenientes do grupo de pesquisa do GPELC.

Na Tabela 3 apresentamos o dimensionamento total dos *corpora* em LP e LI. Nesses *corpora* encontram-se os manuais e os artigos científicos, subdivididos em artigos, dissertações e teses.

Tabela 3 - Dimensionamento dos <i>corpora</i> em LP e em LI.		
	Português	Inglês
Número de textos	1665	1911
Tokens	24.597.712	25.294.698
Types	356.983	356.911
Type/token ratio (TTR)	1,45%	1,41%
TOTAL tokens	49.892.410	

Fonte: Elaborada pelo autor.

A Tabela 3 especifica o dimensionamento de cada *corpora* em LP e LI. O número de *tokens* corresponde aos itens presentes no *corpus* de Linguística Descritiva, no *corpus* de Linguística Aplicada e no *corpus* de manuais de língua de Linguística em LP; o mesmo padrão se aplica para a LI. O tamanho total dos *corpora* em LP e em LI é de 49.892.410 *tokens* ou itens.

Como a metodologia adotada neste trabalho é a de *corpus driven* e de *corpus based*, os *corpora* são direcionados pela árvore de domínio e vice-versa, o que implica numa correspondência biunívoca entre árvore de domínio e os *corpora*. Foi por esta razão que optamos pela inserção das subáreas **Biolinguística** e **Linguística Diacrônica**. Quantitativamente, por exemplo, houve a necessidade de incluir a **Linguística Diacrônica** na árvore, pois, na internet, esse *corpus* é maior do que o de Etimologia, de LH e da Filologia (Apêndice C). Quanto ao percurso de elaboração da árvore de domínios da Linguística, as versões se encontram nos anexos do A ao D.

1.13.Extração de dados do corpus

Esta subseção trata da extração e tratamento dos dados, provenientes do *corpus*, processados pelo WST. Isso significa que a análise de *corpora* específicos permite identificar os candidatos a termos, por meio das listas de palavras, de palavras-chave. Em seguida, o concordanciador nos permite identificar os contextos definitórios e explicativos onde os candidatos a termos ocorrem. Finalmente, a partir da análise qualitativa, extraímos os traços distintivos para a construção da definição terminológica e enciclopédica. É importante

considerar que a designação em Terminologia, conforme a Norma ISO 1087, pode ser realizada por meio de símbolos, nomes ou termos⁵⁷. Nesta pesquisa, nosso foco reside nos termos.

Além do recurso da análise componencial e da sistematização dos dados presentes nas fichas terminológicas do VoBLing, buscamos aplicar os princípios propostos por Fillmore (1975; 1976; 2006) e Ilari (2003) e Souza (2019) para a organização dos traços distintivos nos PD dos verbetes do VoBLing.

1.14.Elaboração dos paradigmas definicionais

O paradigma definicional (PD) faz parte da microestrutura [de vocabulários](#) e descreve os traços distintivos dos ST. Nele estão contidos os conceitos organizados por ordem de maior recorrência; a partir dele, elaboramos um conceito final, utilizado para a construção da definição.

Inicialmente, partimos da análise componencial de Ilari (2003) e do padrão GPDE, (gênero próximo, diferença específica), (PAVEL; NOLET, 2002, p. 25, 27, 122), típico da definição terminológica, para elaborar as definições do vocabulário de Linguística. A plataforma do VoTec (FROMM, 2007) permite que os traços distintivos mais recorrentes sejam sistematizados, o que facilita a identificação desses traços utilizados na construção da definição. Sem embargo, conforme afirma Maciel (2001), a Terminologia tradicional

pressupõe sistemas de conceitos delineados com precisão e denominados univocamente. No entanto, no campo das ciências humanas, tais contornos não são tão nítidos, quanto nos outros domínios [...] De fato, dadas as características próprias dessas áreas, os traços que delimitam a especificidade de seus conceitos se tornam fluídos e dificilmente identificáveis. (MACIEL, 2001, p. 40).

Conscientes dessa realidade, embarcamos no desafio de definir o vocabulário parcial da Linguística. Acreditamos que os dados disponibilizados por meio dos *corpora* e da metodologia da LC e do VoTec (FROMM, 2007) podem trazer subsídios para uma nova metodologia de construção de definição terminológica na área de Ciências Humanas. Ademais, apoiamos nossa prática nos parâmetros da ISO 704 (2000) e na Semântica de *Frames*, de Fillmore (1975; 1976; 2006) e na Terminologia de *Frames* de Faber Benítez et al. (2005). Ademais, seguimos os

⁵⁷ “In terminology work (3.6.1) three types of designations are distinguished: symbols, appellations (3.4.2) and terms (3.4.3)”. Nessa norma, o termo *appellation* é definido como “name, verbal designation (3.4.1) of an individual concept (3.2.2)” (ISO 1087:2000, p. 6), isto é, nome, designação verbal de um conceito individual.

passos propostos pela norma ISO 704 (2000, p. 3), sobre O trabalho terminológico: princípios e métodos, quanto à análise terminológica e associação de conceitos às definições:

Como uma designação não é atribuída a cada objeto individual, a **análise terminológica** não pode iniciar a não ser que o objeto específico em questão corresponda ao conceito representado por meio da **designação** ou **definição**. Portanto, a metodologia usada na análise de terminologias demanda a identificação do contexto ou da área de estudo almejada, identificando propriedades atribuídas aos objetos na área de estudo, determinando essas propriedades abstraídas em características e, então, combinando as características para formar o conceito [...]. (ISO 704: 2000, p. 3.)⁵⁸

Os procedimentos de análise terminológica propostos acima e a aplicação da metodologia da LC fundamentam e sustentam nossa mudança de paradigma da prática terminográfica: deixando o percurso onomasiológico, típico da Terminografia, para adotar o percurso semasiológico, típico da Lexicografia.

O paradigma definicional que construímos atende a dois tipos de definições: a definição terminológica e a definição enciclopédica, ambas escolhidas pelo público-alvo da pesquisa. A primeira está disposta logo após o verbete, seguindo o padrão GPDE, e a segunda está no espaço de notas, seguindo o padrão enciclopédico. Definimos duzentos termos, em português e inglês, da área da Linguística.

Para a construção dos padrões terminológico e enciclopédico, consideramos os traços distintivos mais recorrentes, provenientes das colunas da ficha terminológica do VoBLing, adequando-os ao padrão GPDE na definição terminológica. A definição enciclopédica foi construída a partir de uma segunda possível definição da área ou subárea. Buscamos prover o máximo de informação, de forma organizada, a partir dos dados provenientes dos excertos dos *corpora*.

Em consequência disso, a definição enciclopédica foi mais variável em termos de padrões. Essa variabilidade se justifica pela diversidade de informações disponíveis nos *corpora*, pois abarca correntes, métodos, perspectivas e abordagens diferentes para o estudo do signo e do sistema linguístico, dentro de cada subárea da Linguística. Por exemplo, ao se definir *linguística*, as abordagens imanentista e sócio-histórica são citadas como possibilidades de

⁵⁸ Since a designation is not attributed to every individual object, terminological cannot begin unless the specific object in question corresponds to a concept represented by means of a designation or a definition. Therefore, the methodology used in the analysis of terminologies requires identifying the context or subject field in question, identifying the properties attributed to objects in the subject field, determining those properties which are abstracted into characteristics and then combining the characteristics to form a concept. It may be useful to begin an analysis with those concepts corresponding to concrete objects, since the characteristics are more easily abstracted given that the properties of the objects can be physically observed or examined. (ISO 704:2000, p. 3).

estudo, dependendo da concepção do signo e do sistema linguístico, ou de como a sociedade e a língua se inter-relacionam num dado momento da história.

1.1.20. Proposta de PD baseado em *Frames*

A proposta apresentada adiante é proveniente da organização dos traços distintivos dos verbetes, conforme a teoria da Terminologia de *Frames*, a metodologia do VoTec (FROMM, 2007) e a proposta de Terminografia Pedagógica (FROMM, 2020). Baseados nesses princípios, os traços distintivos mais frequentes foram organizados para compor o PD de cunho terminológico, enquanto os traços distintivos menos frequentes foram organizados para compor o PD de cunho enciclopédico no campo de **Notas** do VoBLing. A terminografia pedagógica de Fromm (2020) fundamenta-se no uso e domínio da estrutura do VoTec (FROMM, 2007) como plataforma terminográfica adequada ao ensino. Outro fator desejável são as caixas de texto que facilitam o acesso do consulente às definições das remissivas e dos termos já cadastrados na plataforma por meio dos *hiperlinks* (FROMM, 2020, p. 772-773.). Já o VoBLing busca tornar a terminologia mais acessível por meio das paráfrases dos conceitos dos novos termos (veja os *ST linguística e teletandem*) (FINATTO, 2020, p. 81).

1.1.1.3 PD terminológico baseado em *Frames*

Primeiramente, no Quadro 5, trazemos a **Proposta 1** de organização dos traços distintivos, extraídos a partir de contextos definitórios do ST **linguística/ linguistics**, provenientes do *corpus* de Linguística. Este PD terminológico se aplica ao primeiro nível da nossa árvore de domínio da Linguística, de padrão GPDE. A fim de ser mais específico, neste passo metodológico, os traços distintivos serão chamados de **descritores**.

Quadro 5 - Proposta 1 - PD terminológico baseado em *Frames*: *linguística/linguistics*.

Nome da área /subárea	1. Descritor 1 - Ciência/disciplina	2. Descritor 2 -O que estuda/faz?	3. Descritor 3 - Método/metodologia	4. Descritor 4 -Para que estuda?	5. Descritor 5 - Perspectiva
linguística	ciência da linguagem	estuda o signo linguístico e a língua, seu funcionamento e sua estrutura	metodologia específica de investigação		pode ser sincrônica ou diacrônica
linguistics	scientific field	studies languages in general as human communication	using particular methods of description	contributing to the general history of ideas	

Fonte: Elaborado pelo autor.

O Quadro 5 apresenta a proposta de um *frame* para um PD terminológico utilizado para a área da Linguística. É importante considerar que os descritores são dependentes e provenientes do *corpus* de estudo, ou seja, nem sempre é possível completar todos os espaços da estrutura de *frames* proposta. Além da metodologia de *Frames*, conjugamos a proposta metodológica do VoBLing (maior recorrência dos traços distintivos organizados em coluna na ficha terminográfica) para a organização dos descritores explicada adiante.

Na primeira coluna, especificamos o nome da subárea a ser definida, informação que, em geral, faz parte da macroestrutura da obra terminográfica. O **descriptor 1** é o hiperônimo que serve como cabeçalho para o início da definição do ST *linguística* e a define como ciência da linguagem ou *scientific field*; o **descriptor 2** trata do objeto de estudo da área ou subárea. No caso da *linguística*, em português, temos o signo linguístico como o objeto de estudo ou *languages in general as human communication*, em inglês. O **descriptor 3** informa sobre o método ou metodologia que se aplica àquela ciência: em português, a *linguística* tem uma metodologia específica de investigação; já em inglês, ela usa *particular methods of description*. O **descriptor 4** informa para que aquela ciência estuda seu objeto de estudo, o que em inglês é *contributing to the general history of ideas*. Por fim, o **descriptor 5** trata da perspectiva de estudo daquela área ou subárea, neste caso, a *linguística* pode estudar o signo linguístico de uma perspectiva sincrônica ou diacrônica.

Por conseguinte, considerando o PD do ST **linguística/linguistics**, sintetizamos a seguinte estrutura: (i) a **definição** do ST, (ii) o **objeto** de estudo, (iii) a **metodologia**, (iv) a **razão** do estudo e (v) a **perspectiva** do estudo. Com base na descrição do ST apresentada, observamos que **linguística** é (i) a ciência da linguagem (definição base); (ii) que estuda o signo linguístico e a língua; (iii) cuja metodologia pode ser sincrônica ou diacrônica; (iv) para contribuir com a história das ideias (razão - em inglês); (v) de perspectiva sincrônica ou diacrônica.

Em segundo lugar, apresentamos a Proposta 2, para o segundo nível da árvore de domínio da Linguística, o PD aplicado à LA com quatro descritores conforme mostra o Quadro 6.

Quadro 6 - Proposta 2 - PD terminológico baseado em *Frames: linguística aplicada/applied linguistics*.

Nome da subárea	1. Descritor 1 - Subárea da L/LD/LA	2. Descritor 2 - O que estuda/faz?	3. Descritor 3 - Modo (Como?)	4. Descritor 4 - Para que estuda?
linguística aplicada	subárea da linguística	estuda os problemas de uso da linguagem humana	por meio de teorias da linguística descritiva e de outras áreas do conhecimento	resolver problemas de comunicação da vida cotidiana e profissional

Applied linguistics	branch of Linguistics	studies real-world language-based problems	developing theoretical and descriptive interdisciplinary research	to solve psychological, pedagogical, social, political, economic, and linguistic issues
---------------------	-----------------------	--	---	---

Fonte: Elaborado pelo autor.

Semelhantemente à Proposta 1, a primeira coluna do Quadro 6 identifica o nome da subárea a ser definida, a LA. O **descriptor 1** introduz o ST LA como subárea da Linguística em LP e como *branch of Linguistics* em LI. O **descriptor 2** define o objeto de estudo da subárea: os problemas de uso da linguagem humana em LP e *real-world language-based problems* em LI. O **descriptor 3**, diferentemente da Proposta 1, especifica o **modo** como são conduzidos os estudos referentes aos problemas de uso da linguagem humana: por meio de teorias da linguística descritiva e de outras áreas do conhecimento em LP e *developing theoretical and descriptive interdisciplinary research* em LI. O **descriptor 4** justifica a razão dos estudos do **descriptor 2**: resolver problemas de comunicação da vida cotidiana e profissional em LP e *to solve psychological, pedagogical, social, political, economic, and linguistic issues* em LI.

Comparando os PD da LA em LP e LI, observamos que há uma representação de mundo distinta, expressa nos traços distintivos empregados pelos falantes dessas línguas. No **descriptor 2**, a ênfase reside em problemas do mundo real em LI, ou seja, o locativo é mais evidente, relacionando a ação ao local em que ela ocorre. Este traço distintivo ocorre uma única vez nos contextos provenientes do *corpus* em LP, e esta ocorrência remete ao excerto de uma tradução de LI para LP: “tomada de decisões no mundo real (p. 1, tradução nossa)” (excerto do VoBling do STF LA). Quanto ao **descriptor 4**, identificamos a visão mais generalizada de falantes de LP (vida cotidiana e profissional), em contraponto com a especificação dos problemas em LI (*psychological, pedagogical, social, political, economic, and linguistic issues*). A recorrência destes traços distintivos confirma a teoria de *Frames* (FILLMORE, 1975) ao evidenciar a associação das cenas aos quadros linguísticos (escolhas lexicais) dos falantes de LP ou LI.

Apesar das diferenças dos traços distintivos utilizados para definir a LA em LP e em LI sejam divergentes, eles foram mais uniformes do que aqueles utilizados para definir a LD (Figura 47). Por esta razão, adotamos o PD da LA como modelo, a fim de aplicá-lo às subáreas da Linguística.

Finalmente, apresentamos a Proposta 3, aplicável ao terceiro nível da árvore de domínio da Linguística, o nível dos ST subordinados às LD e LA. Para exemplificar este paradigma, trazemos o ST *teletandem /tandem language learning*, com 5 descritores, no Quadro 7.

Quadro 7 - Proposta 3 - PD terminológico baseado em *Frames*: *teletandem/tandem language learning*.

ST	1. Descritor 1 (hiperônimo)	2. Descritor 2 (propósito)	3. Descritor 3 (maneira)	4. Descritor 4 (base)	5. Descritor 5 (Como? Para quê?)
<i>teletandem</i>	ambiente virtual	de ensino e aprendizagem de LE	com encontro virtuais síncronos e de breve duração	baseado nos princípios fundamentais de alternância de papéis e de autonomia das parcerias fixas	
<i>tandem language learning</i>	platform	designed to promote learners' [...] on-line tandem language learning	face-to-face, in pairs	based on two main principles of seeking and providing assistance to each other (reciprocity) and (autonomy)	to help their partner achieve learning each other's mother tongue

Fonte: Elaborado pelo autor.

De acordo com o Quadro 7, a coluna à esquerda identifica o ST *teletandem /tandem language learning*. O **descritor 1** é o hiperônimo *ambiente virtual* em LP e *platform* em LI. O **descritor 2** identifica o **propósito** referente ao descritor 1, ou seja, o ambiente virtual ou a plataforma serve a qual propósito? Neste caso, o *teletandem/ tandem language learning* serve ao propósito de ensino e aprendizagem de LE. O **descritor 3** informa a **maneira** como o *teletandem* se realiza *com encontro virtuais síncronos e de breve duração/ face-to-face, in pairs*. O **descritor 4** em forma identifica a **base** sobre a qual o *teletandem* também é construído ou se desenvolve: a autonomia e a reciprocidade. Esses 2 princípios são explicados dentro do PD por meio de paráfrase, sendo a primeira alternância de papéis ou *seeking and providing assistance to each other*. O **descritor 5** responde a 2 perguntas como? E por quê? No caso do *teletandem*, a maneira como o ensino-aprendizagem de língua estrangeira se faz é quando os alunos fazem uso simultâneo da produção e compreensão oral, escrita e de imagens. Em LI, os alunos interagem com o objetivo de *help their partner achieve learning each other's mother tongue*.

Chamamos a atenção para como o ST autonomia é definido em LI como *self-instruction*/auto aprendizado e em LP é definido no PD enciclopédico, da seguinte forma na Nota: “permite que os dois decidam o que, quando, onde e como aprender (autonomia)”. É visível que o conceito de autonomia no *corpus* de LI é diferente do conceito de autonomia no de LP. Em LI, há o viés de autodidatismo (self), enquanto em LP, a não interferência de um professor no período de *chat* do *teletandem* é evidente.

Estas três propostas de PD, apresentadas dos Quadro 5 ao 7, servem como uma proposta para definir os ST da Linguística nos três níveis da árvore de domínio. Obviamente, há momentos em que os descritores disponibilizados pelo *corpus* são distintos daqueles necessários ao preenchimento destes PD. Quando isto ocorre, é necessário adotar outros padrões, os quais denominamos **anomalia**, assunto sobre o qual discorreremos na seção de resultados desta tese. A diferença no uso dos traços distintivos é justificada pela Terminologia de *Frames* sobre como cada cultura/língua interpreta os traços distintivos de um mesmo conceito ou objeto de formas diferentes. A LC permite categorizar e organizar estes traços distintivos e evidenciar a aplicabilidade daquela teoria no contexto deste trabalho.

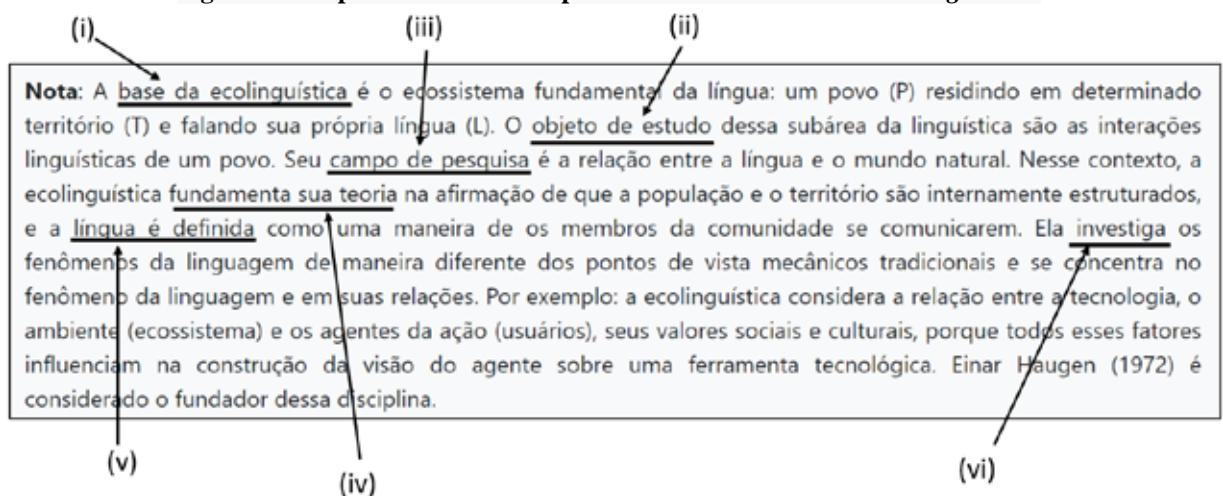
1.1.1.4 PD enciclopédico baseado em Frames

Nesta subseção, descrevemos as propostas de PD enciclopédicos baseados em *Frames*, aplicados à estrutura de **Notas** do VoBLing. Para tanto, usaremos os verbetes *ecolinguística*, *língua*, *biolinguística* e *análise da conversação* como exemplos.

No primeiro padrão, os descritores abordam vários aspectos da *ecolinguística* a partir de uma perspectiva científico-acadêmica. O segundo padrão disponibiliza os descritores que definem *língua* na Linguística, em forma de tópicos. O terceiro padrão é um PD que estabelece uma interlocução com o consulente por meio de perguntas e enumerações que trazem os descritores sobre a *biolinguística*. O último padrão é um PD que traz uma breve cronologia da *análise da conversação*, bem como os descritores presentes nos PD descritos adiante.

A primeira proposta é referente ao ST ecolinguística, descrita na Figura 3.

Figura 3 - Proposta 1 - PD enciclopédico baseado em Frames: ecolinguística.



Fonte: Elaborada pelo autor.

Na **Proposta 1** do PD enciclopédico contida na Figura 3, há seis tipos de descritores: (i) a **base** que sustenta a *ecolinguística*, o ecossistema; (ii) o **objeto de estudo**, as interações linguísticas de um povo; (iii) o **campo de pesquisa** da subárea, a relação entre a língua e o mundo natural; (iv) o **fundamento teórico**, a população e o território são internamente estruturados; (v) a **definição de língua**, maneira de os membros se comunicarem, e (vi) o **que a ecolinguística investiga**, o fenômeno da linguagem e suas relações.

A Figura 4 é a **Proposta 2** e descreve os descritores que fazem parte do ST *língua* em várias perspectivas.

Figura 4 – Proposta 2 - PD enciclopédico baseado em Frames: língua.

Nota: a língua é (i) uma capacidade inata nos seres humanos e a materialização histórica da linguagem; (ii) se manifesta pelo discurso individual como meio de interação entre os homens; (iii) se consolida como instituição humana e se altera conforme os contextos de uso da fala ou da escrita; (iv) concretizada pela interação humana, constituída pela historicidade, refletindo o espírito-nacional e formalizada em documentos.

Fonte: Elaborada pelo autor.

Primeiramente, na Proposta 2, o **descriptor** (i) descreve o conceito de *língua* de uma perspectiva biológica, como capacidade inata dos seres humanos. Depois, *língua* é abordada a partir de uma perspectiva do discurso no **descriptor** (ii), enquanto o **descriptor** (iii) apresenta *língua* como fator social, na forma escrita e falada. Por fim, o **descriptor** (iv) descreve *língua* de um ponto de vista histórico, de identidade nacional e de objeto de registro formal escrito.

Na Figura 5, o PD proposto para a *biolinguística* interage com o consultante por meio de perguntas e mostra o estado da subárea como ciência. Esta é a **Proposta 3** do PD enciclopédico do ST *biolinguística*.

Figura 5 - Proposta 3 - PD enciclopédico baseado em Frames: biolinguística.

Nota: a **biolinguística** concentra-se no estudo da linguagem como uma biologia abstrata, juntando as descobertas e ideias dos vários ramos da biologia e da linguística para apontar de que modo a faculdade da linguagem desenvolveu-se no homo sapiens; a pergunta a ser pesquisada é: qual foi a mudança genética responsável pelo surgimento de uma gramática universal (GU) nos humanos? Ao amplificar a noção de biologia, a linguagem é concebida como característica ou propriedade biológica do ser humano, isto é, um fenômeno natural. De caráter interdisciplinar, a biolinguística busca estudar o pensamento, a sintaxe, a capacidade executiva, a representação simbólica, a memória e a ativação de tecido cerebral no desenvolvimento da linguagem; abordagem ou perspectiva do gerativismo linguístico que se refere às (i) relações cognitivas do ser humano, (ii) ao desenvolvimento da linguagem na espécie humana e (iii) à evolução da linguagem.

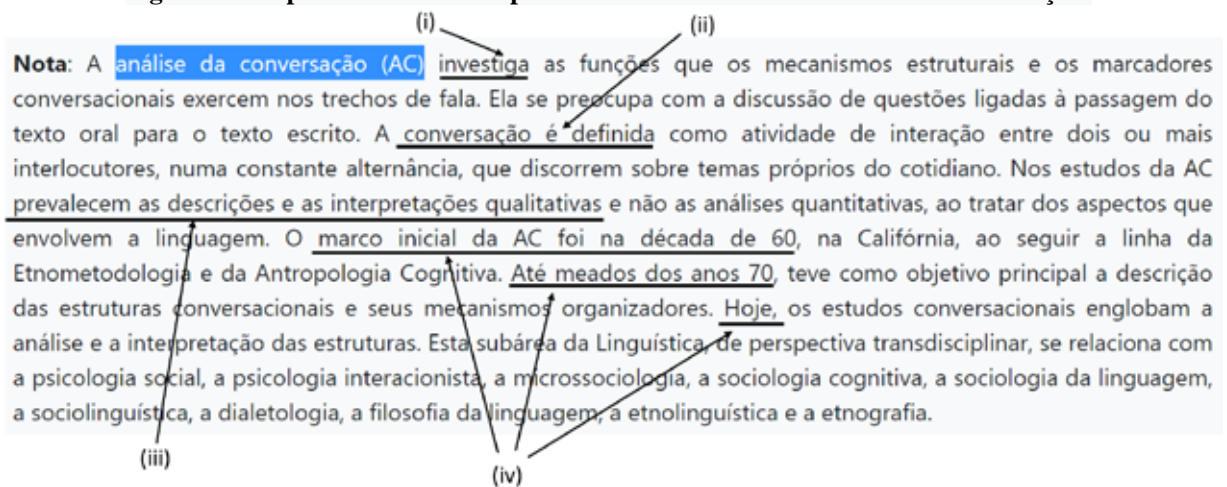
(i) (ii) (iii) (iv)

Fonte: Elaborada pelo autor.

De acordo com a **Proposta 3**, o **descriptor (i)** trata do objeto de estudo da *biolinguística*, a linguagem como biologia abstrata e mostra a intersecção entre a linguística e a biologia. A pergunta de pesquisa da *biolinguística* é apresentada no **descriptor (ii)** por meio de uma frase interrogativa e o **descriptor (iii)** registra a concepção de linguagem como fenômeno natural do ser humano. No **descriptor (iv)** há a abordagem ou a perspectiva a partir da qual a *biolinguística* é embasada, no gerativismo de Chomsky e na teoria da gramática universal (GU).

Finalmente, a Figura 6 descreve o PD da *análise da conversação* com os dados presentes nos PD enciclopédicos anteriores e um breve histórico da subárea.

Figura 6 - Proposta 4 - PD enciclopédico baseado em *Frames*: análise da conversação.



Fonte: Elaborada pelo autor.

O PD da Proposta 4 delinea a definição enciclopédica do ST *análise da conversação*. Na Figura 6 o **descriptor (i)** fornece informações quanto ao objeto de estudo. O **descriptor (ii)** define o conceito de *conversação* nesta subárea. O **descriptor (iii)** descreve como a pesquisa é conduzida, com foco nas interpretações qualitativas dos resultados. O **descriptor (iv)** aborda um breve histórico da disciplina desde seu marco inicial nos anos de 1960, suas alterações nos anos de 1970 e como é conduzida hoje, com perfil transdisciplinar diferente de seu perfil inicial.

Esses paradigmas foram adotados para definir as subáreas e os ST da Linguística. Como explicado previamente, a proposta de *Frames* tem o foco na recorrência dos traços distintivos mais frequentes. A ausência dos traços uniformes nos PD se justifica pela singularidade de escolha que cada autor tomou para produzir o discurso especializado nas produções acadêmicas.

Na seção seguinte, abordaremos como foi feita a compilação e o tratamento do *corpus* de análise desta tese.

5. COMPILAÇÃO E TRATAMENTO DO CORPUS

Nesta seção, faremos um apanhado dos *corpora* que compõem esta pesquisa, sendo eles todos da grande área da **Linguística**, subdivididos, quanto ao conteúdo, nas subáreas da **Linguística Descritiva** e da **Linguística Aplicada**. Em relação ao gênero, todos os *corpora* são provenientes do gênero científico de divulgação e instrucional, subdivididos em artigos científicos, apostilas, livros, manuais e aulas transcritas (ALMEIDA; CORREIA, 2008). Inicialmente, imaginamos ter somente manuais, teses, dissertações e artigos científicos, mas com a **normalização**, encontramos as variedades distintas especificadas acima.

1.15. *Corpora de textos acadêmicos*

Os *corpora* de textos acadêmicos são compostos de artigos científicos das áreas da Linguística Descritiva e da Linguística Aplicada, escritos por profissionais da área de Letras, Linguística, Literatura e outras, publicados em plataformas universitárias, revistas, periódicos e anais de eventos.

Conforme já foi anunciado anteriormente, o objetivo desta pesquisa é construir um protótipo de vocabulário de Linguística que englobe a Linguística Descritiva (LD) e a Linguística Aplicada (LA). A expressão LD foi selecionada para que se opusesse a LA, de forma que houvesse mais clareza na exposição das informações.

Os autores dos *corpora* são de perfis variados: doutores, mestres, especialistas, discentes de graduação e especialistas de outras áreas que tenham publicado na área da Linguística, como acontece em entrevistas e artigos de perfil interdisciplinar. Para exemplificar essa diversidade, mencionamos o artigo *Etimologia Aquática*, escrito por Von Sperling, profissional da área de Engenharia Sanitária e Ambiental; a dissertação da mestra em distúrbios da Comunicação Humana, Vanessa Giacchini, que se enquadra na área da Fonética e Fonologia; o artigo *O articulatório e o fonológico na clínica da linguagem: da teoria à prática*, escrito por fonoaudiólogos, um engenheiro e matemático, e um profissional da Letras.

O perfil de autores profissionais de Letras são professores doutores, mestres, especialistas e discentes de graduação que publicam em periódicos e revistas como *Delta*, *Domínios de Linguagem*, *Ecolinguística*, *Filologia* e *Linguística Portuguesa*, *Intersecciones*, *Philologus*, *Revel*, *SoLetras*, entre outras. Incluímos nessa lista a base de dados do *Scielo*. Anais de congressos também se constituíram fontes de pesquisa, nas quais encontramos artigos

escritos por doutores, mestres, professores e discentes de graduação, quando publicam trabalhos completos de suas apresentações nos eventos científicos de Letras.

Dissertações de mestrado e teses de doutorado foram acessadas a partir do banco de dados das universidades e das bases de dados do governo brasileiro, como a Biblioteca Digital Brasileira de Teses e Dissertações⁵⁹, Catálogo de Teses e Dissertações da CAPES⁶⁰ e Domínio Público⁶¹. A maioria do conteúdo, em formato pdf, pôde ser salva em formato txt. Quando essa opção não existia, optamos por buscar outros arquivos disponíveis para acesso público, não bloqueados.

A LC foi adotada como metodologia e como abordagem pelos seguintes motivos: (i) possibilita a descrição da língua em uso; (ii) possibilita a análise quantitativa e qualitativa de grande volume dos dados linguísticos; (iii) fornece resultados estatisticamente mais precisos dos termos usados pelos especialistas; (iv) permite o uso de recursos computacionais; e (v) é uma metodologia bem solidificada em nível nacional e internacional.

Para uma melhor compreensão da escolha da LC, é importante dizer que sem o processamento de dados, essa abordagem seria impossível devido a dois fatores: (i) a área é extensa, considerando que 49 subáreas foram mapeadas e compiladas; e (ii) a quantidade de dados não seria processada de forma humana no período do doutorado, o que significa um total de 49 milhões de palavras.

O uso desses *corpora* proporciona a disponibilização de dados oriundos do uso real da língua aos usuários, pelo menos, na modalidade escrita. Essa disponibilização será realizada por meio do VoBLing, pois nessa plataforma as definições são construídas a partir de traços distintivos provenientes dos *corpora* acadêmico e dos manuais de Linguística de forma parcialmente automatizada.

Uma limitação que teríamos, caso não adotássemos o VoTec (FROMM, 2007), a LC e os *corpora* acadêmicos, era a de saber se um termo candidato seria ou não uma palavra-chave da área ou subárea estudada, segundo a *lista de palavras-chave*. Além disso, não teríamos como localizar **contextos definitórios** e **explicativos** para a construção de definições adequadas na elaboração da **definição terminológica** (DT) sem os contextos fornecidos pelas linhas de concordância do *WordSmith Tools* 7.0 e 8.0 (SCOTT, 2016; 2020), (WST).

⁵⁹ Biblioteca Digital Brasileira de Teses e Dissertações. Disponível em: <http://bdtd.ibict.br/vufind/>. Acesso em: 01 jul. 2018.

⁶⁰ Catálogo de Teses e Dissertações. Disponível em: <http://catalogodeteses.capes.gov.br/catalogo-teses/#!/>. Acesso em: 01 jul. 2018.

⁶¹ Domínio Público, pesquisa teses e dissertações. Disponível em: <http://www.dominiopublico.gov.br/pesquisa/PesquisaPeriodicoForm.jsp>. Acesso em: 01 jul. 2018.

Segue, no Quadro 8, a taxonomia dos *corpora* de Linguística, de acordo com a proposta de Teixeira (2008).

Quadro 8 - Taxonomia dos *corpora* de artigos científicos: Linguística (português-inglês)

1. Língua	Bilíngue (inglês e português)
2. Modo	Escrito (textos acadêmicos: manuais, artigos científicos, dissertações e teses)
3. Data de publicação	Sincrônico (levantamento realizado entre 2010 e 2018)
4. Seleção	Estático
5. Conteúdo	Especializado (Linguística Descritiva e Linguística Aplicada)
6. Autoria	Falantes nativos e não nativos (inglês e português), individual/coletivo
7. Disposição Interna	Comparável
8. Uso na pesquisa	Estudo (análise terminológica/terminográfica)
9. Nível de Codificação	Com cabeçalhos, sem etiquetas

Fonte: Teixeira (2008).

O Quadro 8 traz a taxonomia/tipologia dos *corpora* que utilizamos neste trabalho, o que explicamos a seguir. Eles são *corpora*:

1. **bilíngues**, abarcam as línguas portuguesa e inglesa;
2. **escritos**, contrastam com o oral;
3. **sincrônicos**, trazem o registro de dada língua em um momento específico da história, enquanto os diacrônicos cobrem períodos diferentes dessa língua - os textos dos *corpora* em epígrafe abarcam o período de 2010 a 2018, considerado o período de busca e digitalização dos manuais;
4. **estáticos**, ou seja, não permitem alteração, isto é, a inclusão ou exclusão de textos;
5. **especializados**, objetivam a linguagem de especialidade da Linguística Descritiva e da Linguística Aplicada, contrastando, portanto, com os de língua geral;
6. **autoria**, escritos em LP de português europeu, africano ou brasileiro; e a autoria de LI é de falantes de inglês de várias nacionalidades;
7. **comparáveis**, pois os *corpora* de português não são uma tradução dos *corpora* de inglês e vice-versa. Tagnin (2015) os define da seguinte forma: “textos originais em duas (ou mais) línguas, numa determinada área de domínio” (TAGNIN, 2015, p. 26);
8. *corpora* de **estudo**, usados para estudos nesta pesquisa;
9. **codificação**, com cabeçalhos, sem etiquetas.

As subáreas que haviam sido compiladas pelos discentes também foram compiladas pelo pesquisador deste trabalho. Este acervo totalizou aproximadamente 49 milhões de palavras nas duas línguas, sendo 46,4 milhões de itens a partir de textos acadêmicos e aproximadamente 2,6 milhões de itens a partir dos manuais. Como metodologia da primeira etapa do trabalho completo de compilação, tivemos as seguintes etapas:

- a) uma “árvore de domínio” da Linguística foi apresentada aos discentes;
- b) os discentes foram orientados a optar por uma subárea da Linguística, cujo *corpus* ainda não houvesse sido coletado;
- c) sob a supervisão do professor Fromm, os discentes pesquisaram na internet os textos disponíveis nessa subárea (em português e em inglês), salvando-os em formato txt. Os

textos em txt foram processados pelo WST, por meio das ferramentas *lista de palavras* com aplicação de *stoplists*, *lista de palavras-chave* e concordanciador.

- d) Realizada 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), disponível em <http://treino.votec.ileel.ufu.br/>. Os discentes 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 no Quadro 9, a seguir.

Quadro 9 - Manuais de Linguística em inglês e português

MANUAIS DE LINGÜÍSTICA EM LI		
	Nome da obra	Autor
1.	<i>General Linguistics: An Introductory Survey - Third Edition</i>	Robins
2.	<i>The Handbook of Linguistics</i>	Aronoff e Rees-Miller
3.	<i>The Handbook of Applied Linguistics</i>	Davies e Elder (Editors)
4.	<i>The Handbook of English Linguistics</i>	Aarts & McMahon
5.	<i>Linguistics</i>	Widdowson
MANUAIS DE LINGÜÍSTICA EM LP		
1.	Manual de Expressão Oral e Escrita	J. Mattoso Câmara Jr.
2.	Fundamentos da Linguística Contemporânea	Edward Lopes
3.	Introdução à Linguística: objetos teóricos	José Luiz Fiorin (Org.).
4.	Introdução a Linguística II: princípios de análise	José Luiz Fiorin (Org.).
5.	Introdução à linguística: domínios e fronteiras, v. I	Fernanda Mussalim, Anna Christina Bentes (Org.).
6.	Introdução à linguística: domínios e fronteiras, v. II	Fernanda Mussalim, Anna Christina Bentes (Org.).
7.	Introdução à linguística: domínios e fronteiras, v. III	Fernanda Mussalim, Anna Christina Bentes (Org.).
8.	Introdução à Linguística Teórica	John Lyons
9.	Linguística e comunicação	Roman Jakobson
10.	Manual de Linguística: subsídios para a formação de professores	BRASIL
11.	Manual de Linguística Românica	Vidos
12.	Manual de Linguística: fonologia, morfologia e sintaxe	Schwindt (Org.).
13.	Manual de Linguística	Pais et al.

Fonte: Elaborado pelo autor.

O tamanho desses *corpora* foi equilibrado com aproximadamente 1,1 milhão de palavras/*tokens* em cada língua. O uso de manuais de Linguística se justifica como aprimoramento da dissertação de mestrado do pesquisador. A indicação para ampliarmos os

corpora da pesquisa e incluirmos esses manuais foi feita para que tivéssemos um *corpus* no qual os contextos definitórios e explicativos fossem mais frequentes que em artigos científicos, dissertações e teses.

Quanto à metodologia usada para a padronização desses *corpora*, seguimos os passos adotados para os *corpora* de teses, dissertações e material instrucional, isto é, fizemos a limpeza, o redimensionamento e, por fim, o balanceamento dos manuais, conforme explicamos a seguir.

A limpeza consistiu na retirada do conteúdo denominado *outside matter*, subdividido em *front matter*, *middle matter* e *back matter* (YAMAMOTO, 2015, p. 51), dos manuais em formato txt. Para ajustar esses padrões aos manuais de Linguística, eliminamos o nome das obras e do(s) autor(es), a apresentação e a parte introdutória como constituintes do *front matter*. Quanto ao *middle matter*, eliminamos os quadros, as leituras complementares (ao final dos capítulos), as bibliografias complementares e os comentários de rodapé. Para finalizar, relacionada ao *back matter*, eliminamos os exercícios ou atividades didáticas, as bibliografias, os sumários e os índices remissivos das obras em LP e em LI.

Concluída a limpeza, fizemos a leitura dos *corpora* com o WST para quantificá-los. O parâmetro adotado para quantificar o tamanho dos *corpora* foram os dados disponibilizados na aba *statistics* e nos da coluna *tokens used for Word list*, conforme a Figura 7 abaixo:

Figura 7 - Dados estatísticos dos *corpora* de manuais.

Figura 7 - Dados estatísticos dos corpora de manuais.

Statistics List (C:\Users\issam\OneDrive\Normalizado Guilherme e Márcio - versão de ajuste\Wordlists_ML_PT_M1-M13\Manuais_L_PT_M1-M13_files\list)

N	text file	file size	tokens (running words) in text	tokens used for sum of word list	types (distinct words)	type/token ratio (TTR)	standardized TTR	STTR std. dev.	STTR basis	mean word length (in)	word length std. dev.	se...	mean...	
	Overall	13.816.132	1.101.544	1.079.337	0	47.4394,40%	41...%	5...6	1.000	5,00	3,34	7...0	3...8	
1	ML...11	675.882	58.548	56.149	0	7.04012...%	40...%	5...8	1.000	4,52	3,10	1...8	3...3	
2	ML...T2	1.080.166	84.201	81.759	0	10.39812...%	43...%	5...6	1.000	5,06	3,35	2...2	2...8	
3	ML...T7	2.240.330	174.382	171.893	0	13.8968,08%	43...%	5...2	1.000	5,23	3,45	5...4	6...2	
4	ML...T6	975.784	75.132	74.199	0	9.18112...%	42...%	5...1	1.000	5,28	3,50	2...7	2...2	
5	ML...XT5	899.078	69.782	68.394	0	8.72812...%	42...%	5...0	1.000	5,20	3,39	2...8	2...8	
6	ML...T4	1.075.436	87.624	85.687	0	9.73911...%	38...%	5...2	1.000	4,85	3,19	3...7	2...9	
7	ML...T3	1.193.624	96.257	95.040	0	10.51611...%	40...%	5...4	1.000	4,99	3,25	4...7	2...8	
8	ML...T8	2.388.022	188.670	185.024	0	13.0907,07%	40...%	5...0	1.000	5,06	3,36	6...8	5...1	
9	ML...T12	646.372	56.214	54.590	0	6.48311...%	38...%	5...1	1.000	4,42	3,14	1...2	2...0	
10	ML...T1	517.762	38.324	37.617	0	5.93015...%	45...%	5...4	1.000	4,91	3,29	1...2	2...1	
11	ML...13	923.252	80.072	77.791	0	7.6709,86%	37...%	6...5	1.000	4,54	3,25	2...6	2...5	
12	ML...10	603.362	46.103	45.318	0	7.06615...%	43...%	5...3	1.000	5,25	3,46	30914...	4...6	
13	ML...T9	507.062	46.235	45.876	0	8.74417	%48	%4	0	1.000	5,23	3,37	1...0	2...5

frequency alphabetical **statistics** filenames notes

14 entries Row 1 T S < > Help

Fonte: Elaborada pelo autor.

De acordo com a Figura 7, vemos o dimensionamento dos *corpora* de manuais de Linguística em LP, onde há treze (13) manuais, de tamanhos diversos, totalizando 1.079.337 *tokens* utilizados na lista de palavras.

Finalizando o procedimento de padronização dos *corpora* de manuais, reduzimos os *corpora* de LI, para igualar aos de LP. Este corte se fez necessário, pois os *corpora* de manuais em LI eram de 1.230.061 *tokens* e os de LP de 1.079.337 *tokens*. Equilibrados os *corpora* de manuais em LP e LI, obtivemos os dimensionamentos de 1.079.337 *tokens* para a LP e de 1.099.776 *tokens* para a LI. Somando os dois *corpora* e dividindo por dois, chegamos ao dimensionamento médio de 1.089.556 *tokens*.

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

- (i) compará-lo ao *corpus* de LD e LA;
- (ii) identificar qual dos *corpora* trará mais **contextos definitórios** (DUBUC, 1992; AUBERT, 2001; PAVEL; NOLET, 2002; BARROS, 2007); e
- (iii) usá-los para a construção de definições na plataforma do VoBLing.

Um passo metodológico importante nesta pesquisa foi o de padronizar as denominações que são usadas ao descrevermos os diferentes tipos de *corpora*. Primeiramente, especificarei e descreverei os *corpora* que compõem esta pesquisa. Em seguida, explicarei como chamaremos cada componente desses *corpora*.

Os componentes dos *corpora* dessa pesquisa são:

- (i) um *corpus* de Linguística Descritiva, em LP, composto de 29 subáreas, da Análise da Conversação à Terminologia, organizado em ordem alfabética;
- (ii) um *corpus* de Linguística Descritiva, em LI, composto de 29 subáreas, da Análise da Conversação à Terminologia, organizado em ordem alfabética;
- (iii) um *corpus* de Linguística Aplicada, em LP, composto de 18 subáreas, do Bilinguismo e Multilinguismo à Tradução, organizado em ordem alfabética;
- (iv) um *corpus* de Linguística Aplicada, em LI, composto de 18 subáreas, do Bilinguismo e Multilinguismo à Tradução, organizado em ordem alfabética;
- (v) um *corpus* de Manuais de Linguística, em LP, composto de 13 manuais, classificados numericamente de um a doze;
- (vi) um *corpus* de Manuais de Linguística, em LI, composto de 5 manuais, classificados numericamente de um a cinco.

Denominaremos *corpora* de Linguística a esses seis *corpora* independentes, compostos por artigos científicos e manuais.

Ao nos referirmos a cada *corpus* individualmente, trataremos de *corpus*, *subcorpora* e *subcorpus*. Por exemplo, temos o *corpus* da Linguística Descritiva, e seus *subcorpora*, sendo que cada área será um *subcorpus*. Esse parâmetro se aplica para a Linguística Descritiva e para a Linguística Aplicada, em português e em inglês.

Já em relação aos manuais, adotamos a seguinte denominação:

- (i) os *corpora* de manuais, ao tratarmos dos *corpora* de LP e LI juntos;
- (ii) o *corpus* de manuais será usado para tratarmos do conjunto de manuais de uma única língua, quer seja LP ou LI; e
- (iii) ao referirmo-nos a cada manual, como livro, em específico, componente do *corpus* de manuais de uma língua específica, usamos o termo *arquivo*. Por exemplo: No *corpus* de manuais de LI, há os manuais denominados *arquivos* 1 a 5.

Os verbetes são construídos a partir de *corpora* autênticos, que serão organizados por entradas em português e em inglês. Essa forma de trabalho é referenciada 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; FINATTO, 2004, p. 27, grifo nosso).

Antes que o tratamento de dados fosse feito, procedemos à correção dos manuais em formato txt, codificação Unicode. Esse procedimento se fez necessário, pois a maioria dos manuais foi escaneado devido ao limitado número deles disponível na internet. Após serem escaneados, eles foram salvos em formato txt com o auxílio do *OmniPage16*, software pago que transforma arquivos pdf em txt, graças à funcionalidade *Optical Computer Reading* (OCR). Os arquivos em txt foram abertos no programa *Word* e corrigidos com a ferramenta de correção textual em LP e LI.

Após o tratamento dos dados (construção da *lista de palavras*, *lista de palavras-chave*, análise dos cotextos via concordanciador), os traços distintivos foram organizados em fichas terminológicas disponíveis no VoBLing.

Para que fosse feita a **padronização** do *corpus* de artigos acadêmicos, em primeiro lugar, partimos da organização das subáreas propostas na árvore de domínio da Linguística de Fromm (2018). Os *corpora* que foram organizados estavam em uma pasta do *OneDrive* do orientador, posteriormente compartilhada com o pesquisador para que a **padronização** fosse realizada. O que descrevemos a seguir são os procedimentos que foram feitos rumo a essa **padronização**:

1. as pastas foram organizadas de acordo com a árvore de domínio (FROMM, 2018); as

pastas cujas subáreas não estavam presentes nessa árvore foram excluídas, pois o recorte metodológico que focamos neste trabalho é restrito às 49 subáreas da Linguística, acrescidas das duas subáreas inseridas a partir desta pesquisa: a **Biolinguística** e a **Linguística Diacrônica**. Ademais, em 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 que foi mantida;

2. o próximo passo foi o dimensionamento do *corpus* contido em cada pasta: baixamos os arquivos das subáreas da Linguística Aplicada ou da Linguística Descritiva e fizemos seu processamento com o WST para determinar o número de palavras contido em cada subárea. Os dados obtidos estão relacionados do **Apêndice D** ao **Apêndice G**.

Na execução do passo 1, ocorreu que tínhamos a subpasta **Pesquisa Narrativa** na pasta de LA/LI, mas esta não constava na árvore de domínios. Além dela, havia a subárea de **Linguística de Contato** na **árvore de domínio**, cuja pasta era 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 foi feita a coleta de *corpus* da subárea da **Linguística de Contato**.

A área da **Pesquisa Narrativa** (PN) foi excluída, pois, em consulta bibliográfica, observamos que a PN é concebida como metodologia (DUARTE NUNES, 2018) ou abordagem epistemológica de investigação qualitativa (GONÇALVES; NARDI, 2016) de caráter interdisciplinar, não sendo especificamente da grande área da Linguística.

Com relação ao passo 2, no qual dimensionamos o tamanho dos *corpora*, identificamos quatro 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 que não continham arquivos. A visualização deste dimensionamento está disponível dos **Apêndice D** ao **Apêndice G**. A **padronização** do número de 500 mil palavras por subárea se deve à formatação original do projeto, na qual esse número foi considerado como representativo da subárea, já que responderia pela dimensão de um *corpus* médio, conforme Teixeira (2008, p. 184).

Diante do exposto quanto ao passo 2, três ações foram necessárias: a primeira foi a de reduzir a quantidade de arquivos nas pastas que continham mais de 1 milhão de palavras. Para tal ação, fizemos uma análise dos textos utilizando o WST e mantivemos os arquivos nos quais a presença de palavras-chave foi maior. A segunda ação foi aumentar o *corpus* nas pastas que continham um menor número de palavras a fim de que ele atingisse o número padrão de 500

mil palavras. A terceira ação foi compilar os textos para aquelas áreas que ainda não os possuíam.

Quanto ao dimensionamento dos *corpora* de manuais, na primeira parte da pesquisa não nos foi possível prever seu tamanho, já que, àquele momento, não podíamos prever a quantidade de *corpora* disponíveis na internet ou ainda se todos teriam que ser digitalizados. Até aquele momento, tínhamos compilado aproximadamente 320 mil palavras em português e 1,2 milhões em inglês. Esses dados mostravam que o *corpus* em português era menor que o em inglês, o que nos demandou a tarefa de equilibrar esses *corpora*, após a finalização do processo de compilação.

Compilamos por meio da ferramenta de busca do Google os manuais em inglês e em português, já digitalizados, e observamos que havia uma incidência maior de manuais em inglês do que em português. Essa diferença se dá porque os manuais em inglês apresentam maior número de *tokens* que os manuais em português. Logo, em nossos *corpora* de manuais, a quantidade de *tokens* em LI é menor que em LP.

A falta de manuais digitalizados em português mostrou a necessidade de se escanear as obras nesse idioma, o que fizemos por várias horas na biblioteca da UFU. As etapas de compilação dos manuais de Linguística foram:

- 1) seleção dos manuais que continham claramente as subáreas da Linguística; e
- 2) seleção dos manuais que eram direcionados para um público-alvo de iniciantes (checagem da introdução).

Após a seleção das obras, seguiu-se para a digitalização delas em formato pdf, já que esse era o único tipo de formato de arquivo disponível, à mão, que escaneasse as obras rapidamente. Nesse caso, as obras foram escaneadas por um escâner planetário disponível na biblioteca da Universidade Federal de Uberlândia. A capacidade de escaneamento foi de 100 páginas a cada dez minutos, aproximadamente.

Em seguida, as obras foram processadas pelo programa *OmniPage16*, que transformou os arquivos de formato pdf em txt. Nesse caso, as obras foram carregadas por completo e depois processadas pelo programa.

Durante o processo inicial de digitalização das obras, o escaneamento foi feito por páginas, em vez de ter sido feito como um arquivo único devido às configurações prévias da máquina de escâner. Nesse caso, para que não se perdesse o trabalho de escaneamento, lançamos mão de uma ferramenta on-line⁶² por meio da qual podíamos carregar até vinte

⁶² O site usado foi o <http://combinepdf.com/>.

arquivos para que fossem combinados em arquivo único. Como na obra em questão havia 131 páginas (em arquivo), o procedimento foi realizado repetidas vezes até que chegássemos ao arquivo único.

Nessa seleção de manuais em português, houve uma questão que 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 composição textual, observamos que ele é um manual rico em conceitos, o que nos interessa para cumprir o objetivo deste trabalho. Esses conceitos podem ser úteis na seleção de contextos explicativos para deles extrairmos traços distintivos e, posteriormente, construirmos as definições dos ST. Por isso, optamos por mantê-lo em nossa lista de obras compiladas que compõem nosso *corpus* de manuais.

Com a compilação dos dois tipos de *corpora* concluída, fizemos uma análise comparativa para observar qual deles trazia mais contextos definitórios úteis para a construção das definições, e a possível eleição dos candidatos a termos. Contudo, ambos os *corpora*, o de artigos acadêmicos e os de manuais, foram usados para o processo terminográfico de escolha de candidatos a ST, a termos e a contextos definitórios. A seleção de candidatos a termos foi norteada por textos que traziam o ST inserido em contextos explicativos ou definitórios, gerando, assim, o vocabulário de especialidade. Como exemplos de contexto explicativo e definitório para os ST *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 behavior of men and women in society, in interaction with their fellows.* (em negrito)

Para ST em português temos 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 ideia) 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 retratam dois tipos de contextos definitórios por meio da utilização dos verbos SER e TRATAR. Observa-se que os exemplos 1, 2 e 3 definem o termo por meio do verbo SER, enquanto o exemplo 4 usa o verbo TRATAR para explicar um dos tipos de textos dos quais a Linguística se ocupa: dos textos em prosa.

1.1.21. Digitalização e correção dos *corpora*

O trabalho de digitalização das obras é mais árduo do que a busca dos termos na internet, por ser (i) mais mecânico e (ii) pelo fato de a tecnologia disponível até o momento não ser tão rápida. A variável que tivemos que considerar quanto ao tempo é se a imagem escaneada ficaria legível ou embaçada ou distorcida. Caso a imagem ficasse 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 *OmniPage16*.

O procedimento foi, de certa forma, moroso, já que esse programa demandou um computador com um bom processador e uma grande capacidade de memória, o que nem sempre o pesquisador teve disponível à mão. Ao usar o *OmniPage16*, como os arquivos têm uma extensão maior que 100 (cem) páginas, foi necessário certificar-se de que as obras tinham sido decodificadas em toda sua extensão, para que não ficassem incompletas.

Além do uso do *OmniPage*, fizemos uma tentativa de usar o serviço pago do PDF *services*⁶³. Contudo, nossos arquivos eram grandes (50.000 KB), e o *site* não conseguia fazer o carregamento. Devido a essa limitação, optamos por desistir de usar esse recurso.

A busca por uma nova ferramenta se fez necessária, pois a **decodificação** de caracteres feita pelo OCR do *OmniPage16* produziu muitos erros, o que demandou um trabalho maior para a correção no *Word*.

O *Acrobat Pro DC* foi a ferramenta escolhida para a transformação dos arquivos de imagem em pdf para txt e surtiu um resultado melhor que os outros programas. A decodificação e identificação de caracteres do português por meio desse programa foram mais eficientes que nas outras ferramentas.

Em todas essas ferramentas, configuramos a língua para o português brasileiro a fim de possibilitar a decodificação de diacríticos. Para efeito de padronização, todas as palavras foram digitadas com letras minúsculas. Como podemos observar no Quadro 10, a acentuação realmente foi um fator determinante na decodificação das palavras.

Quadro 10 - Problemas ortográficos corrigidos no *corpus* de manuais em português – junho/2018.

	Forma decodificada (OCR)	Forma corrigida
1.	Historia	História
2.	Experiencia	Experiência
3.	Lingilistica	linguística
4.	Lingthstica	linguística
5.	Linguística	linguística
6.	Lingoisticos	linguísticos
7.	Começa	começa

⁶³ Serviço pago do Adobe Acrobat.

8.	Investigagao	investigação
9.	vac)	vão
10.	sac)	são
11.	regides	regiões
12.	op6e	opõe
13.	lung:a°	função
14.	atualizageo	atualização
15.	to-lo	tê-lo
16.	sec:06116ª	sequência
17.	lingaisticos	linguísticos
18.	seq0oncias	sequências
19.	colocanla	colocá-la
20.	idonticos	idênticos
21.	humanism°	humanismo
22.	sinai	sinal
23.	corn	com
24.	direr	dizer
25.	loma	toma
26.	camp	campo
27.	especiffr cidade	especificidade
28.	sign% cado	significado
29.	ril.dd,° _	modo
30.	assirn	assim
31.	alg l l mas	algumas
32.	enquanio	enquanto
33.	form la	forma
34.	fo l les	fontes
35.	ttaços	traços
36.	enh·e	entre
37.	cttio	cujo
38.	pennitinclo	permitindo
39.	cxc l l l plifica r	exemplificar
40.	flnal	final

Fonte: Elaborado pelo autor.

O Quadro 9 mostra que é possível observar os seguintes problemas de decodificação: (i) decodificação dos sinais diacríticos do português – do 1 ao 20; e (ii) decodificação de letras em geral – do 21 ao 30. Na decodificação de letras, observamos que as letras **m**, **f**, **o** em final de palavras foram as mais frequentes.

Além desses casos, observamos que há uma limitação de decodificação da imagem das letras do 30 ao 40: rn/m, 11/u, i/t, n1/m, 11/nt, T/ra, h·/trete. Os erros de decodificação não eram estáveis, no sentido de que uma decodificação **rn** sempre corresponderia a **m**, por exemplo. Caso esses padrões fossem recorrentes, poderíamos usar a ferramenta **Substituir** do **Bloco de Notas**, o que não foi possível por dois motivos:

1. a decodificação poderia representar um padrão recorrente na língua: (30) rn, como em **moderna** e **materna**;
2. o padrão de decodificação não era recorrente: (31) 11/u e (34) 11/nt; e (35) tt/tr e (37) tt/uj.

No Quadro 10, registramos a quantidade de horas e minutos utilizadas nas atividades de tratamento dos *corpora*.

Quadro 11 - Dimensionamento do tempo de digitalização dos *corpora* (parcial).

Atividade	Tempo horas	Tempo minutos
Digitalização de manuais	18	1080
Correção ortográfica de manuais	324	19.440
Normalização e limpeza dos <i>corpora</i>	653	39.180
TOTAL	995	59700

Fonte: Elaborado pelo autor.

O Quadro 10 mostra o tempo investido na realização de cada uma das atividades de preparação do *corpus*, a saber: (i) a digitalização de manuais, realizada na biblioteca da UFU; (ii) a correção ortográfica que corresponde ao Manual de Linguística, de Cidmar Teodoro Pais (1979); ao Manual de Introdução à Linguística, v. I e II, de Fiorin (2002, 2003); ao Manual de Expressão Oral e Escrita, de Mattoso Câmara (1991); e ao Manual de Linguística, de Schwindt (2014); e (iii) à normalização dos *corpora* que correspondeu à organização das pastas, à padronização do nome dos arquivos, à limpeza e ao equilíbrio dos *corpora*.

Para fazermos a correção dos erros de decodificação, copiamos o conteúdo do arquivo em txt e o salvamos em um documento do *Word*. Em seguida, selecionamos todo o texto e aplicamos a ferramenta de correção de língua portuguesa para a correção ortográfica.

1.1.22.A plataforma *ToGatherUp*

Devido à dimensão dos *corpora* desta pesquisa, foi necessário considerar o aspecto de manutenção dos arquivos em segurança para evitar perdas e danos. A solução que tivemos foi armazená-los nas nuvens, o que poderia ser feito usando as ferramentas disponíveis no mercado como *Google Drive* do *Google*, e *OneDrive* da *Microsoft*. Outra solução era usarmos plataformas disponíveis como *e-Termos* e *Sketch Engine*, pagas ou gratuitas, mas não faziam parte do nosso projeto.

Concomitantemente à condução desta pesquisa, Oliveira (2019)⁶⁴, desenvolveu uma plataforma chamada *ToGatherUp*, sob a orientação do pesquisador Fromm, cujo objetivo é:

auxiliar pesquisadores que precisam gerenciar a coleta de textos para construção de *corpus* de estudo. Ela possui uma interface web que permite a inclusão, o gerenciamento e a visualização de informações estatísticas sobre

⁶⁴ Dissertação disponível em <http://repositorio.ufu.br/handle/123456789/25433>, intitulada *ToGatherUp*: um protótipo de ferramenta para a construção de corpora

os textos e que facilita as rotinas do trabalho de armazenamento de arquivos. (OLIVEIRA, 2019)⁶⁵

Diante da possibilidade de uso dessa plataforma, optamos por adotá-la como parte da metodologia do **VoBLing** em função: (i) da dimensão dos *corpora*, de 49 milhões de itens, (ii) da segurança em relação ao armazenamento de dados e (iii) da praticidade na nomenclatura dos arquivos.

compreensão sobre o funcionamento dessa ferramenta, descreveremos alguns componentes e passos da *ToGatherUp*. Na página inicial, conforme a Figura 8, o usuário encontra a tela de acesso restrito aos pesquisadores. Nesta tela, o pesquisador insere seu usuário e sua senha, fornecidos pelo administrador da plataforma.

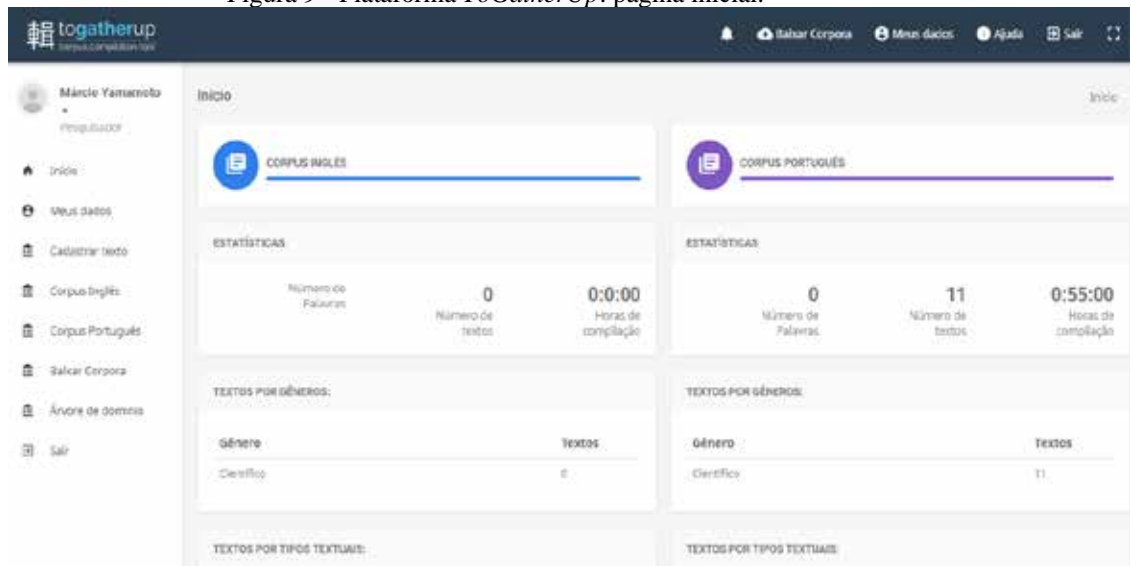
Figura 8 - Plataforma *ToGatherUp*: página de acesso.



Fonte: Elaborada pelo autor.

Na Figura 8 visualizamos os dados do pesquisador e como se faz o acesso à plataforma. A página seguinte, após a verificação do usuário e senha, disponibiliza uma visão geral do projeto desenvolvido e dos *corpora* armazenados na plataforma (cf. Figura 9).

⁶⁵ Disponível em <https://www.ileel.ufu.br/togatherup/linguistica/sobre.php>. Acesso em: 28 jun. 2018.

Figura 9 - Plataforma *ToGatherUp*: página inicial.

Fonte: Projeto do autor na plataforma *ToGatherUp*.

A Figura 9 disponibiliza os dados de Início, isto é, os dados do projeto dispostos em três colunas. A coluna da esquerda traz os ícones: **Meus dados**, **Cadastrar texto**, **Corpus Inglês**, **Corpus Português**, **Baixar Corpora**, **Árvore de Domínio** e **Sair**. O ícone **Meus dados** disponibiliza o nome do pesquisador, seu *status* - ativo (neste caso) - e o quadro para alteração de senha. Por serem dados corriqueiros, não apresentaremos a tela dessa página.

O ícone seguinte, **Cadastrar textos**, disponibiliza as opções para o cadastro dos textos, especificando suas características, conforme retratado na Figura 10, a seguir.

Figura 10 - Plataforma *ToGatherUp*: cadastro de texto.

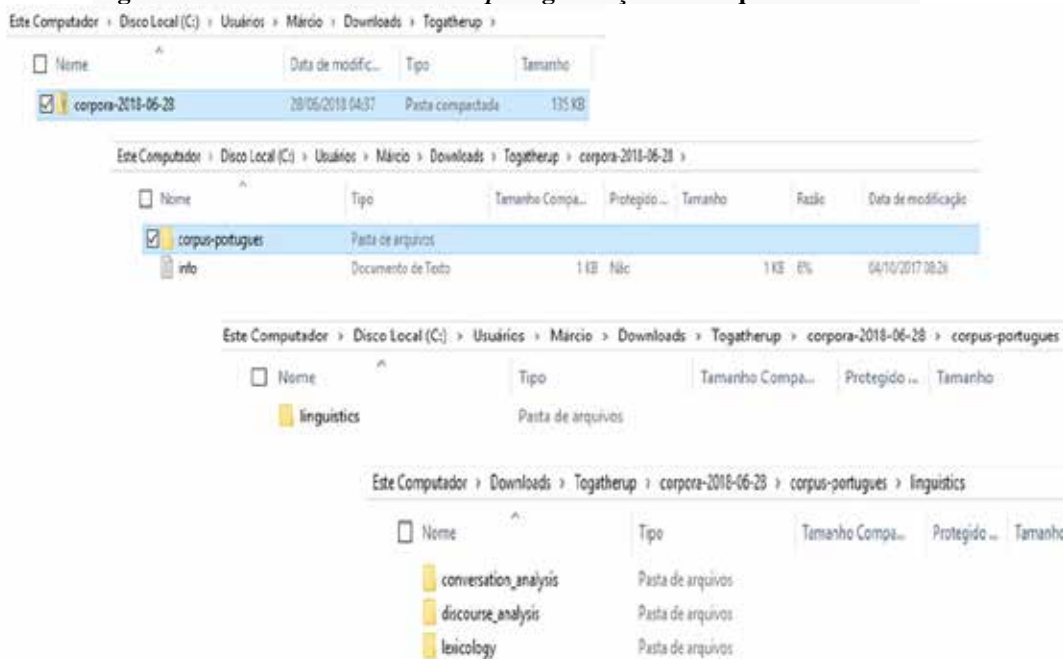
Fonte: Projeto do autor na plataforma *ToGatherUp*.

Conforme disposto na Figura 10, do **Cadastro do Texto**, temos o Formulário de Cadastro de Texto, no qual o pesquisador deve inserir os seguintes dados referentes aos *corpora*: (i) **Subárea**, conforme disponibilizada na árvore de domínio da Linguística (**Erro! Fonte de referência não encontrada.**); (ii) a **Fonte**, onde o pesquisador deve informar o *link* de onde o arquivo foi compilado; (iii) o **Título**, do artigo, da dissertação, da tese ou do manual; (iv) o **Tempo de extração**, já pré-definido pelo criador da plataforma em 5 minutos; (v) a **Língua**, português ou inglês; (vi) o **Gênero textual**, personalizado para este projeto como Científico; (vii) o **Meio de distribuição**: impresso (PR) ou internet (IN); (viii) os **Tipos Textuais**: artigo científico (AC), dissertação (DS), Manual (MA) e Tese (TS); e a caixa de carregamento de arquivos (ix) a seleção do **arquivo para envio**, na qual o pesquisador usa o botão **Escolher arquivo** para buscar e selecioná-lo em seu computador. Ao pressionar o botão **Registrar**, em vermelho, a plataforma faz o carregamento do arquivo para sua base de dados na nuvem.

Por **legibilidade**, concebemos o fato de o arquivo, salvo em formato pdf, permitir que o conteúdo seja salvo em formato txt. Em caso afirmativo, sua legibilidade será maior que a daqueles arquivos que não permitem tal ação e a transformação do texto em formato pdf para o txt será mais rápida. Arquivos que não permitem tal ação, mas não possuem restrição quanto a direitos autorais e uso para pesquisa, foram decodificados por meio de *sites* da internet que possibilitam transformar arquivos de pdf para txt. Caso o texto não seja legível, é necessário buscar outros textos para que os direitos autorais não sejam violados.

Quanto aos itens iv, vii e viii, faremos algumas observações. O tempo de extração (iv) gira numa média de 5 minutos, podendo sofrer variação para mais ou para menos dependendo da pesquisa. As variáveis a serem consideradas são: (1) se a área possui grande acervo disponível na internet para acesso público; (2) se é uma área mais consolidada da Linguística ou se é mais nova, como a **Ecolinguística**; e (3) qual gênero foi selecionado. A questão do gênero interfere no tempo necessário para se transformar um arquivo de formato pdf para txt. Teses e dissertações requerem mais tempo que um artigo científico com média de 20 páginas. Um detalhe quanto ao arquivo é sua **legibilidade**.

A Figura 11 disponibiliza a sistematização de como os *corpora* são organizados na plataforma.

Figura 12 - Plataforma ToGatherUp: organização dos arquivos baixados.

Fonte: Elaborada pelo autor.

A Figura 12 mostra a sequência de como a plataforma organiza o corpus e sua disposição ao ser baixado. No canto superior esquerdo, visualizamos a pasta geral, nomeada *corpus do português*. Dentro da pasta geral, vemos a pasta nomeada *Linguistics*, em cujo interior se encontram as pastas equivalentes às subáreas: *Conversation Analysis*, *Discourse Analysis*, e *Lexicology*. Com o trabalho concluído, é possível visualizar as pastas correspondentes às subáreas da Linguística Descritiva e da Linguística Aplicada em LP e LI.

Finalmente, ao abrirmos a pasta das subáreas, vemos os arquivos individuais, em txt, conforme mostra a Figura 13.

Figura 13 - ToGatherUp: visualização final dos arquivos em txt.

Computador > Disco Local (C:) > Usuários > Márcio > Downloads > Togetherup > corpora-2018-06-28 > corpus-portugues > linguistics > conversation_analysis							
<input type="checkbox"/> Nome	Tipo	Tamanho Comp...	Protegido ...	Tamanho	Razão	Data de modificação	
PT-LG-CI-AC-IN-28Jun2018-3	Documento de Texto	14 KB	Não	37 KB	65%	28/06/2018 03:58	
PT-LG-CI-AC-IN-28Jun2018-4	Documento de Texto	25 KB	Não	67 KB	63%	28/06/2018 03:59	
PT-LG-CI-AC-IN-28Jun2018-5	Documento de Texto	17 KB	Não	43 KB	62%	28/06/2018 04:00	
PT-LG-CI-AC-IN-28Jun2018-6	Documento de Texto	17 KB	Não	43 KB	62%	28/06/2018 04:00	
PT-LG-CI-AC-IN-28Jun2018-7	Documento de Texto	22 KB	Não	60 KB	64%	28/06/2018 04:03	
PT-LG-CI-AC-IN-28Jun2018-8	Documento de Texto	8 KB	Não	19 KB	60%	28/06/2018 04:07	
PT-LG-CI-AC-IN-28Jun2018-9	Documento de Texto	5 KB	Não	11 KB	56%	28/06/2018 04:23	
PT-LG-CI-AC-IN-28Jun2018-10	Documento de Texto	9 KB	Não	19 KB	57%	28/06/2018 04:33	
PT-LG-CI-AC-IN-28Jun2018-11	Documento de Texto	8 KB	Não	19 KB	60%	28/06/2018 04:34	

Fonte: Elaborada pelo autor.

Na Figura 13, é possível ver os arquivos que compõem o *corpus* da Análise da Conversação, em LP, enumerados de 3 a 11, para título de exemplificação neste trabalho.

Em relação à divergência dos acordos ortográficos, é necessário explicar como lidamos com o registro dos títulos dos textos na plataforma. Por exemplo, em relação ao trema, alteramos a escrita nos registros dos textos para o padrão do *ToGatherUp* – novo acordo ortográfico - e mantivemos o padrão do acordo original do *corpus* compilado. Isto é, o nome do artigo na página de visualização será sem trema e no arquivo baixado será com trema (linguística vs. Lingüística).

1.1.23.A plataforma *ToGatherUp* vs WST: codificação textual

Oliveira (2019) ressalta que a codificação textual é um aspecto de suma importância para o êxito no processamento dos dados, para os linguistas que lidam com esses dados e para a codificação e decodificação de várias línguas por meio do formato *Unicode UTF-8*.

McEnery e Xiao (2005, p. 58-72) fazem uma abordagem histórica sobre a criação da codificação UTF e mencionam a implicação de seu uso em *corpora* linguísticos. O termo Unicode (*Unification code*) foi primeiramente usado por Joe Becker da companhia Xerox. O Unicode é concebido com 16 bits (UTF-16), sendo dois bytes para cada caractere, permitindo espaçamento de até 65536 caracteres. Além do formato UTF-16, foram criados os formatos UTF-8, UTF-7 e UTF-32. O formato Unicode foi criado para ser usado em várias plataformas e abranger a codificação e a decodificação de várias línguas, dentre elas o chinês, o japonês e o coreano. Atualmente, apoiados pela Microsoft, a Unicode lançou a sua quarta versão padrão. Devido à proeminência da Microsoft no mercado mundial da informática, os autores recomendam que pesquisas linguísticas adotem este padrão.

Escolhido o formato, esses autores recomendam considerar o formato de transformação a ser adotado. O formato Unicode define a identidade de cada caractere, seu valor numérico (*code point*), e formula como esse valor é representado em bits: quando o caractere é armazenado em um arquivo de computador ou transmitido via intranet ou internet. Esses tipos de formulações são conhecidos como formatos de transformação Unicode, abreviados como UTF.

Nos manuais do WST 7.0 (SCOTT, 2019, p. 531) e do WST 8.0 (SCOTT, 2020, p. 529, 540), o autor enumera os formatos provenientes do Unicode Consortium:⁶⁷ (i) Unicode, Unicode (Big-Endian) (gerados para alguns Mac ou programas do Unix); (2) Unicode (UTF-7)

⁶⁷ O Unicode Consortium é uma organização sem fins lucrativos, fundada para desenvolver, expandir e promover o uso do Padrão Unicode e os padrões de globalização relacionados que especificam a representação do texto em produtos de software modernos e outros padrões. Disponível em <http://www.unicode.org/standard/WhatIsUnicode.html> Acesso em 26 ab. 2020.

e (iii) Unicode (UTF-8), e ressalta a importância da aparência de cada caractere. Ademais, justifica que adotou a codificação Unicode a partir da versão 3 no *Text Converter*, pois permitia o processamento de mais línguas (SCOTT, 2020, p. 562).

Se a codificação não estiver correta, o WST não pode classificar as palavras corretamente ou mesmo reconhecer onde uma palavra começa e onde ela termina. Portanto, é mais seguro converter para Unicode, especialmente se o pesquisador processar textos em alemão, espanhol, russo, grego, polonês, japonês, persa, árabe e outras línguas com acentos e de origem não latinas. Devido ao fato de que o português usa uma acentuação diferente do inglês, como o acento circunflexo e o til, a codificação dos textos em Unicode é recomendável e necessária. Finalmente, como o WST pode trabalhar com várias línguas além do inglês, Scott (2020) recomenda que seja usado o formato **Unicode (little-endian)**, **UTF16**, por abarcar mais línguas que requerem uma codificação acima de 8 bits. Esse aspecto é relevante para a língua portuguesa devido à acentuação dos caracteres.

Foi exatamente nesse aspecto de codificação e decodificação de textos em formato Unicode UTF-16 e UTF-8 que houve a corrupção parcial dos arquivos. Durante o carregamento dos textos da máquina do pesquisador para a plataforma *ToGatherUp*, os arquivos foram corrompidos. A fim de resolver esse problema, inicialmente carregamos os arquivos no formato UTF-16 LE, padronização proveniente do WST, após o uso da ferramenta *Text Converter*. A ferramenta *Text Converter* foi utilizada para padronização dos arquivos, porquanto os *corpora* haviam sido salvos em padrões diferentes, ora UTF-8, ora ANSI⁶⁸ etc.

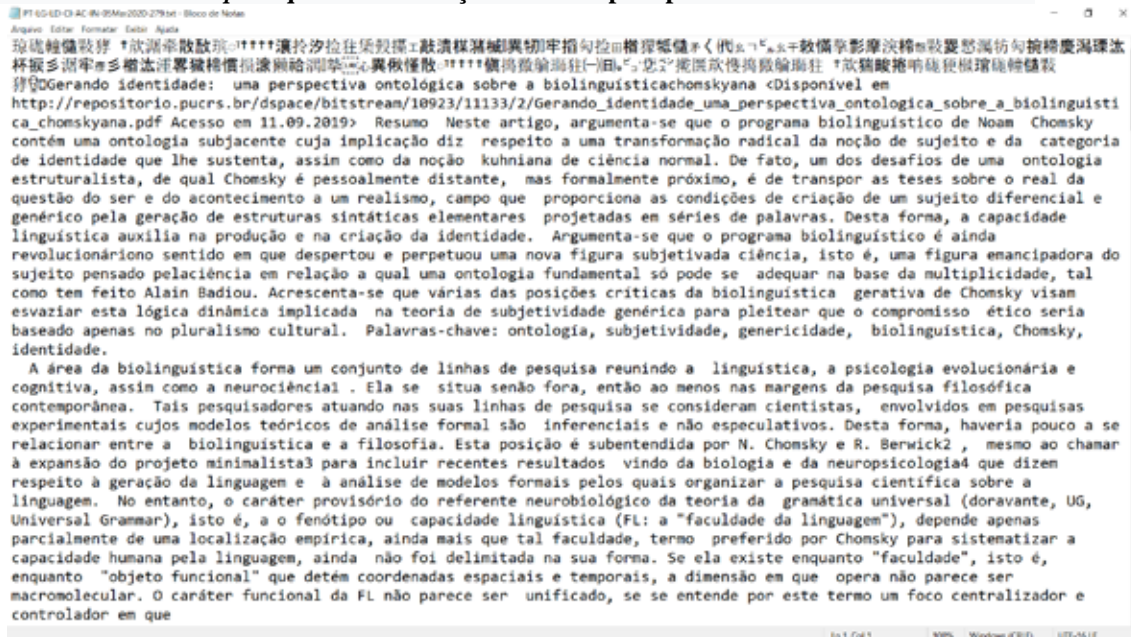
O carregamento dos *corpora* seguia normalmente, até que, ao fazer o *download* dos *corpora*, identificamos dois problemas em parte dos arquivos. Parte deles estava corrompida com duas características: (i) o cabeçalho inserido automaticamente pelo *ToGatherUp* aparecia codificado em chinês quando aberto pelo Notepad, aplicativo do Windows; (ii) o espaçamento entre caracteres do texto era duplicado, apesar de o arquivo originalmente carregado não apresentar excesso de espaçamento entre os caracteres. O sistema operacional da máquina de origem dos arquivos era o Windows, enquanto a plataforma do *ToGatherUp* usava o sistema Unix. Em busca da resolução dos problemas, grosso modo, baixamos os *corpora*, corrigimos e

⁶⁸ ANSI: termo geral usado para se referir à página de código padrão no sistema (geralmente Windows). É mais apropriadamente chamado de Windows-1252, e é uma extensão do conjunto de caracteres ASCII, porque inclui todos os caracteres ASCII e códigos de 128 caracteres adicionais. O motivo dessa diferença é que a codificação "ANSI" é de 8 bits em vez de 7 bits, assim como o ASCII (atualmente, o ASCII é quase sempre codificado como bytes de 8 bits e o MSB é definido como 0). Disponível em <https://www.it-swarm-pt.tech/pt/character-encoding/o-que-e-o-formato-ansi/957862898/>. Acesso em 23 jul. 2020.

salvamos os arquivos em UTF-8 e os carregamos novamente. Descreveremos os procedimentos corretivos em detalhes adiante.

O primeiro procedimento adotado para a correção do cabeçalho em chinês, visível na Figura 14, foi baixar o arquivo e abri-lo com o aplicativo Notepad do Windows.

Figura 14 – ToGatherUp: arquivo com cabeçalho corrompido pós-download de base Windows.



Fonte: Elaborada pelo autor.

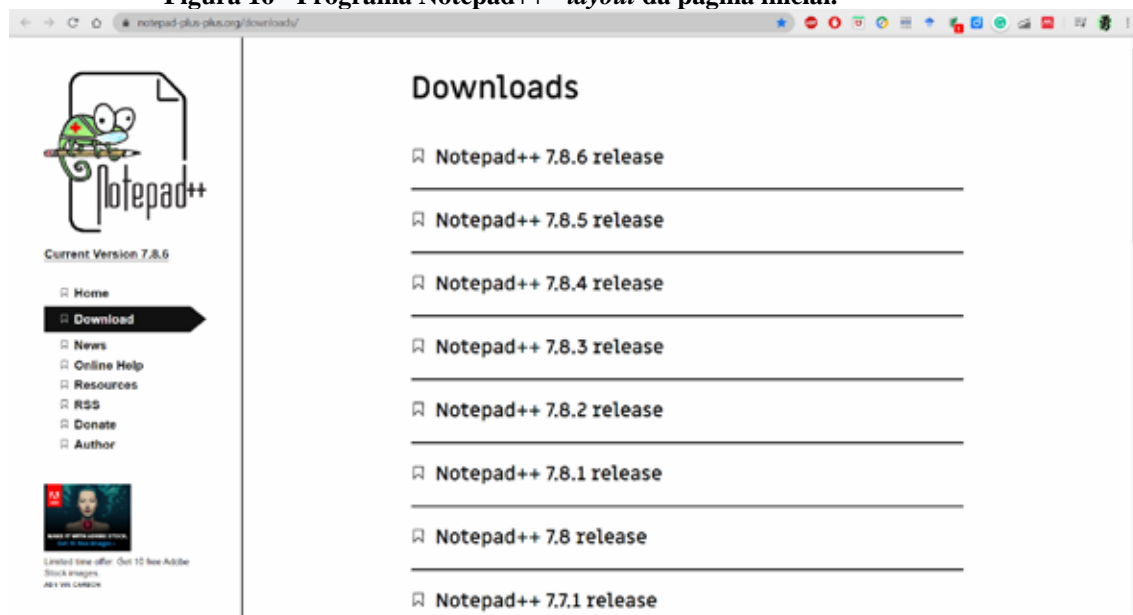
Na Figura 14 o cabeçalho aparece codificado em chinês, pois a codificação de entrada foi o UTF-16 LE, mas a plataforma do *ToGatherUp* fora concebida para receber arquivos somente em formato UTF-8 e insere o BOM automaticamente. Isto é, durante o processamento do *software* TUP⁶⁹ na plataforma, houve a confusão na decodificação explicada previamente (MCENERY; XIAO, 2005, p. 69) devido à diferença de codificação de entrada e de saída no percurso de mudança de plataformas de armazenamento.

Primeiramente, abrimos o arquivo a partir do *corpus* baixado para identificá-lo. Como a codificação do arquivo proveniente da plataforma era diferente da codificação do arquivo salvo no *OneDrive*, foi necessária a abertura do arquivo para identificá-lo por meio do título que havia abaixo do cabeçalho em chinês visível na Figura 15.

Na Figura 15 é possível visualizar a codificação do arquivo proveniente da plataforma do *ToGatherUp* em formato UTF-16 LE.

⁶⁹ O modelo TUP, do inglês *Technology, Usability and Pedagogy*, é um modelo de *checklist* usado para avaliar e selecionar um software educacional, por meio de uma abordagem interdisciplinar, englobando aspectos como tecnologia, usabilidade e pedagogia (PERFOLL JÚNIOR; MODRO, 2016, p. 3). http://www.ceavi.udesc.br/arquivos/id_submenu/787/ademar_perfoll_junior.pdf.

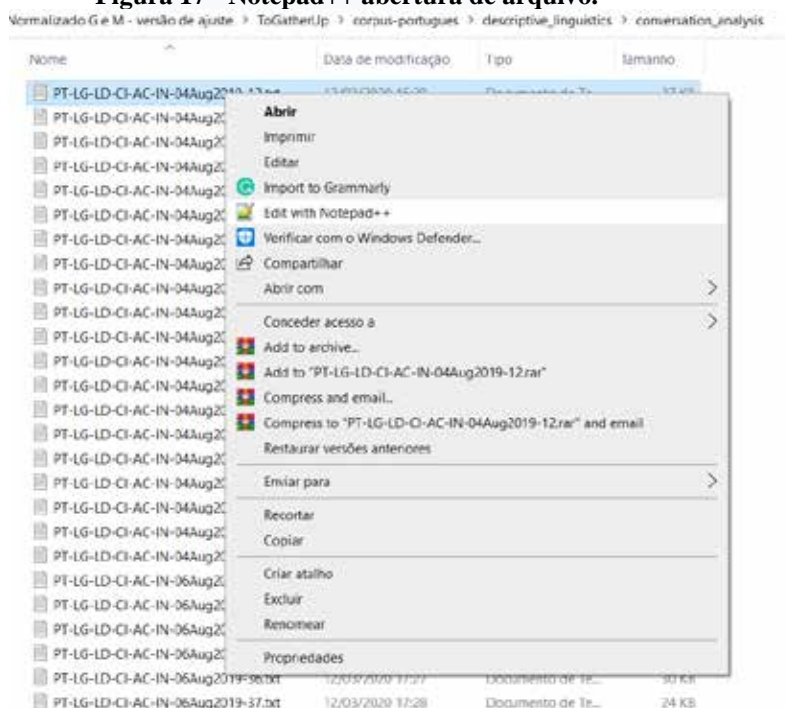
Figura 16 - Programa Notepad++ - *layout* da página inicial.



Fonte: Elaborada pelo autor.

Na Figura 16, visualizamos a página da internet que disponibiliza o Notepad++ para *download*. O Notepad++ abre o arquivo baixado do *ToGatherUp* e decodifica o cabeçalho corretamente, quando se usa o Notepad do Windows. Nas Figuras 13 e 14 descrevemos este processo.

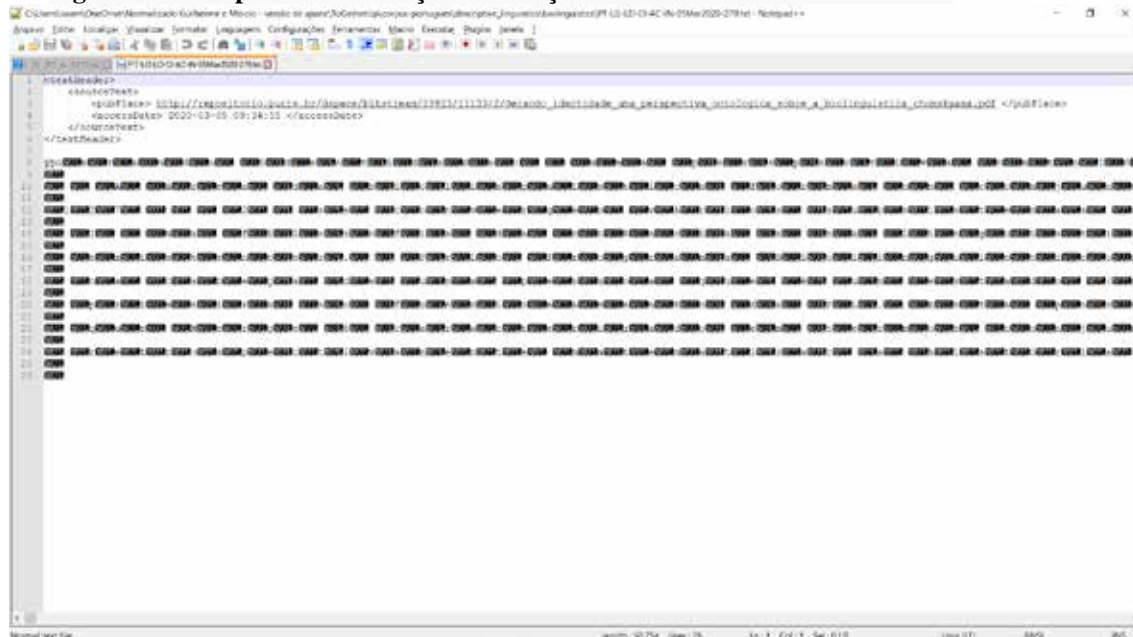
Figura 17 - Notepad++ abertura de arquivo.



Fonte: Elaborada pelo autor.

Após selecionar a opção de editar o arquivo com o Notepad++ (Figura 17), é possível ler o cabeçalho, tal qual configurado na plataforma do *ToGatherUp*. A parte do arquivo que corresponde ao texto aparece codificada, de forma ilegível, ao contrário do que acontece com o cabeçalho, conforme mostra a Figura 18.

Figura 18 - Notepad++ visualização do cabeçalho e do texto ANSI de base Unix.



Fonte: Elaborada pelo autor.

Para finalizarmos o processo de correção dos arquivos corrompidos, copiamos o cabeçalho do arquivo aberto no Notepad++ e o substituímos no lugar do cabeçalho em chinês do arquivo baixado do *ToGatherUp*. Em seguida o salvamos em formato UTF-8. Na base do Unix, na plataforma do *ToGatherUp*, a codificação equivalente ao UTF-8 do Windows é a codificação ANSI, como exposto no item 5.4.1, visível no canto inferior direito do rodapé da Figura 18.

O segundo tipo de arquivo corrompido apresentava o cabeçalho legível, mas o texto trazia espaçamento duplo ou triplo entre os caracteres, conforme se pode verificar na Figura 19.

Figura 19 – ToGatherUp: arquivo com texto corrompido pós-download de base Unix.

```

<sourceText>
  <pubPlace> http://www.letramento.iel.unicamp.br/publicacoes/artigos/aquisicao_de_linguagem_MariaSilvia.pdf
</pubPlace>
  <accessDate> 2019-08-10 00:59:26 </accessDate>
</sourceText>
</textHeader>

yba AQUISIÇÃO DE LINGUAGEM E PROCESSO DE SUBJETIVAÇÃO

  <Acessado em 07/11/11. Disponível em http://www.letramento.iel.unicamp.br/publicacoes/artigos/aquisicao_de_linguagem_MariaSilvia.pdf>

  Resumo: A partir da constatação de Leontiev de que é necessário o dar conta do princípio heurístico presente na aquisição da linguagem infantil, sem nos determos apenas nos aspectos algorítmicos, os quais vêm o domínio da linguagem de um ponto de vista estratégico, propomos a expansão do enfoque transcategorial, de forma a vir a abranger o aspecto pragmático inerente à linguagem. Chamamos a atenção, no entanto, para o fato de que não basta a enfatizar esse aspecto pragmático apenas do ponto de vista situacional, o que nos levaria a incorrer no mesmo erro implícito no princípio algorítmico. Se, por um lado, o aumento progressivo na extensão frasal pode ser visto do ponto de vista de um domínio sintático crescente, e não apenas do ponto de vista da expansão vocabular, por outro, essa relação transcategorial pode ser expandida, para vir a incluir na mesma visada a construção progressiva da subjetividade e da alteridade no universo cognitivo infantil. Palavras-chave: subjetividade; identidade; linguagem; pragmático; heurístico; transcategorial. Há quase trinta anos
  
```

Fonte: Elaborada pelo autor.

Para a correção deste problema de decodificação, abrimos o arquivo a partir do *corpus* baixado e mantivemos o cabeçalho legível. Em seguida, abrimos o arquivo original alocado no *OneDrive*, copiamos todo o conteúdo usando os atalhos Control+A, Control+C e Control+V no arquivo baixado. O arquivo foi salvo em formato UTF-8, conforme a Figura 20.

Figura 20 – ToGatherUp: arquivo com texto corrigido pós-download de base Unix ANSI para UTF-8.



Fonte: Elaborada pelo autor.

Na Figura 21 é possível visualizar o texto da Figura 20 corrigido e salvo em formato UTF-8. No canto inferior direito das imagens, contidas na Figura 21, há a codificação das fontes ANSI e UTF-8 e o sistema Unix, da plataforma do *ToGatherUp*.

Apesar do cumprimento desses passos, não nos foi possível corrigir o erro, pois a plataforma recebia o arquivo e inseria o cabeçalho em formato UTF-8 a partir da base Unix. Além da inserção do cabeçalho, a plataforma tentava transformar os arquivos de formato UTF-8 de base Windows, para arquivos UTF-8 de base Unix. Nesse percurso, ocorria a corrupção da codificação e posterior decodificação, quando baixados. Finalmente, concluímos que seria necessário alterar a formatação da plataforma para receber arquivos UTF-16 LE e fazer a inserção do cabeçalho. Essa formatação foi possível pela adição do comando `$string = mb_convert_encoding($string, "UTF-16LE")`. A partir dessa alteração, a plataforma passou a ter uma segunda formatação para recepção de arquivos: a de receber arquivos em formato UTF-16 LE, além da opção original de recepção de arquivos em formato UTF-8.

1.16. Normalização dos corpora

A normalização de *corpora* é parte do processo metodológico na criação de obras terminográficas. Por normalização, entende-se que os *corpora* sejam equilibrados em termos de **tamanho** e **tipo** e que suas características sejam as mesmas, isto é, número de *tokens*, gênero (artigos científicos, dissertações, teses, material instrucional e manuais de Linguística) e, no nosso caso, atendam às características listadas no Quadro 1 (p. 53). A esse ponto da pesquisa, já temos os *corpora* de manuais em inglês e português compilados e disponíveis para processamento linguístico pelo WST. Adiante, apresentamos a organização dos *subcorpora* de Linguística Descritiva em português na Figura 21.

Figura 21 – Linguística: pastas com o *corpus* em português.



Fonte: Elaborada pelo autor.

Na pasta da Figura 21 estão armazenados e organizados os arquivos que compõem as 29 subáreas da Linguística Descritiva elencadas. Como exposto anteriormente, cada subárea traz 500 mil palavras, partindo da **Análise da Conversação** até à **Terminologia**.

As áreas cuja compilação foi preciso iniciar do zero foram quatro: a **Etimologia**, a **Filologia**, a **Biolinguística** e a **Linguística Diacrônica**, sendo as duas últimas inseridas a partir desta pesquisa. Houve também outras áreas cujos *corpora* precisaram ser completados até o número de 500 mil itens. O dimensionamento detalhado das subáreas está elencado no **Apêndice D**.

Os **Apêndices D e G** trazem dados importantes sobre o *corpus* de Linguística/Português e serviram para mensurar o trabalho que já tinha sido desenvolvido e dimensionar o que teve que ser desenvolvido. As cores descrevendo a pré-análise dos *subcorpora* referem-se ao tamanho, em específico, de cada subárea da Linguística. As cores significam: (i) suficiente - cor verde; (ii) excessivo - cor vermelha, acima de 500 mil; (iii) inexistente - amarelo, foi necessário realizar a compilação do *corpus* da subárea do zero; e (iv) insuficiente - cor alaranjada, foi necessário completar o *corpus* da subárea até que atingisse o padrão da pesquisa (500 mil itens). O Quadro 11 apresenta as cores e os significados utilizados no **Apêndice D**.

Quadro 12 - Dimensionamento dos *subcorpora*: significados e cores

Avaliação da dimensão em relação ao padrão de 500 mil itens		Cores
(i)	Suficiente	
(ii)	Excessivo	
(iii)	Inexistente	
(iv)	Insuficiente	

Fonte: Elaborado pelo autor.

Esta pré-análise serviu ao propósito de dimensionar a quantidade de trabalho que foi desenvolvido para que o tamanho dos *corpora* fosse completado e equilibrado. Ademais, a análise permitiu que se visualizasse a padronização pela qual seriam submetidos os *corpora*.

Nos Apêndices D, E, F e G, a coluna 1 lista as subáreas como elas se encontram na árvore de domínio da Linguística (FROMM, 2018) e as áreas adicionais (**Biolinguística**, **Linguística Diacrônica**) seguem a descrição conforme mostra o Apêndice H. As novas informações sobre os *subcorpora* são: (i) **dimensão**; (ii) **codificação**; e (iii) **pré-análise** do *subcorpus* coletado em relação ao objetivo da pesquisa - 500 mil palavras/itens.

Quanto à **dimensão**, os números da segunda coluna descrevem o tamanho atual do *subcorpus*, referente a cada subárea, considerando-se que nosso objetivo foi padronizá-lo, de forma que tivessem aproximadamente 500 mil itens cada.

A terceira coluna da tabela traz a **codificação** do *subcorpus*, cujas informações são separadas pelo sinal de *underline*, abreviada da seguinte forma: **L_AL_PT_A_TXTS#**.

O primeiro código, a letra **L** maiúscula, corresponde à subárea da Linguística ou Linguística Descritiva, contrapondo com a subárea da Linguística Aplicada (LA), abaixo da grande área de Linguística Geral. Essas duas subáreas serão correspondentes na LP e na LI.

O segundo código corresponde à subárea da Linguística em foco, nesse caso a **Aquisição da Linguagem**, representada pela sigla **AL**. Em geral, recorreremos às abreviaturas, privilegiando as consoantes, seguindo o padrão de abreviação da LI, de forma a termos uma codificação mais concisa, ex.: **DL** para **Dialetologia**, **FNT** para **Fonética**, **LdC** para **Linguística de Contato** e **SM** para **Semântica**.

O terceiro código, **PT**, indica a língua dos *corpora*, podendo ser **PT** para a LP e **EN** para a LI. Essa codificação é importante também para o uso e a configuração do WST, quando da confecção das listas de palavras-chave. Em caso de divergência de línguas no *subcorpus* e na configuração do console, um erro acontece, de forma que a confecção da lista é impossibilitada devido à divergência das línguas.

O quarto código representa o **gênero textual** dos arquivos, quais sejam: **artigos científicos (A)**, **dissertação (D)**, **monografia (MG)**, **manuals (ML) de Linguística**; **material instrucional (MI)**, podendo subdividir-se em aula ou apostila/aula (**MIA**), **pôster (MIP)**, ou **livro (MIL)**.

O quinto código representa o número de cada texto, neste caso, grafado pelo símbolo #. No *subcorpus* organizado, essa codificação fica visível com o número dos arquivos/textos conforme a Figura 22.

Figura 22 - Visão parcial da codificação do *subcorpus* de Linguística em português: subárea da Aquisição da Linguagem.

eDrive > Normalizado Guilherme e Márcio - versão de ajuste > Port_Corpus de Linguística > Linguística > Aquisição da Linguagem

Nome	Status	Data de modific...	Tipo	Tamanho
L_AL_PT_A_TXT65	✓	13/06/2018 01:52	Documento de Te	56 KB
L_AL_PT_A_TXT66	✓	13/06/2018 01:54	Documento de Te	10 KB
L_AL_PT_A_TXT67	✓	13/06/2018 06:34	Documento de Te	109 KB
L_AL_PT_A_TXT68	✓	13/06/2018 05:08	Documento de Te	34 KB
L_AL_PT_A_TXT69	✓	13/06/2018 05:12	Documento de Te	26 KB
L_AL_PT_D_TXT2	✓	13/06/2018 02:04	Documento de Te	97 KB
L_AL_PT_D_TXT4	✓	13/06/2018 02:13	Documento de Te	265 KB
L_AL_PT_D_TXT33	✓	13/06/2018 03:34	Documento de Te	393 KB
L_AL_PT_D_TXT39	✓	12/06/2018 04:22	Documento de Te	455 KB
L_AL_PT_D_TXT57	✓	12/06/2018 22:54	Documento de Te	359 KB
L_AL_PT_D_TXT70	✓	13/06/2018 06:35	Documento de Te	399 KB
L_AL_PT_MG_TXT21	✓	13/06/2018 03:02	Documento de Te	55 KB
L_AL_PT_MG_TXT62	✓	13/06/2018 01:43	Documento de Te	132 KB
L_AL_PT_ML_TXT3	✓	13/06/2018 02:09	Documento de Te	211 KB
L_AL_PT_MIA_TXT15	✓	13/06/2018 02:54	Documento de Te	66 KB
L_AL_PT_MIL_TXT36	✓	13/06/2018 03:49	Documento de Te	752 KB
L_AL_PT_MIP_TXT27	✓	13/06/2018 03:22	Documento de Te	7 KB

Fonte: Elaborada pelo autor.

A Figura 22 traz a codificação da nomenclatura dos arquivos que seguirá o mesmo padrão para o português e para o inglês. Seu objetivo é facilitar a identificação e a descrição dos *corpora* dentro do escopo desta pesquisa. A princípio, não esperávamos encontrar tamanha diversidade de gêneros textuais, pois a concepção original do projeto previa somente artigos, dissertações e teses. Ocorre que, à medida que a limpeza e a normalização dos *corpora* foram feitas, encontramos outros gêneros, como pôsteres, aulas e resumos, algo vantajoso, já que permitiu a diversificação dos gêneros descritores dos vários discursos acadêmicos da Linguística.

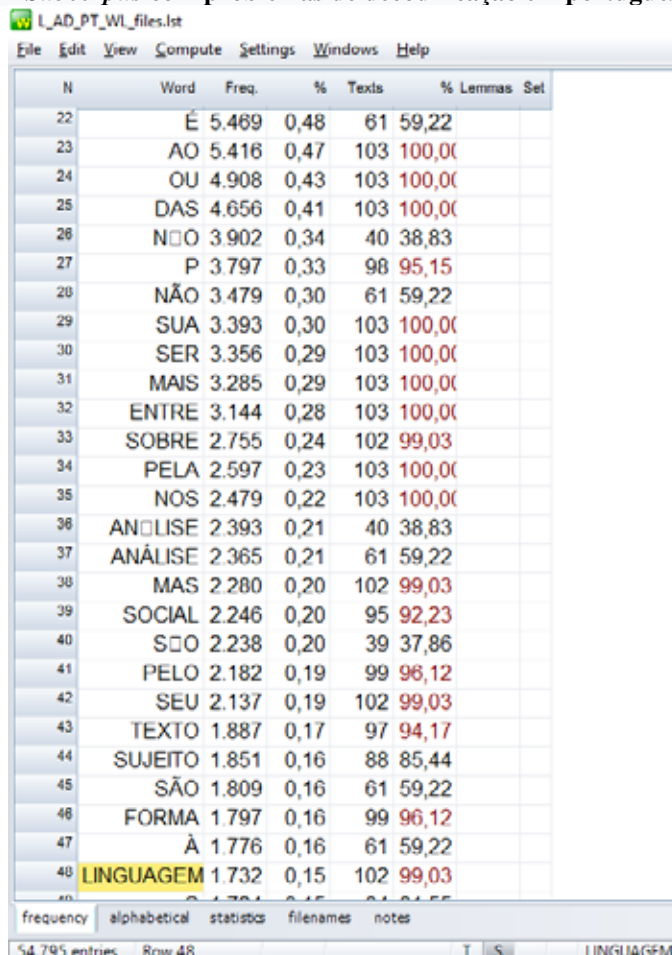
Nos Apêndices D e E, apresentamos o dimensionamento dos *corpora* iniciais de Linguística Descritiva e de LA em português; os Apêndices F e G trazem o dimensionamento dos *corpora* da LD e da LA em inglês, antes de serem padronizados.

1.1.24. Padronização do *corpus*

A busca por **padronizar** os *corpora* teve por objetivo limpar e deixar todos os arquivos com as mesmas características. Houve *corpora* cujas fontes não estavam codificadas de maneira uniforme e legível para que o WST em português fizesse a leitura. Dessa forma, a depender de como o participante do projeto havia salvado o arquivo, isso podia ou não interferir nas leituras feitas pelo WST (Unicode, ANSI, UTF etc.). Isto é, arquivos salvos em formato txt, cuja codificação escolhida foi ANSI, não eram decodificados corretamente quanto aos sinais diacríticos da língua portuguesa. Dessa forma, foi necessário converter a codificação dos arquivos de ANSI para Unicode. A codificação ANSI é a antiga ASCII (ISO 646), codificação

de 8-bits para línguas europeias, compatível com o padrão UTF-8 (MCENERY; XIAO, 2005, p. 61, 70). Essas inconsistências podem ser vistas na Figura 23, a seguir.

Figura 23 - Subcorpus com problemas de decodificação em português.



N	Word	Freq.	%	Texts	% Lemmas	Set
22	É	5.469	0,48	61	59,22	
23	AO	5.416	0,47	103	100,00	
24	OU	4.908	0,43	103	100,00	
25	DAS	4.656	0,41	103	100,00	
26	NAO	3.902	0,34	40	38,83	
27	P	3.797	0,33	98	95,15	
28	NAO	3.479	0,30	61	59,22	
29	SUA	3.393	0,30	103	100,00	
30	SER	3.356	0,29	103	100,00	
31	MAIS	3.285	0,29	103	100,00	
32	ENTRE	3.144	0,28	103	100,00	
33	SOBRE	2.755	0,24	102	99,03	
34	PELA	2.597	0,23	103	100,00	
35	NOS	2.479	0,22	103	100,00	
36	ANALISE	2.393	0,21	40	38,83	
37	ANALISE	2.365	0,21	61	59,22	
38	MAS	2.280	0,20	102	99,03	
39	SOCIAL	2.246	0,20	95	92,23	
40	SO	2.238	0,20	39	37,86	
41	PELO	2.182	0,19	99	96,12	
42	SEU	2.137	0,19	102	99,03	
43	TEXTO	1.887	0,17	97	94,17	
44	SUJEITO	1.851	0,16	88	85,44	
45	SAO	1.809	0,16	61	59,22	
46	FORMA	1.797	0,16	99	96,12	
47	A	1.776	0,16	61	59,22	
48	LINGUAGEM	1.732	0,15	102	99,03	

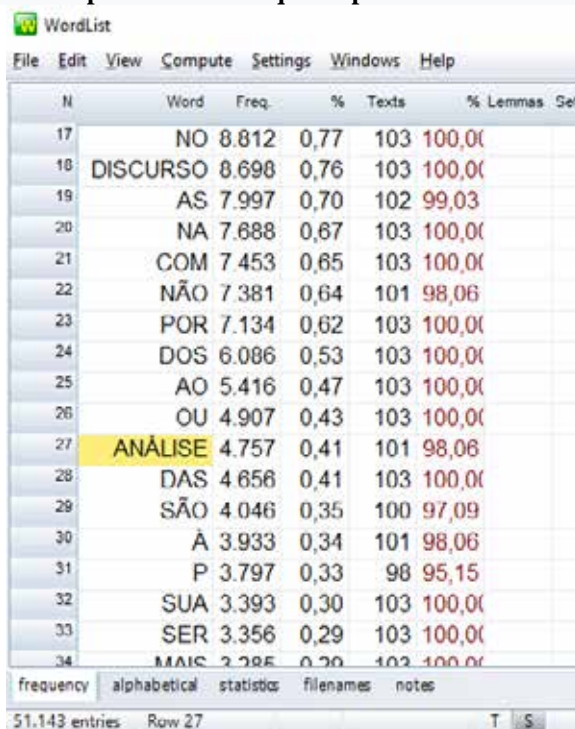
Fonte: Elaborada pelo autor.

A Figura 23 mostra os erros de decodificação: observemos as linhas 36 e 37, que trazem a mesma palavra, mas uma decodificada corretamente e a outra decodificada incorretamente para a língua portuguesa. Nesse caso, a dificuldade para o programa foi a decodificação da palavra com o acento agudo. Em contrapartida, os exemplos de número 26 e 40 trazem as palavras **NÃO** e **SÃO** decodificadas incorretamente, pois o til (~) não foi decodificado. Lembramos que essa não é uma limitação do WST, mas sim uma questão de procedimento metodológico quanto à forma de salvamento e codificação dos arquivos.

Para que essas imperfeições de decodificação fossem sanadas, o utilitário **Conversor de textos** (*Text converter*) foi utilizado. Carregamos os arquivos originais no WST, abrimos o *Text Converter* e salvamos os arquivos em outra pasta (BERBER SARDINHA, 2009).

Concluído o procedimento de conversão dos textos, processamos outra *lista de palavras* pelo WST. Dessa vez, observamos que os caracteres acentuados e as cedilhas estavam codificados corretamente.

Figura 24 - Lista de palavras com arquivos padronizados: Unicode.



N	Word	Freq.	%	Texts	% Lemmas	Set
17	NO	8.812	0,77	103	100,00	
18	DISCURSO	8.698	0,76	103	100,00	
19	AS	7.997	0,70	102	99,03	
20	NA	7.688	0,67	103	100,00	
21	COM	7.453	0,65	103	100,00	
22	NÃO	7.381	0,64	101	98,06	
23	POR	7.134	0,62	103	100,00	
24	DOS	6.086	0,53	103	100,00	
25	AO	5.416	0,47	103	100,00	
26	OU	4.907	0,43	103	100,00	
27	ANÁLISE	4.757	0,41	101	98,06	
28	DAS	4.656	0,41	103	100,00	
29	SÃO	4.046	0,35	100	97,09	
30	À	3.933	0,34	101	98,06	
31	P	3.797	0,33	98	95,15	
32	SUA	3.393	0,30	103	100,00	
33	SER	3.356	0,29	103	100,00	
34	MAIS	2.285	0,20	102	100,00	

frequency alphabetical statistics filenames notes
51.143 entries Row 27 T S

Fonte: Elaborada pelo autor.

Comparadas as duas listas de palavras, observamos, na Figura 24, a diferença no número de entradas e a decodificação correta dos itens, sendo a segunda *lista de palavras* menor que a primeira. Essa diferença se justifica pelo fato de que as entradas na primeira lista eram maiores, pois o WST contava as entradas com acento e cedilhas e as semelhantes, não decodificadas corretamente, como entradas distintas.

Outra característica padronizada foi o registro da fonte dos *corpora*. Essa ação se fez necessária porque havia subáreas cujos *corpora* não traziam suas fontes. Por consequência, foi necessário realizar a busca da fonte na internet e o registro dela (*link* do artigo na *web*) e da data de coleta. Nesse caso, a data da coleta foi registrada como a data da busca do *link* do arquivo, já em 2018.

Adiante, tratamos dos procedimentos de padronização desses *corpora*.

1.1.25. Padronização do tamanho dos *corpora*

Na primeira fase do dimensionamento dos *corpora*, havia três tipos de *corpora*: os *subcorpora* de aproximadamente 500 mil itens, os de tamanho superior e os de tamanho inferior. Em busca da redução dos superiores, adotamos dois procedimentos para reduzi-los: (1) primeiro, o de caráter quantitativo e (2) segundo o de caráter qualitativo, objetivando sua padronização.

O primeiro procedimento, o de caráter quantitativo, consistiu em estabelecer a média de arquivos nas pastas equilibradas (500 mil itens) das subáreas da Linguística Descritiva e da Linguística Aplicada. Dentre as subáreas já compiladas e equilibradas, a média era de 55 textos. Logo, esse foi o número almejado de textos nas subáreas a serem equilibradas. Foi interessante chegar a esse dado, sabendo-se que a maioria dos gêneros era composta de artigos científicos e que a extensão de artigos é, em geral, padronizada para publicação em eventos e revistas com uma média de 20-30 páginas.

A variável desse padrão quantitativo aconteceu quando no *subcorpus* havia uma **quantidade maior ou menor de dissertações e teses (variável 1)**. Como esses gêneros são mais extensos que os artigos científicos e não têm uma dimensão padronizada, a presença deles interferiu na necessidade de aumentar ou diminuir a dimensão do *subcorpus*.

O segundo procedimento, de caráter qualitativo, consistiu na escolha dos arquivos nos quais houve a menor recorrência de palavras-chave no *subcorpus* em questão. Para conduzir a redução de arquivos, utilizamos a ferramenta *Plot*⁷¹ do concordanciador do WST. Para acessá-la, foi necessário fazermos primeiramente (1) a *lista de palavras*, e (2) em seguida, a *lista de palavras-chave*. Então, (3) selecionamos a palavra-chave proveniente da *lista de palavras-chave* mais recorrente, (4) abrimos o concordanciador e, finalmente, (5) fizemos a visualização do *Plot* daquela palavra-chave para selecionar os arquivos a serem eliminados. Explicitaremos esse procedimento em detalhes adiante.

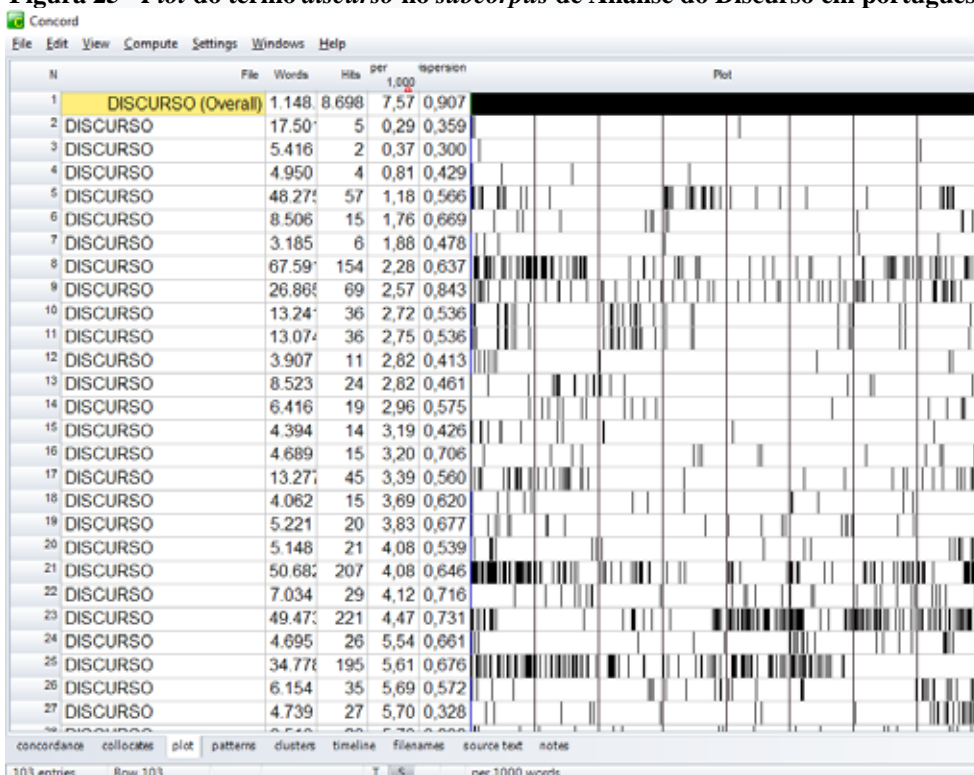
A partir da *lista de palavras-chave*, selecionamos as 10 palavras-chave mais recorrentes no *subcorpus*. Em seguida, selecionamos a primeira palavra-chave e acessamos o concordanciador. A partir dele, acessamos o *Plot*, ferramenta que representa graficamente as ocorrências do termo dentro dos textos selecionados. Em seguida, selecionamos a coluna **per 1000** e fizemos a lista dos dez textos nos quais as palavras-chave recorriam menos, isto é, textos

⁷¹ Inicialmente, o *plot* foi criado para identificar a porcentagem de palavras-chave no texto. No entanto, nesta pesquisa, os dados utilizados se referem à porcentagem da frequência de um texto específico no *corpus* de uma subárea da Linguística. A razão pela qual o *plot* pode ser aplicado nesse caso é porque ele pode identificar a frequência de palavras-chave em um determinado texto do *corpus*, quer seja alta ou baixa. Conforme mostrado na Figura 25, na quarta coluna, ele contém dados do termo **discurso** em cada 1000 itens no *corpus* de AD.

nos quais as palavras-chaves tinham uma **densidade lexical menor (variável 2)**. Com um clique duplo sobre o termo acessamos os códigos dos textos, anotamos os números deles e repetimos o procedimento com as outras palavras-chave (BERBER SARDINHA, 2009; SCOTT, 2016).

Vejamos a janela do *Plot* para o termo **discurso** no *subcorpus* de Análise do Discurso na Figura 25.

Figura 25 - Plot do termo *discurso* no *subcorpus* de Análise do Discurso em português.



Fonte: Elaborada pelo autor.

Na Figura 25, observamos que o termo **discurso** recorre menos frequentemente em alguns textos (números 2, 3, 4, 7 etc.) que em outros. Anotamos o número dos textos nos quais o termo era menos recorrente e repetimos o mesmo procedimento com as nove palavras-chave restantes da *lista de palavras*, visíveis na Figura 26.

Figura 26 - Lista de palavras-chave da Análise do Discurso (parcial – português).

N	Key word	Freq	%	Texts	RC. Freq
1	DISCURSO	4.607	0,88%	58	125.594
2	ANÁLISE	2.622	0,50%	58	289.316
3	LINGUAGEM	1.141	0,22%	58	56.492
4	DISCURSOS	733	0,14%	51	24.003
5	LINGÜÍSTICA	552	0,11%	45	9.958
6	SUJEITO	845	0,16%	52	58.362
7	TEXTO	1.057	0,20%	56	157.323
8	TEXTOS	650	0,12%	52	45.410
9	RACISMO	407	0,08%	5	10.329
10	IDEOLOGIA	432	0,08%	41	16.341
11	SENTIDOS	424	0,08%	50	17.068
12	CORPUS	330	0,06%	25	7.812
13	SENTIDO	939	0,18%	58	193.292
14	LÍNGUA	548	0,11%	46	50.049
15	INTERDISCURSO	159	0,03%	25	243
16	ACD	164	0,03%	7	404
17	QUE	13.336	2,56%	58	12...63
18	RELAÇÕES	771	0,15%	54	149.615
19	ENUNCIÇÃO	219	0,04%	39	2.562
20	MÍDIA	420	0,08%	24	31.748

Fonte: Elaborada pelo autor.

Na Figura 26, é possível verificar as palavras-chave que serviram ao propósito da redução do *subcorpus* da Análise do Discurso. Os textos nos quais esses termos eram menos recorrentes foram escolhidos para a primeira seleção que nos levaria à escolha dos textos a serem eliminados do *subcorpus*. A essa lista, denominamos **lista de baixa densidade terminológica**⁷², visto que, nela, a recorrência das principais palavras-chave da subárea foi menor.

Em seguida, chegamos a uma lista de textos e, dentre esses arquivos, observamos que havia alguns nos quais as palavras-chave eram mais recorrentes que em outros. Anotamos o número desses textos e fizemos uma segunda lista: a lista que marcaria os textos a serem eliminados. Essa lista foi denominada **lista redutora** e a exigência era a de que seus textos tinham que ter uma recorrência mínima de duas vezes na *lista de palavras-chave* selecionadas (neste caso, dez). O objetivo final foi o de atingir a média de 55 arquivos (dependendo da

⁷² Usamos a expressão densidade terminológica de Barros (2004, p. 191), para designar textos nos quais as palavras-chave daquela subárea foram mais frequentes. Originalmente, o conceito foi refinado por Halliday (1985, p. 64), na LC, representando a proporção das palavras lexicais sobre as palavras gramaticais de um texto.

subárea – **variável 1** - e da consistência terminológica da seleção – **variável 2**) e um total de aproximadamente 500 mil itens.

Esse procedimento de adequação do dimensionamento dos *subcorpora* foi adotado para todas as subáreas da Linguística que excederam o limite de 500 mil itens. Contudo, a redução do número de itens não garantiu o balanceamento dos *corpora*, já que eles não estavam limpos.

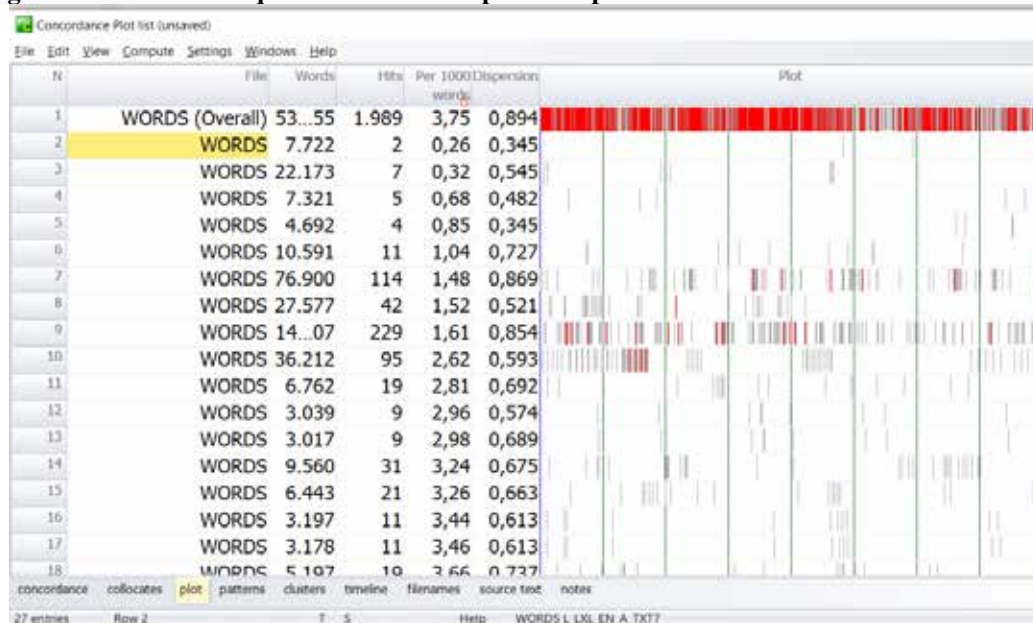
Como conclusão desta primeira fase de normalização dos *corpora*, podemos chegar a alguns conceitos novos quanto a esse procedimento. Os conceitos são das **variáveis** e a **nomenclatura das listas** para exclusão de arquivos.

A **Variável 1** é aquela na qual a quantidade de textos de um *subcorpus* pode ser maior ou menor, dependendo dos gêneros textuais que o compõem. A **Variável 2** refere-se ao perfil dos textos a serem eliminados, dependendo da densidade terminológica do texto. Isto é, quanto menor for a densidade terminológica do texto, maior será a probabilidade de esse texto ser excluído do *subcorpus* que ele compõe.

Incluso no conceito da **Variável 2** ou da densidade terminológica do texto, nós teremos duas listas: (i) a **lista de baixa densidade terminológica**, na qual elencamos os textos com menor recorrência das principais palavras-chave de uma subárea, candidatos à exclusão; e (ii) a **lista redutora**, isto é, a lista que marca os textos selecionados para exclusão do *subcorpus*, devido à *baixa densidade terminológica*.

Adiante, exemplificamos como este procedimento foi conduzido no *subcorpus* de Lexicologia em LI. Feitas as listas de palavras e de palavras-chave, selecionamos as dez palavras-chave mais recorrentes (maior frequência e chavicidade). A palavra-chave mais recorrente na lista foi *Words*; assim, abrimos o item *Words* no concordanciador, clicamos na coluna **per 1000** e selecionamos os dez textos nos quais a palavra *Words* teve menor recorrência. Ao selecionarmos a linha da palavra-chave, o nome do arquivo é disponibilizado no rodapé da janela dessa ferramenta, conforme mostra a Figura 27.

Figura 27 - Nome do arquivo identificado a partir da palavra-chave *Words* no concordanciador.



Fonte: Elaborada pelo autor.

A Figura 27 mostra a linha dois selecionada e o nome do txt 17 codificado (arquivo L_LXL_EN_A_TXT17, no rodapé da imagem). Em seguida, construímos um quadro com o número dos dez textos e as 10 palavras-chave que compuseram a lista de *baixa densidade terminológica* (por exemplo, *words* na Figura 27).

As 10 palavras-chave mais recorrentes no *subcorpus* de Lexicologia foram: *words*, *thesaurus*, *meaning*, *language*, *term*, *word*, *equivalence*, *dictionary*, *senses* e *expressions*. Estas palavras-chave foram mais recorrentes nos textos 21, 25, 14 e 17 dentre os 100 elencados.

Após a eliminação dos textos ou arquivos excedentes, fizemos a *lista de palavras* no WST novamente para adequar o tamanho do *subcorpus* ao tamanho padrão. Esse procedimento foi repetido até que houvesse aproximadamente 500 mil itens nesse *subcorpus*.

As variáveis previamente explicadas só se aplicam em contextos em que o pesquisador precisa organizar *corpora* desequilibrados e busca a redução de um *subcorpus* que excede o padrão de dimensionamento escolhido para a condução da pesquisa.

A abordagem descrita acima trata da diminuição do tamanho dos *corpora*, a **Fase 1**. Outra fase do procedimento de padronização, a **Fase 2**, é o aumento dos *corpora* menores, ou mesmo a compilação a partir do zero para aqueles inexistentes.

Nesse estágio da pesquisa, as subáreas que ofereceram perfil peculiar foram a da Ecolinguística e a da Etimologia, pois, após a busca por artigos, dissertações e teses, houve uma característica paradoxal entre essas subáreas. A subárea de Ecolinguística cresceu muito quando comparada à primeira fase do projeto de compilação (2010-2016), nas pesquisas de graduação

e de pós-graduação. Em contrapartida, a subárea da Etimologia tem tido pouca produção científica.

Na primeira fase da pesquisa, na subárea da Ecolinguística, a bibliografia era de menor representatividade numérica, com a maioria dos trabalhos de autoria do professor Hildo Honório do Couto (UnB). No segundo momento de compilação, foi possível observar uma ampliação no número de dissertações e teses nessa subárea e a presença de outros profissionais conduzindo programas de pós-graduação, inclusive a existência de uma revista, a *Ecolinguística: Revista Brasileira de Ecologia e Linguagem*⁷³. Consequentemente, é uma área em franco crescimento no cenário atual da Linguística no Brasil. Também vale ressaltar que essa subárea é bem difundida no exterior, em países como a Inglaterra, a Alemanha e os Estados Unidos.

Diferentemente, a subárea da Etimologia, hoje representada no Brasil pelo professor Mario Viaro (USP), apresentou um número limitado de artigos, dissertações e teses no país. Essa escassez foi visível pelo número de páginas disponibilizadas (17) no *Google*, quando usamos a descrição *Linguística*Etimologia*artigos: pdf*. Para esta pesquisa, exaurimos as páginas disponibilizadas, sem que atingíssemos o alvo de 500 mil itens. Isso nos fez explorar mais a fundo as opções de páginas com a variação de dissertações e teses. Por fim, o fator que contribuiu para que completássemos esse limite foi a presença de teses e dissertações em português europeu.

Em primeiro lugar, para que atingíssemos o limite mínimo de 500 mil itens para a subárea da Etimologia, fizemos a pesquisa padrão, usando as descrições *Linguística*Etimologia*teses/dissertações/artigos:pdf* ou *Etimologia*teses/dissertações/artigos:pdf*. O resultado do número de páginas do Google foi satisfatório, mas nem todos os documentos eram da área da Etimologia como subárea da Linguística. Nesse contexto, houve textos acadêmicos que tratavam da etimologia de termos dentro de outras áreas da Ciência, sem o rigor metodológico proposto pela Etimologia.

Em segundo lugar, foi necessário baixar os arquivos e buscar as palavras-chave contidas nele. Finalmente, caso o arquivo não trouxesse Etimologia como palavra-chave nos resumos, buscamos o termo no arquivo, usando o atalho *Control+F* e digitando o termo Etimologia. Após essa busca, fizemos a leitura dos contextos nos quais o termo ocorria para verificar se o arquivo estaria adequado ao propósito desta pesquisa a fim de prosseguirmos com os próximos passos, a saber, a integração desse arquivo ao nosso *subcorpus* de Etimologia.

⁷³ Revista disponível em <http://periodicos.unb.br/index.php/erbel>. Acesso em: 10 jul. 2018.

Durante os procedimentos descritos anteriormente, foi possível constatar que muitos artigos do gênero acadêmico usam o termo etimologia para explicar termos de sua área, como o Direito e a Pedagogia. A pesquisa padrão, o *download* dos arquivos, a busca da palavras-chave e a leitura dos contextos fizeram com que a busca por esse *subcorpus* fosse mais demorada comparada com os procedimentos tomados para os demais *subcorpora*. Isso não só demonstra a escassez de estudos conduzidos na área de Etimologia no Brasil, mas também pode demonstrar uma queda nos estudos dessa subárea no Brasil.

Na próxima seção, detalharemos como os processos de limpeza e padronização dos *subcorpora* foram conduzidos.

1.1.26. Limpeza dos *corpora*

Após a redução ou aumento dos *subcorpora*, representando cada subárea da Linguística Descritiva, procedemos à limpeza dos textos. O resultado obtido foi que o *subcorpus* foi reduzido em tamanho e ficou abaixo do padrão estabelecido. Consequentemente, na segunda fase da padronização, uma nova coleta de textos foi feita, a limpeza foi realizada e os dados foram agrupados aos *subcorpora* da pesquisa.

A limpeza consistiu na retirada de fragmentos de textos ou informações que continham as seguintes características:

1. Nomes de eventos: Anais do VII Congresso Internacional da Abralín, Curitiba, 2011;
2. Citações em língua estrangeira;
3. Agradecimentos;
4. Limpeza dos trechos falados em artigos, teses e dissertações, presentes na subárea Análise da Conversação;
5. Dados dos quadros dos textos;
6. Resumos em língua estrangeira (*abstract*, *resumen*, *résumé*);
7. Biografias e depoimentos;
8. Artigos em LE.

A limpeza dos *corpora* foi conduzida pelo uso de três programas: o **Bloco de Notas**, o *Word* e o *WST*. O **Bloco de Notas** foi usado na primeira etapa, ao padronizarmos a estrutura dos arquivos da seguinte forma: título do texto, metadados: data e endereço de acesso (*link*), e o corpo do texto sem as referências bibliográficas ou bibliografia. Quanto ao *link* do arquivo original, foi necessário fazer uma revisão, já que muitos haviam expirado ou mudado de endereço. Neste caso, foi necessário fazer uma segunda busca do arquivo, colocando-se o título do texto entre aspas no Google. Algumas mudanças encontradas foram: (i) o endereço da revista, (ii) a taxonomia do endereço dos repositórios de dissertações e teses das bibliotecas

universitárias e, (iii) em outros casos, os arquivos só se encontravam disponíveis em plataformas como Academia.edu e *Research Gate*, disponibilizados pelos autores.

O segundo programa utilizado foi o *Word* para a correção ortográfica dos textos que não foram decodificados corretamente pelo *Acrobat Reader*. Neste caso corrigimos as palavras e eliminamos os espaços extras que havia entre os caracteres. Tanto no *Word* quanto no **Bloco de Notas** usamos a opção **Substituir** para a limpeza dos metadados dos artigos e das revistas, tais como título e autor dos artigos, nomes, data de publicação, volume e número das revistas.

Finalmente, usamos o conversor de textos (*Text Converter*), ferramenta do WST para padronizar a codificação dos arquivos do formato txt para Unicode, de forma que a leitura do *corpus* não fosse prejudicada quando da confecção das listas de palavras. Isso porque, com um *corpus* sem padronização, o número de palavras sofre alteração, pois os itens que levam os acentos agudo, grave e circunflexo e a cedilha não são decodificados como *tokens*, e, com isso, a contagem final de palavras diminui. Outra vantagem do uso dessa ferramenta de padronização é a remoção das quebras de linhas, uniformizando o formato dos textos de uma só vez, sem que isso fosse feito manualmente de forma individualizada.

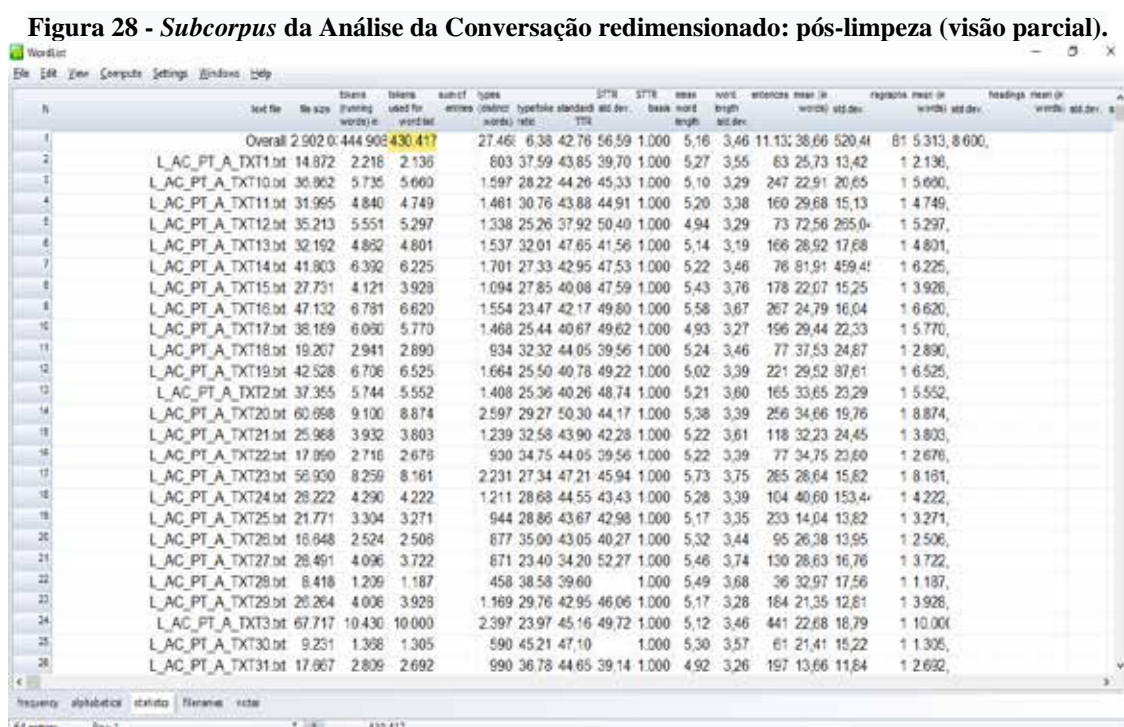
A checagem da validade do endereço de cada arquivo separadamente foi outro procedimento necessário quando da limpeza dos *corpora*. Este procedimento foi necessário, pois parte dos *corpora* foi coletado de 2010 a 2015, bem antes, portanto, da normalização final dos dados realizada entre 2018 e 2019. Devido a este considerável espaço de tempo entre a coleta dos dados e a respectiva limpeza final, vários *links* foram atualizados, teses e dissertações foram alocadas para outros servidores e novas URL disponibilizadas.

Para que localizássemos o novo endereço dos arquivos, primeiramente copiamos o *link* registrado e verificamos sua validade. Em segundo lugar, caso o endereço estivesse inválido, copiamos trechos aleatórios do texto, inserimo-los entre aspas e refizemos a busca do novo endereço do texto no *Google*. Do contrário, caso o *link* antigo direcionasse para um novo *link*, o que normalmente é mais rápido, conferimos se o texto era o mesmo e depois realizamos uma cópia e o respectivo salvamento do arquivo em nosso banco de dados.

As alterações existentes foram: (1) as URL de artigos foram atualizadas nas páginas dos cursos de pós-graduação para revistas acadêmicas; (2) as teses ou dissertações foram alocadas para o site do domínio público: <http://www.dominiopublico.gov.br/pesquisa/PesquisaObraForm.jsp> ou para a plataforma do *Scielo*, artigos foram realocados para os *sites* da Academia.edu ou *Research Gate* e, finalmente, outras URL caíram em desuso, tornando o arquivo indisponível na internet.

Quando os endereços dos arquivos haviam sido alterados, e as URL estavam indisponíveis, bem como o arquivo em si, foi necessário descartar o arquivo coletado e fazer nova compilação.

A Figura 28 traz exemplos com dados estatísticos do *subcorpus* da Análise da Conversação, mencionado na seção anterior.



Fonte: Elaborada pelo autor.

Os dados do *subcorpus* da Análise da Conversação estão parcialmente descritos após os dois procedimentos da normalização: o redimensionamento dos arquivos e a limpeza dos dados. A dimensão do *subcorpus* era de 430.417 itens. Nesse caso, foi necessário limpar alguns arquivos da *lista redutora* que haviam sido eliminados e retorná-los ao *subcorpus* a fim de mantermos a padronização de 500 mil itens por arquivo

Nesse procedimento de aumento do *subcorpus*, estabelecemos dois parâmetros: (i) a *dimensão do arquivo* e (ii) a *densidade terminológica* do arquivo. Quanto à *dimensão*, buscamos fazer uma relação de dimensão dos arquivos com a alteração do número de itens. Se a necessidade era de um número grande de itens, buscamos resgatar os arquivos maiores e vice-versa. O segundo parâmetro, o da *densidade terminológica*, serviu para a escolha do arquivo, já que aquele que apareceu menos vezes nas listas de *baixa densidade terminológica* seria mais útil para a extração de conceitos e, provavelmente, para a consequente definição das entradas do protótipo de Vocabulário de Linguística.

Quanto à lista de *baixa densidade terminológica*, lembramos que foram feitas no mínimo dez, com vinte componentes, sendo que dez era o número de palavras-chave que levaram à análise da plotagem (*Plot*). Já o número vinte era a média do número de arquivos que excediam a média dos arquivos nos *subcorpora*, antes de serem equilibrados.

Nos Apêndices D - K, apresentamos o dimensionamento dos *corpora* das subáreas da Linguística, em português e em inglês, antes e depois de serem padronizados.

1.1.27. Linguística Descritiva – dimensionamento dos *corpora* em português e em inglês

Os apêndices D e F trazem os dados dos *subcorpora* da LD em língua portuguesa (LP) e em língua inglesa (LI). A denominação *inicial* serve para identificar os *subcorpora* antes de serem padronizados. Em contrapartida, os *subcorpora* denominados “final” identificam os *subcorpora* já padronizados.

Como podemos observar, os *corpora* são das subáreas com arquivos acima de 500 mil itens. É relevante chamar a atenção para a questão do número de itens da subárea de **Linguística Matemática** (LM), cujos itens ficaram abaixo de 500 mil, pois foi a única subárea da LD em LP que não conseguiu ter produção acadêmica para atingir o padrão de 500 mil.

1.1.28. Linguística Aplicada – dimensionamento dos *corpora* em português e em inglês

Os apêndices E, G, I, K e M trazem os dados dos *subcorpora* da LA em língua portuguesa (LP) e em língua inglesa (LI). As denominações “inicial” e “final” seguem o mesmo padrão de uso da LD.

Além da normalização dos *corpora*, foi necessário considerar o aspecto do arquivamento desses *corpora* para formar um banco de dados. Nesta pesquisa optamos pela plataforma *ToGatherUp*, descrita na Seção 4.3.2.

1.17.Extração e tratamento de dados do corpus: WST

Nesta seção, abordamos a condução da extração e do tratamento do *corpus* de pesquisa pelo WST. É importante lembrar que esta pesquisa dispõe de um *corpus* de Linguística, subdividido em 3 *subcorpora*, quais sejam: (i) um *subcorpus* de LD; (ii) um *subcorpus* de LA; e um *subcorpus* de manuais de Linguística. Outro fator importante a ser considerado é que a língua de partida para as análises de pesquisa é a LP, haja visto que o público-alvo é composto por alunos iniciantes do Curso de Letras brasileiros.

Na Tabela 4, descrevemos o dimensionamento de cada *subcorpus* em LP, explicitando o número de textos e dados estatísticos de *tokens* e *types*. Esses *subcorpora* são fruto do processo de redimensionamento do *corpus* de Linguística, provenientes do projeto de pesquisa inicial de Fromm (FROMM; YAMAMOTO, 2013), descrito na seção 5.2. As listas de palavras foram confeccionadas com a aplicação da *stoplist* e da *lemma list*⁷⁴ da Linguística e das subáreas, unindo os termos com e sem trema. Nesse dimensionamento há o *corpus* de Linguística, o qual contém o *subcorpus* de LD, de LA e o de manuais de Linguística.

Tabela 4 – Dimensionamento dos subcorpora de Linguística em LP.

	Linguística	Linguística Descritiva	Linguística Aplicada	Manuais de Linguística em LP
Número de textos	1665	1082	570	13
Tokens	24.597.712	14.661.999	8.856.847	1.089.748
Types	356.983	285.969	154.366	44.407
Type/token ratio (TTR)	1,45%	1,95%	1,74%	4,07%

Fonte: Elaborada pelo autor.

Os dados da Tabela 5 são provenientes das listas de palavras da Linguística. O *corpus* de Linguística corresponde a soma total nos textos provenientes da LD, da LA e dos manuais de Linguística. O *corpus* de linguística serviu à extração das palavras chaves que foram definidas na área da linguística, conforme listadas na Figura 29. Nesta pesquisa, listamos os 10 termos mais chaves, de acordo com o padrão de chavicidade BIC⁷⁵ e definimos o termo *língua* no VoBLing. Os subcorpora de LD e LA receberam o mesmo tratamento, porém, a partir deles, selecionamos 5 ST para serem listados e definimos

A lista de palavras-chave da Linguística e de suas 2 subáreas são apresentadas conforme a Figura 29 adiante. Elas são utilizadas para mostrar o processo empregado na seleção de palavras-chave definidas no VoBLing. Não há necessidade de criar uma lista de palavras-chave para os manuais, porque eles são úteis apenas para procurarmos contextos definitórios e

⁷⁴ Esta lista de itens lematizados foi confeccionada com os ST da linguística e das subáreas de LD e La, grafados com trema e sem trema, com as variações de linguística, biolinguística, ecolinguística, neurolinguística e sociolinguística.

⁷⁵ BIC score - uma alternativa ao valor de p. O BIC usa a pontuação de probabilidade do log e o tamanho dos dois *corpora* (de pesquisa e de referência) em sua fórmula. O valor de p em 0,1 é utilizado nas pontuações do BIC para avaliar a chavicidade. As pontuações do BIC podem ser interpretadas assim: abaixo de 0 -> não confiável; 0-2 -> vale uma menção; 2-6 -> evidência positiva; 6-10 -> Forte; mais do que 10 -> muito forte (SCOTT, 2020; https://lexically.net/downloads/version8/HTML/keywords_calculate_info.html?anchor=bic_score).

explicativos quando não há o número suficiente de contextos nos *subcorpora* de LD e LA. Lembramos que esta busca é realizada pela ferramenta Concordanciador.

Figura 29 - Listas de palavras-chave da Linguística, LD e LA: pré-limpeza (visão parcial).

R	Key word	Freq.	% Texts	BIC
1	LÍNGUA	85.424	0,34%	1.527
2	LÍNGUAS	30.343	0,12%	1.547
3	LINGÜÍSTICA	20.357	0,08%	795
4	LINGÜAGEM	36.493	0,14%	1.436
5	PORTUGUÊS	26.080	0,10%	1.093
6	TRADUÇÃO	20.602	0,08%	719
7	PALAVRAS	29.247	0,11%	1.418
8	VERBO	11.993	0,05%	633
9	TEXTO	32.728	0,13%	1.229
10	GRAMÁTICA	11.290	0,04%	883
11	FALA	22.381	0,09%	1.226
12	LINGÜÍSTICA	12.458	0,05%	830
13	ESCRITA	16.695	0,07%	1.025
14	TEXTOS	17.561	0,07%	1.067
15	LINGÜÍSTICO	7.294	0,03%	630
16	LEXICAL	8.142	0,03%	794
17	SEMÂNTICA	8.960	0,04%	727
18	FALANTE	8.378	0,03%	863
19	FALANTES	8.102	0,03%	909
20	VERBOS	7.620	0,03%	532

R	Key word	Freq.	% Texts	BIC
1	LÍNGUA	43.137	0,28%	975
2	LINGÜÍSTICA	20.119	0,13%	501
3	LINGÜAGEM	23.175	0,15%	939
4	LÍNGUAS	14.677	0,10%	688
5	PORTUGUÊS	16.679	0,11%	693
6	VERBO	9.242	0,06%	443
7	PALAVRAS	19.322	0,13%	939
8	FALA	15.130	0,10%	829
9	GRAMÁTICA	7.179	0,05%	577
10	LEXICAL	5.858	0,04%	513
11	VERBOS	5.905	0,04%	372
12	SEMÂNTICA	6.305	0,04%	541
13	FALANTE	5.757	0,04%	607
14	SÍLABA	4.816	0,03%	231
15	TEXTO	18.762	0,12%	753
16	VOGAL	4.803	0,03%	229
17	ENUNCIADO	6.082	0,04%	449
18	TRADUÇÃO	9.420	0,06%	381
19	VOGAIS	4.541	0,03%	243
20	LINGÜÍSTICO	4.124	0,03%	412

R	Key word	Freq.	% Texts	BIC
1	LÍNGUA	39.089	0,43%	542
2	LINGÜÍSTICA	10.933	0,12%	255
3	LÍNGUAS	13.883	0,15%	449
4	ENSINO	19.965	0,22%	448
5	ALUNOS	18.499	0,20%	413
6	TRADUÇÃO	11.046	0,12%	329
7	APRENDIZAGEM	10.032	0,11%	384
8	LINGÜAGEM	10.984	0,12%	484
9	PORTUGUÊS	9.129	0,10%	390
10	ALUNO	9.057	0,10%	353
11	TEXTOS	8.638	0,09%	434
12	ESCRITA	7.520	0,08%	421
13	INGLÊS	8.212	0,09%	371
14	TEXTO	12.998	0,14%	463
15	PROFESSORES	10.104	0,11%	399
16	LINGÜÍSTICAS	2.901	0,03%	205
17	LETRAMENTO	3.300	0,04%	133
18	LEITURA	8.717	0,10%	417
19	PROFESSOR	10.896	0,12%	388
20	AULA	6.699	0,07%	349

Fonte: Elaborada pelo autor.

Na Figura 29, as listas de palavras-chave estão organizadas com os seguintes códigos: L_Overall_orig - a grande área da Linguística; LD_Overall_orig - a subárea da Linguística Descritiva e LA_Overall_orig para a Linguística Aplicada. A repetição dos termos comuns como: língua, linguística, linguagem, palavras, verbo, texto e fala é evidente. É visível observar como os termos relacionados ao ensino são recorrentes na LA, por exemplo: alunos, professores, aprendizagem, escrita, formação etc.

O parâmetro de corte adotado foi de 10 ST para a Linguística, 5 ST para a LD e 5 ST para a LA. Os ST foram selecionados baseados na chavacidade e o de maior chavacidade foi definido no VoBLing, sendo *língua* para Linguística, *falante* para LD, *ensino* para LA e um ST para cada uma das subáreas da LD e da LA, em LP e em LI.

Observamos que nas três listas de palavras-chave da Linguística, o ST *língua* encontra-se no topo como termo de maior frequência e chavacidade. Para definir em qual área ou subárea da Linguística a palavra-chave seria alocada, procedemos a uma análise quantitativa da chavacidade dos termos. Na Figura 29, marcamos o índice padrão de chavacidade com a seta vermelha: o BIC, como acontece no WST 7.0 e no 8.0, haja vista que ambas as versões foram utilizadas nesta pesquisa.

O resultado da alocação dos 10 termos da Linguística, baseado na maior chavacidade, segue abaixo na Figura 30.

Figura 30 - Termos chave da Linguística (10 termos).

LINGUÍSTICA		
	palavra-chave	chavacidade
1	LÍNGUA	396078,16
2	LINGUAGEM	123389,57
3	TRADUÇÃO	72431,98
4	PALAVRAS	63322,04
5	VERBO	56105,59
6	TEXTO	55450,15
7	GRAMÁTICA	54450,15
8	FALA	54020,62
9	ESCRITA	51132,29
10	FALANTE	41842,27

Fonte: Elaborada pelo autor.

Dos 10 termos acima, definimos *língua* no VoBLing. Para a escolha dos ST das subáreas da LD e da LA, utilizamos todos os *subcorpora* de cada área (LD ou LA) para indicar os termos mais chave e os alocamos (desde que já não constassem como termos da grande área de Linguística) em cada uma dessas duas subáreas. No caso em que os contextos definitórios ou explicativos não eram suficientes, utilizamos o *corpus* de manuais para extração de outros contextos.

Posteriormente às análises quantitativas, chegamos às listas de palavras-chave da Figura 31, na qual apresentamos os cinco termos de maior chavacidade, presentes nos *subcorpora* de LD e da LA.

Figura 31 - Lista de palavras-chave da LD e LA (visão parcial).

LD_overall_PT_KWL_files.kws (Unicode)									
N	Key word	Freq	%	Texts	RC	Freq	RC	%	SIC
1	SILABA	4.816	0,03%	n/a	993	31.710,52			
2	VOGAL	4.803	0,03%	n/a	1.155	31.071,58			
3	ENUNCIADO	6.082	0,04%	n/a	5.571	30.817,83			
4	LÉXICO	4.355	0,03%	465	1.723	26.266,55			
5	CORPUS	5.610	0,04%	455	7.812	25.176,58			
6	DISCURSO	14.206	0,09%	704	125.594	22.692,21			

LA_Overall_PT_KWL_07_2020.kws (Unicode)									
N	Key word	Freq	%	Texts	RC	Freq	RC	%	SIC
1	ENSINO	19.965	0,22%	448	142.965	0,02%	55.979,08		
2	ALUNOS	18.499	0,20%	n/a	121.125	0,02%	54.661,88		
3	APRENDIZAGEM	10.032	0,11%	384	34.575		40.492,45		
4	LETRAMENTO	3.330	0,04%	133	1.402		23.198,10		
5	LEITURA	8.717	0,10%	417	68.942		22.973,61		

Fonte: Elaborada pelo autor.

A seleção dos ST da Figura 31 foi feita após a eliminação manual dos ST selecionados para a área da Linguística, contidos na Figura 30. O mesmo procedimento foi adotado para as subáreas da LD e da LA, de forma que os ST selecionados anteriormente para uma área específica, fossem eliminados, não voltando a se repetir em nenhuma subárea.

No próximo subitem, tratamos do lançamento dos dados na plataforma do VoBLing.

1.18.Tratamento do corpus - VoBLing

O VoBLing é um ambiente *web* de gerenciamento terminográfico bilíngue, embasado na arquitetura original do VoTec (FROMM, 2007). Iniciamos esta pesquisa utilizando a versão 1.4 da plataforma, agora descontinuada, e terminamos com a versão 1.5.1⁷⁶ quando da conclusão da pesquisa. Fizemos modificações quanto à inserção de *hyperlinks* para os termos que já foram definidos no VoBLing, disponibilizamos uma janela *pop-up* para as definições de termos sob o *hyperlink* **Veja também** e uma disposição vertical para a ontologia na página inicial.

A versão 1.5.1 continua a disponibilizar várias formas de visualização (normal e descritiva) presentes na plataforma original, e as consultas serão nos módulos total, tradutor e modular, dependendo do objetivo do consulente. 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 sequencial. O modo **descritivo** apresenta os dados em forma hierárquica, conforme mostra a Figura 32 a seguir.

Figura 32 - VoBLing: visualização parcial do ST *linguística* no modo Descritivo.

PORTUGUÊS

Voltar ao resultado da busca

linguística

ciência da linguagem que estuda o signo lingüístico e a língua, seu funcionamento e sua estrutura com uma metodologia específica de investigação podendo ser analítica e quando trata de seu desenvolvimento passando que circunscreve no presente, ou diacrônica, uma explicação histórica das mudanças sofridas pela língua.

Nota: a linguística prende para as ciências naturais e pode ser abordada sob o ponto de vista das ciências sociais como ciência autônoma, pois concebe as línguas como sistemas que mudam com o tempo; essa ciência estuda a vida da língua, investiga como ela atua nas inter-relações entre os membros de uma comunidade ao se comunicarem entre si oral e gestualmente. A linguística pode usar algumas abordagens para o estudo da língua como a imanentista, focada no formato interno da língua, e a sócio-histórica, na qual a língua mantém relação com a história da sociedade em que ela se constitui. O fenômeno linguístico pode ser representado pelos níveis de fonética, fonologia, morfologia, lexicologia, sintaxe, semântica e pragmática. Durante a segunda guerra mundial, a posterior Revolução Industrial da Inglaterra, e a ascensão dos EUA como potência político-militar econômica e cultural impulsionam a linguística para o ensino de línguas.

Etimologia: De linguista e subo-íco. Proveniente do francês linguistique aparece já em 1844. Empregada por Herculano no Panorama, VII, 292-4 (Lúcia de Vasconcelos, Lições de História Portuguesa, 7).

Categoria Gramatical: substantivo

Gênero: feminino

Número: singular

Corpus

- Posição na Ordem de Frequência: 11
- Nº de Ocorrências do termo: 32015
- Nº de Ocorrências do fraseologismo: 0

Ontologia: Linguística

Variações Morfosintáticas: Nada encontrado

Sinónimos: ciência da linguagem

Antónimos: Nada encontrado

Hipónimos: ciências naturais

Co-hipónimos: Nada encontrado

Hiperênime: linguística descritiva/linguística aplicada/fala

Exemplos

1. A tese de que, no pensamento saussureano, a Semiologia precede à Linguística e é explicada, dentre outros pontos, pela reflexão filosófica que o genealogista fez sobre a linguagem. Dessa reflexão, deriva que a arbitrariedade do signo, tem como o valor a ele relacionada, não pode ter correspondência a partir da compensação entre sistemas semiológicos diferentes e não apenas do estudo da língua por ele mesmo.

Fonte: Elaborada pelo autor.

Na Figura 32, vemos os dados apresentados de forma separada para o termo *linguística*, facilitando ao leitor a localização de dados específicos como categoria gramatical, gênero e número mais rapidamente.

⁷⁶ Plataforma do VoBLing disponível em <http://vobling.votec.ileel.ufu.br/>. Acesso em 13 de jul. de 2020.

Quanto à estrutura interna, de acesso restrito aos pesquisadores, apresentamos as telas para descrever seu funcionamento. A primeira tela do sistema nos dá a visão geral dos ST cadastrados e se constitui como a página onde podemos registrar novos ST. Os ST já cadastrados são aqueles que aguardam a aprovação do administrador da plataforma, apresentados na Figura 33.

Figura 33 - VoBLing: tela inicial de registro de termos.

análise do discurso	Português	06/02/2020	Marcio Issamu Yamamoto	editar
discourse analysis	Inglês	06/02/2020	Marcio Issamu Yamamoto	editar
conversation analysis	Inglês	05/02/2020	Marcio Issamu Yamamoto	editar
análise da conversação	Português	03/02/2020	Marcio Issamu Yamamoto	editar
linguagem	Português	30/01/2020	Marcio Issamu Yamamoto	editar
linguística	Português	23/01/2020	Marcio Issamu Yamamoto	editar
língua	Português	07/11/2019	Marcio Issamu Yamamoto	editar
language	Inglês	07/11/2019	Marcio Issamu Yamamoto	editar
ecolinguística	Português	17/09/2019	Marcio Issamu Yamamoto	editar
ecolinguistics	Inglês	17/09/2019	Marcio Issamu Yamamoto	editar
linguística descritiva	Português	17/09/2019	Marcio Issamu Yamamoto	editar
descriptive linguistics	Inglês	17/09/2019	Marcio Issamu Yamamoto	editar
linguística aplicada	Português	17/09/2019	Marcio Issamu Yamamoto	editar

Fonte: Elaborada pelo autor.

Na Figura 33, temos os STF **linguística descritiva** e *descriptive linguistics* como termos que estão em análise.

Na página da **Ontologia**, houve modificações quanto ao *layout* de inserção de dados, quando do cadastro dos ST, o que apresentamos nas Figuras 34 e 35.

Figura 34 - VoTec: tela de cadastro e ontologia do termo *etymology* em 2015.

Fonte: Elaborada pelo autor.

Figura 35– VoBLing: tela de cadastro e ontologia dos termos em 2020.

Fonte: Elaborada pelo autor.

Na Figura 34, apresentamos a tela da **Ontologia** na versão de 2015 e na Figura 35 a versão atual. Na versão anterior do VoTec (YAMAMOTO, 2015), havia os três campos acima retratados: *Termo*, *Língua* e *Ontologia*. Na versão atual, há os campos *Termo*; o campo *Língua* foi substituído por *Escolha um tipo*, que é a escolha do idioma (LP ou LI); e o campo *Ontologia* foi substituído pelo campo *Grande área*.

A Figura 35 mostra que quando o pesquisador seleciona uma grande área, um novo campo se abre para que as subáreas desejadas sejam selecionadas. Comparando este projeto (que envolve a Linguística Descritiva e a Linguística Aplicada) com a versão anterior (YAMAMOTO, 2015), pode-se dizer que a versão atual é mais densa e envolve 49 subáreas da Linguística.

Figura 36 - VoBLing: cadastro da ontologia do termo *ecossistema*.

Fonte: Elaborada pelo autor.

A Figura 36 detalha a **Ontologia** do termo *ecossistema* no VoBLing 1.5.1. Registramos o termo *ecossistema* com as seguintes especificações: (1) Língua: português; (2) Grande área:

Linguística; (3) Subárea 1: Linguística Descritiva e (4) Subárea 2: Ecolinguística. Preenchidos esses campos, passamos para o próximo passo que é o registro dos **Contextos**, visível na Figura 38.

Figura 37 – VoBLing: tela de cadastro de Contextos.

Exemplo	Conceito	Fonte	Ações
*... a língua é real, viva, altera-se conforme os contextos de uso; desempenha funções conforme a intenção e o interlocutor.	real, viva, altera-se conforme os contextos de uso; desempenha funções conforme a intenção e o interlocutor	PER 2 02/12/2013	editar excluir
A linguagem é uma capacidade inata nos seres humanos, enquanto a língua é uma materialização social e histórica dessa capacidade.	materialização social e histórica da linguagem	PER 2 15/06/2015	editar -

Fonte: Elaborada pelo autor.

A Figura 37 mostra a tela do campo **Contextos** e os campos de preenchimento: exemplo, conceito, fonte e data de coleta do *corpus*. Apesar de não haver necessidade de alteração quanto ao *layout* da versão anterior, a mudança foi em relação à abordagem da quantidade de contextos a serem registrados pelo pesquisador. Na versão atual, aumentamos a quantidade de contextos registrados no VoBLing devido ao tamanho maior do *corpus*. Em 2015, o dimensionamento era de 1 milhão de itens, pois abarcou somente a subárea da LH (YAMAMOTO, 2015), enquanto na versão atual (1.5.1) abarcamos a Linguística e suas subáreas.

É importante mencionar uma nuance sobre os acrônimos no registro dos contextos. Como o público-alvo primário da obra são os discentes dos anos iniciais dos cursos de Letras, consideramos relevante e didático explicitar o significado dos acrônimos para promover a compreensão completa deles em seu contexto. Por exemplo, no registro dos contextos da subárea de *First Language Acquisition* (Aquisição da Linguagem), a sigla *Second Language Acquisition* (SLA) foi recorrente. Logo, escrevemos o nome da subárea por extenso e indicamos o acrônimo em parênteses.

Outro fator relevante a ser esclarecido é a questão da ortografia. As definições foram redigidas a partir do Acordo Ortográfico da Língua Portuguesa de 1990⁷⁷. Contudo, os *corpora* compilados cobrem um período do novo e do antigo acordo ortográfico. Diante desse fato, optamos por manter a ortografia do texto compilado, de acordo com o padrão original, a fim de mantermos a originalidade do texto. Essa foi a mesma postura adotada com os padrões de textos de português europeu, africano e do Brasil.

O mesmo procedimento foi adotado nos textos de língua inglesa (LI) de padrão britânico e norte-americano. Isto é, mantivemos a ortografia original do texto nos excertos e atualizamos a ortografia para o padrão de inglês americano nos PD do VoBLing.

O próximo passo, depois do registro de **Contextos**, é preencher a página de elaboração da microestrutura. Na Figura 38, apresentamos o registro do termo *linguística descritiva* e comentamos as diferenças entre a versão 1.5.1 e a versão de 2015 com as respectivas alterações na aba **Dados**.

Figura 38 - VoBLing: visualização da aba Dados – *linguística descritiva*.

The screenshot shows the 'Dados' (Data) tab in the VoBLing interface. It contains several input fields and buttons for entering linguistic information. The 'Ontologia' field is set to 'Linguística'. The 'Categoria Gramatical' is 'Substantivo', 'Número' is 'Singular', and 'Gênero' is 'Feminino'. The 'Lemas' field contains a list of variations: 'linguística; linguísticas; lingüística; lingüísticas; linguística descritiva; lingüística descritiva'. The 'Etimologia' field contains 'linguística - veja linguística; descrição - do latim descriptio'. There is an 'Áudio' section with a button 'Escolher arquivo' and a status 'Nenhum arquivo selecionado'. The 'Corpus' section shows 'Posição na Ordem de Frequência do termo base' as 11, 'N° de Ocorrências do termo base' as 32815, and 'N° de Ocorrências do fraseologismo' as 63. There are also 'ouvir' and 'excluir' buttons.

Fonte: Elaborada pelo autor.

A Figura 38 apresenta os campos de preenchimento para o termo *linguística descritiva* na aba **Dados**. As alterações foram: acrescentamos o campo (1) *Lemas*, onde inserimos as possíveis variações dos termos como *linguística* e *linguísticas*; *lingüística* e *lingüísticas*;

⁷⁷ Disponível em <http://www.portaldalinguaportuguesa.org/index.php?action=acordo&version=1990b>. Acesso em 14 ab. de 2020.

lingüística descritiva e linguística descritiva; (2) *Etimologia*, compilados a partir de dicionários de etimologia em português e em inglês; (3) *Áudio*, para pronúncia dos termos; (4) *n.º de ocorrências do termo/fraseologismo*, campo para inserção da frequência dos fraseologismos⁷⁸.

Quanto à etimologia dos ST, utilizamos o Dicionário Etimológico da Língua Portuguesa de Nascentes (1955) para a LP e o *Comprehensive Etymological Dictionary of the English Language* de Klein (1971) e o *Origins: A Short Etymological Dictionary of Modern English* de Partridge (1966) para a LI. Como a quantidade de verbetes do dicionário de Nascentes (1955) foi insuficiente para a demanda desta pesquisa, baseado na formação do pesquisador e nos estudos de LH do mestrado, optamos por adotar, o *Diccionario crítico etimológico castellano e hispánico* de Corominas (1984), quando o étimo do ST era de origem latina ou grega, como em **descrição** - do latim *descriptio*, utilizado para o STF *lingüística descritiva*, e como em **eco** - do grego *oikos*, utilizado para o ST *ecolinguística*. Observamos, em nossa análise do estado da arte dos dicionários de Linguística, que Nascentes (1955) também foi adotado por Massaud (1995) como referência para a produção do dicionário de Termos literários. No caso em que a etimologia de um componente do STF já havia sido carregada na plataforma, e fazia parte das remissivas, utilizamos a marca **veja**: como em Linguística Descritiva: “lingüística - veja linguística; de descrição - do latim *descriptio*”.

Para preencher os campos **n.º de ocorrências do termo** e o **n.º de ocorrências do termo/fraseologismo**, usamos os seguintes parâmetros no caso do termo *lingüística descritiva*. Primeiramente, o campo **n.º de ocorrências do termo** é preenchido com os dados provenientes da lista de palavras dos *corpora* para o termo *lingüística*, que será considerado o termo base para a *lingüística descritiva*. Ou seja, partimos dos substantivos para preencher este campo. Em segundo lugar, a partir dos dados provenientes do concordanciador, que traz o número de ocorrências do fraseologismo Linguística Descritiva, e de forma a evitar a confusão na hora do preenchimento da aba *Dados*, é importante dizer que na versão 1.4 preenchíamos o campo **n.º de ocorrências do termo/fraseologismo** e na versão 1.5, o campo é chamado **n.º de ocorrências do fraseologismo**.

A Figura 39 mostra como os traços distintivos são organizados. Na sequência, explicamos como é feito o preenchimento e como o quê? foi utilizada nesta pesquisa.

⁷⁸ Neste trabalho, consideramos fraseologismos todas as expressões cuja extensão sejam acima daquelas compostas pela junção de duas UL, como por exemplo, Linguística Descritiva. Para considerar fraseologia, baseamos em Fromm (2020, p. 769), “qualquer agrupamento de duas ou mais palavras cuja frequência combinada seja destacada pelos programas de análise lexical.”

Figura 39 - VoBLing: visualização da aba Traços distintivos – *linguística descritiva*.

#	A	B	C	D	E	F	G
1	aborda fon...						
2		explica as...					
3			fundamenta...				
4					originou a...	intuito de...	
5				linguistic...			
6				linguistic...			
7		documenta ...			sublinha da...		
8						opôs-se à ...	
9							concepção ...
10						pressupost...	
11			de base em...				
12			favorece a...				
13	fonética e...						
14					a Linguíst...		Linguistic...
15					domínios d...		

Fonte: Elaborada pelo autor.

A Figura 39 retrata a organização dos traços distintivos do ST *linguística descritiva*. Os traços que são semelhantes são organizados na mesma coluna, de forma a tornar mais prática a visualização dos traços semelhantes mais recorrentes. Como iniciamos o uso da plataforma cientes de que haveria a definição terminológica e a definição enciclopédica, optamos por deixar a última coluna exclusivamente para os dados que compõem o paradigma enciclopédico, correspondente à coluna G da figura. O *pop-up* visível nesta figura possibilita a visualização rápida dos traços sem que sejam selecionados manualmente. Neste caso, o cursor é deixado sobre o traço e, automaticamente, a linha com os dados se torna visível.

Na aba **Semântica**, marcamos o box **Termo dicionarizado** para registrar que o ST consta definido em dicionários, escolhemos se a definição é coincidente (de forma completa ou parcial) ou não, conforme se pode visualizar no canto superior direito da Figura 41.

Figura 40 – VoBLing: visualização da aba Semântica - *linguística descritiva*.

The screenshot shows the 'Semântica' tab in the VoBLing interface. It contains the following fields and values:

- Termo Dicionarizado?**: ☒
- Definições Coincidentes?**: A dropdown menu with options 'Parcial', 'Sim', 'Não', and 'Parcial' (selected).
- Fonte**: Wiktionary
- Definição Dicionarizada**: estudo do mecanismo pelo qual uma dada língua funciona, num dado momento, como meio de comunicação entre os seus falantes, e da análise da estrutura, ou configuração formal, que nesse momento a caracteriza
- Hipêrônimo de**: semântica; biolinguística
- Hipônimo de**: língua
- Co-hipônimo de**: linguística aplicada
- Sinônimo(s)**: (empty field)
- Antônimo(s)**: linguística chomskiana
- Notas**: (empty text area)

Fonte: Elaborada pelo autor.

Após a classificação do tipo de definição, selecionamos a fonte que originou os dados. Na plataforma, esses dados servem de apoio ao pesquisador, já que não são disponibilizados ao consulente. Nos outros campos da página de visualização, utilizamos os dados provenientes dos contextos de ocorrência do ST (aba **Contextos**) para preencher os dados semânticos. Esses dados são: hipêrônimo de, hipônimo de, co-hipônimo de, sinônimo e antônimo. O campo **Notas** é reservado à comunicação entre o pesquisador e o administrador da página.

A aba denominada **Termo equivalente** registra a relação a ser feita no sistema entre os ST de LP com o de LI. No exemplo a seguir, a Figura 41 retrata a ligação da *linguística descritiva* ao seu equivalente em LI, *descriptive linguistics*.

Figura 41 – VoBLing: termo equivalente - *linguística descritiva*.

The screenshot shows the 'Termo Equivalente' tab in the VoBLing interface. It contains the following elements:

- Text instruction**: O termo equivalente deve ser um termo cadastrado em inglês, com a mesma ontologia e que ainda não esteja marcado como equivalente de outro termo.
- Form fields**: 'Buscar Termo Equivalente:' followed by a text input field and a 'Buscar' button.
- Table**: A table with one row containing the term 'descriptive linguistics' and an 'Apagar' button.

Fonte: Elaborada pelo autor.

Na aba seguinte, a de **Termos remissivos**, serão buscados os ST que foram elencados no mesmo campo semântico do ST que está sendo definido e que já tenham sido aprovados pelo administrador.

Figura 42 – VoBLing: aba Termos remissivos do ST - *linguística descritiva*.

Os termos remissivos devem ser termos cadastrado em português e com os respectivos cadastro já aprovados pelo administrador.

Buscar Termo Remissivo:

Termo	Apagar
língua	
linguística	
linguística aplicada	
análise da conversação	
biolinguística	

Fonte: Elaborada pelo autor.

A Figura 43 mostra os termos remissivos do ST *linguística descritiva* e os ST relacionados com os hipônimos *análise da conversação*, *biolinguística* e *língua*; o hiperônimo *linguística* e o co-hipônimo *linguística aplicada*. Na visualização disponível ao consulente, esses ST estão disponíveis na tela em forma de *pop-up* e *hyperlink*.

Excepcionalmente, o ST *linguística descritiva* traz várias remissivas por se encontrar abaixo do topo da árvore de domínio. Contudo, por uma questão pedagógica e didática, optamos por manter o número de remissivas reduzido para que a tela *pop-up* não ficasse poluída para o consulente. Diante disso, em primeiro lugar, optamos por limitar o número de remissivas ao mínimo de uma e ao máximo de 3, caso haja relações semânticas nos contextos carregados no VoBLing, ou seja, caso haja existência de hiperônimos e hipônimos. Em segundo lugar, consideramos a presença de co-hipônimos, sinônimos e antônimos. O candidato a escolha foi aquele equivalente ao próximo nível de subordinação ou superordenação na árvore de domínio.

A aba **Informações enciclopédicas** disponibiliza o campo a ser preenchido com as informações enciclopédicas sobre o ST, provenientes de um *site* de acesso público. Para a maioria dos verbetes, utilizamos a Wikipédia, salvo as exceções, quando o próprio pesquisador elaborou a definição, conforme mostra a Figura 43.

Figura 43 – VoBLing: aba Informações enciclopédicas - *linguística descritiva*.

Definição	Artigo	Fonte	Link	Apagar
estudo do mecanismo pelo qual uma dada língua funciona, como meio de comunicação entre seus falantes.	Linguística Descritiva	Linguística em foco		

Fonte: Elaborada pelo autor.

Quando o pesquisador não encontra os dados disponíveis na Wikipédia, um novo *site* pode ser cadastrado para as informações enciclopédicas (Figura 43). No caso da *linguística descritiva*, a Wikipédia não disponibiliza essa definição, o que nos levou ao *site* Linguística em foco.

Quanto à aba **Multimídia**, houve modificações relevantes e didáticas com a possibilidade de inserção de outras formas de informações de caráter distinto dos textos verbais, conforme mostra a Figura 44.

Figura 44 - VoBLing: visualização da aba Multimídia - *linguística descritiva*.

Fonte	Tipo	Link	Apagar
Youtube - O que é lingu	Vídeo		
Imagem autoral	Figura		

Fonte: Elaborada pelo autor.

Conforme a Figura 44, na aba **Multimídia** podemos inserir imagens e mídias. A figura mostra a possibilidade de inserção de: árvore hiperbólica/grafos, cartograma, diagrama, figura, linha do tempo, *links* para internet, nuvem lexical, pronúncia, áudio (aba Dados) e vídeo. Nesta pesquisa, inserimos (i) figuras com base no recorte da árvore de domínio da Linguística para os verbetes definidos; e (ii) vídeos do Youtube com explicações sobre as subáreas da Linguística,

caso haja, na aba *Multimídia*. A inserção do recurso de áudio no VoBLing é visível na Figura 45.

Figura 45 - VoBLing: visualização da aba Dados/áudio - *linguística descritiva*.

The screenshot displays the VoBLing interface with the 'Dados' tab selected. The 'Ontologia' field is set to 'Linguística'. The 'Categoria Gramatical' is 'Substantivo' and 'Número' is 'Singular'. The 'Gênero' is 'Feminino'. The 'Lemas' field contains the text 'linguística;linguísticas;lingüística;lingüísticas;linguística descritiva;lingüística descritiva'. The 'Etimologia' field contains 'linguística - veja linguística: de descrição - do latim descriptiōne'. The 'Var.' and 'Morfossintáticas' fields are empty. The 'Áudio' section at the bottom has a button 'Escolher arquivo', the text 'Nenhum arquivo selecionado', and buttons 'ouvir' and 'excluir'.

Fonte: Elaborada pelo autor.

Os arquivos de áudio, com a pronúncia dos ST em LP e LI são carregados na aba **Dados**. A Figura 45 retrata o arquivo de áudio da *linguística descritiva*, em formato mp3, gravado em LP. O mesmo tipo de arquivo foi disponibilizado para a LI, gravado por professores assistentes⁷⁹ de ensino de LI da Comissão Fulbright, provenientes dos Estados Unidos. Os arquivos de áudio foram gravados no exterior e enviados ao Brasil via *GoogleDocs*.

As imagens foram produzidas a partir de um *site* da internet: o *AirMore Mind*⁸⁰. O *AirMore Mind* disponibiliza o serviço de acesso gratuito para a construção de mapas mentais. O acesso ao serviço é possível via cadastro no *site*, via Gmail ou via Facebook para uma conta com recursos limitados, caso se opte pelo acesso não pago.

⁷⁹ Ruben Adery - Linguistix Pronunciation – Disponível em: <https://www.youtube.com/channel/UC3wikulG1obZp9k1sH2cH3Q>. Acesso em 20 jul. de 2020; Email: ruben@linguistixpro.com; Instagram @linguistixpro; Linguistics Pronunciation – Site: <https://www.linguistixpro.com/> Email: ruben@linguistixpro.com e Gautam Ramesh: gautam.ramesh1@gmail.com.

⁸⁰ AirMore Mind: disponível em <https://mind.airmore.com/login>. Acesso em 20 maio de 2020.

Figura 46 – VoBLing: imagens da Linguística e LD.



Fonte: Elaborada pelo autor.

Na Figura 46, temos o padrão das imagens inseridas na aba *Multimídia* para representar a posição do ST na árvore de domínio da Linguística. As imagens correspondem à posição teórica que adotamos e à definição dos ST no VoBLing, e os dados são provenientes do *corpus*. Na parte esquerda da imagem, há o item superordenado, a área da Linguística. À direita, há os itens subordinados: as subáreas da LD e da LA.

Em relação à última tela do banco de dados, inserimos o campo **Nota**, além dos campos já existentes **Conceito final** e **Definição**, conforme a Figura 48 a seguir. O campo **Conceito final** é um espaço destinado ao esboço da definição final. Primeiramente, o pesquisador usa este espaço para organizar os traços distintivos úteis à construção do PD do ST. Em seguida, no campo **Definição**, o pesquisador redige o PD final, conforme padrão adotado para a pesquisa, e aguarda a aprovação do administrador para disponibilizá-la aos consulentes.

Figura 47 - VoBLing: Campo Nota.

Fonte: Elaborada pelo autor.

No campo **Nota**, são inseridos os dados que fazem parte da definição enciclopédica e no campo *definição* foi construída a definição terminológica no padrão GPDE, atendendo ao resultado da pesquisa com o público-alvo. Como resultado da pesquisa, houve a predominância

do paradigma enciclopédico e, em seguida, do paradigma terminológico, conforme apresentado na Tabela 1, p. 61.

Em relação aos dados provenientes do *corpus*, utilizados na construção da macro e da microestrutura, explicamos os procedimentos adotados para seu tratamento por meio do programa WST.

Inicialmente, os candidatos a verbetes foram identificados a partir de dados estatísticos e probabilísticos, pelas ferramentas lista de palavras e lista de palavras-chave do WST. O candidato a verbete se constitui como verbete a partir da recorrência do termo nos *subcorpora*, em inglês e português, e a partir da existência de contextos definitórios e explicativos que permitem a construção da definição. Por fim, os verbetes são definidos seguindo a metodologia do VoBLing, com base teórica da Terminologia de *Frames*, e disponibilizados para visualização na página inicial do VoBLing.

A microestrutura disponibiliza os seguintes dados organizados: (i) *categoria gramatical*⁸¹; (ii) *hiperonímia*; (iii) *hiponímia*; (iv) *co-hiponímia*; (v) *sinonímia*; (vi) *antonímia*; (v) *posição do termo/verbo*; e (vi) *fraseologismo* no *corpus* e a (vi) *definição* fazem parte da microestrutura. A organização dos conceitos que compõem a estrutura da definição terminológica seguiu o padrão GPDE, anteriormente explicitado na seção 4.5, página 73.

Como escolhemos dez (10) contextos definitórios ou explicativos de ocorrência dos ST, a quantidade de hiperônimos e hipônimos foi grande. Por conseguinte, para escolher qual desses possíveis ST inserir na plataforma, estabelecemos o padrão de um ST para atender ao aspecto pedagógico da obra.

Caso o *corpus* trouxesse mais informações do que as que seriam usadas na definição de padrão terminológico, disponibilizamos um campo à parte reservado para **NOTA**. No campo **NOTA**, organizamos o padrão próximo à definição lexicográfica ou enciclopédica, sendo que a primeira se interessa mais por uma **descrição linguística** do ST e a segunda, pela descrição da **coisa** ou **fenômeno**. Essa foi uma possibilidade surgida da necessidade de construção das definições, caso o *corpus* trouxesse esse tipo de informação.

⁸¹ A categoria gramatical contém os seguintes dados: número (singular, plural ou dual), gênero (masculino, feminino ou neutro), lemas (ex.: linguística e lingüística), etimologia e áudio (pronúncia do termo em português e inglês).

6. O PARADIGMA DEFINICIONAL (PD) BASEADO EM FRAMES

A metodologia terminográfica do VoTec (FROMM, 2007) proporciona uma mudança de paradigmas do *savoir faire* da Terminologia, devido à adoção da LC. A Terminologia e os trabalhos terminográficos adotam uma perspectiva onomasiológica em sua grande maioria, partindo do conceito para se chegar ao termo. Entretanto, o fato de o VoBLing ser baseado em *corpus*, o ponto de partida são os candidatos a termos provenientes das listas de palavras-chave processadas via WST. Isto significa que partimos dos possíveis candidatos a termos para identificar os conceitos ou traços distintivos que definem o termo, o que caracteriza uma perspectiva semasiológica.

Nesta seção, descrevemos os procedimentos de construção do PD baseado em *Frames*, aplicados à estrutura do VoBLing. Os dados são provenientes dos *corpora* de Linguística, das subáreas da Linguística Descritiva e da Linguística Aplicada.

Considerando os princípios da Terminologia de *Frames*, expostos na seção 2.3.2.2, página 35, e na organização dos traços distintivos proposta por Ilari (2003), ambos baseados no princípio da maior frequência no *corpus*, propomos a definição disponível na Figura 48. Esta figura apresenta a visualização prévia do termo *linguística*, pertencente à grande área da Linguística, como teste da visualização do VoBLing 1.5.

Figura 48 - VoBLing: visualização normal do termo *linguística*.



Fonte: Elaborada pelo autor.

Com o objetivo de possibilitar maior acessibilidade terminológica do PD ao público-alvo, estruturamos o PD *linguística* por meio de dois apostos que explicam os adjetivos **sincrônica**: quando trata de um desenvolvimento passado que continua no presente; e **diacrônica**, uma explicação histórica das mudanças sofridas pela língua, foram explicadas por

meio de apostro. Esta estratégia objetiva maior contato e consequente familiarização do aluno iniciante de Letras à terminologia da Linguística. O apostro do qual falamos pode ser tomado como uma mini definição de **sincrônica** (trata de um desenvolvimento passado que continua no presente) e **diacrônica** (explicação histórica das mudanças sofridas pela língua), proveniente do *corpus* de Linguística, identificada a partir da busca pelo ST **sincronia** e **diacronia**. A busca foi realizada com a expressão ST+ **ser** como em *sincronia é*, utilizando a ferramenta Concordanciador (FINATTO, 2020, p. 81-84).

A Figura 49 retrata como o ST *linguística* foi definido para refletirmos sobre a importância da padronização do PD. Houve a necessidade dessa padronização para que houvesse um formato homogêneo para todos os ST de cada subárea da Linguística, dos ST e dos STF de cada subárea da Linguística Descritiva e da Linguística Aplicada.

Com base nesses procedimentos, chegamos à proposta de PD, organizando os traços distintivos presentes no *corpus*. Apresentamos a ficha terminológica da plataforma do VoBLing, na qual os traços mais frequentes encontram-se relacionados nas colunas, conforme demonstrado na Figura 49.

Figura 49 – VoBLing: traços distintivos de *linguística* (em colunas).

#	A	B	C	D	E
1					Semiologia...
2	ciência	objeto e u...			
3	ciência au...	descrição ...		representa...	
4		objeto é a...			
5	ciência da...	estudo da ...	investiga ...		
6					abordagem...
7	ciência ma...			perde para...	
8	ciência no...	estudo da ...			
9		linguas sa...			linguístic...
10			objeto de ...		
11	campo	estuda a l...		impulsiona...	Durante a ...

Fonte: Elaborada pelo autor.

Da mesma forma como foi apresentado na Figura 39, página 132, a última coluna foi reservada aos traços distintivos que seriam utilizados para a construção do PD enciclopédico (na Figura 49, encontra-se na coluna E) do VoBLing, diferente daquela proposta original de Fromm (2007). Essa alteração se justifica, pois, a quantidade de traços distintivos presentes no

corpus de Linguística tomado para esta pesquisa é maior do que o *corpus* utilizado por Fromm em 2007. Além disso, a quantidade de contextos definitórios selecionados pelo pesquisador para esta pesquisa foi maior que dez contextos por ST. A organização dos traços distintivos, organizados em uma coluna única, evita que a tela fique muito grande, tornando o trabalho do pesquisador mais prático.

Figura 50 – VoBLing: visualização parcial do STF Linguística Descritiva no modo normal.



Fonte: Elaborada pelo autor.

O PD *descriptive linguistics* inicia-se com a definição *subarea of Linguistics that describes...* em inglês, diferente da definição em português, conforme mostra a Figura 50. Essa estrutura distinta, no início da definição, mostra a necessidade de padronização do PD. Ou seja, foi vital propormos uma estrutura padrão que fosse aplicada a todos os PD, ao definirmos as subáreas da Linguística. Nesse caso, a melhor opção seria: *subarea of Linguistics that describes...* Além desses dados, identificamos os traços distintivos que permeiam a definição da subárea a partir da estrutura nuclear dos verbos e da descrição. Ex.: (EN) *describes...*; *grounded in...*; *conceives...* que **descreve** a língua, **baseada** em observações empíricas sistemáticas e **concebe** a língua como um sistema simbólico altamente complexo.

No intuito de tratar das estruturas distintas de PD, faremos descrições considerando os princípios de analogia e de anomalia⁸². A opção por esses princípios se justifica, pois, embora haja uma tentativa de padronização, os padrões de PD não serão sempre os mesmos. Essa certeza repousa no fato de que os dados são muitos, contudo nem sempre os traços distintivos

⁸² Nesta pesquisa, concebemos analogia como o padrão de PD que pode ser aplicado aos ST, conforme a nossa proposta baseada em *frames*. Anomalia é a estrutura de PD que não segue o modelo padrão baseado em *frames*.

encontrados serão os mesmos para todas as subáreas da LD e da LA, nem mesmo haverá a mesma recorrência dos traços distintivos dentro do *corpus*.

Para explicar melhor o conceito de analogia e anomalia, trazemos as contribuições de Robins (1997) ao tratar do tema. Segundo o autor, os filósofos discutiam a estrutura da língua grega e seus paradigmas e concebiam analogia como os padrões recorrentes no mesmo grupo de palavras, suas terminações e a relação de forma e significado⁸³. Quando o sistema da língua não permitia tais regularidades, havia a anomalia, estabelecendo, assim, a quebra do padrão ideal dentro do sistema linguístico, o que chamamos de exceção à regra. Da mesma forma, os ST, cujos traços distintivos não forem condizentes com o padrão proposto, comporão a anomalia.

Na próxima seção de resultados, exemplificamos esses padrões contextualizados à prática terminográfica do VoBLing.

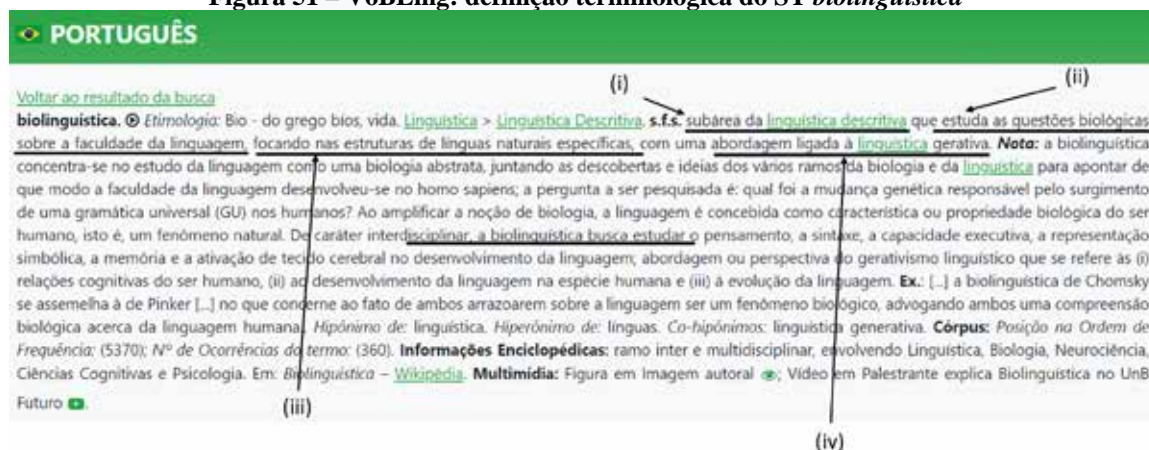
⁸³ “As regularidades buscadas pelos analogistas eram aquelas de paradigmas formais em que palavras com o mesmo *status* gramatical tinham as mesmas terminações morfológicas e estrutura tônica; [...] aquelas envolvendo as relações entre forma e significado, nas quais se podia esperar que palavras comparáveis morfológicamente tivessem significados comparáveis e "análogos", e vice versa.”(ROBINS, 1997, p. 20).

7. RESULTADOS

Neste capítulo, fazemos a comparação entre os paradigmas propostos no subitem 4.5.1, **Proposta de PD baseado em Frames** (p. 73) e os PD construídos quando da definição dos verbetes do VoBLing. Os PD foram comparados aos paradigmas baseados em Frames, conforme proposição feita nos Quadros 5, 6 e 7 e, com base nesses aspectos, descrevemos a analogia e a anomalia dos ST.

A primeira definição analisada é a do ST *biolinguística* e o paradigma aplicado a esta definição é a Proposta 2 do PD terminológico (Quadro 6). A Proposta 2 é composta pelos seguintes descritores: nome da subárea + descritor 1 - subárea da L/LD/LA + descritor 2 - o que estuda/faz? + descritor 3 - modo (como?) + descritor 4 - para que estuda? Esta proposta foi selecionada, pois é a proposta aplicada às subáreas da Linguística. O PD construído para a *biolinguística* está exemplificado na Figura 51.

Figura 51 – VoBLing: definição terminológica do ST *biolinguística*

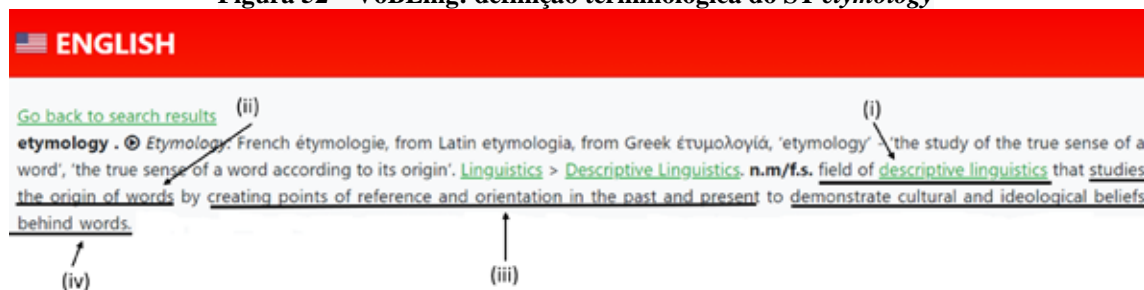


Fonte: Elaborada pelo autor.

De acordo com a Figura 51, notamos que a definição de *biolinguística* é composta por quatro descritores. Dentre os quatro descritores, três seguem o paradigma proposto e se enquadram no padrão de **analogia**, mas o quarto difere da proposta original - é uma **anomalia**. Os descritores de um a três respondem às perguntas do PD da Proposta 2: a *biolinguística* é (i) uma **subárea** da LD que (ii) estuda (o quê?) as questões biológicas sobre a faculdade da linguagem (iii) (como?) focando nas estruturas de línguas naturais específicas. O traço distintivo (iv) (**para que** estuda?) resta em aberto. Em seu lugar, há a identificação da abordagem adotada pela *biolinguística*, abordagem da linguística gerativa.

O segundo exemplo analisado é a definição de *etymology*, baseado no mesmo paradigma do ST anterior, a Proposta 2 do PD terminológico, já que se trata de outra subárea da LD. A definição de *etymology* se encontra na Figura 52.

Figura 52 – VoBLing: definição terminológica do ST *etymology*



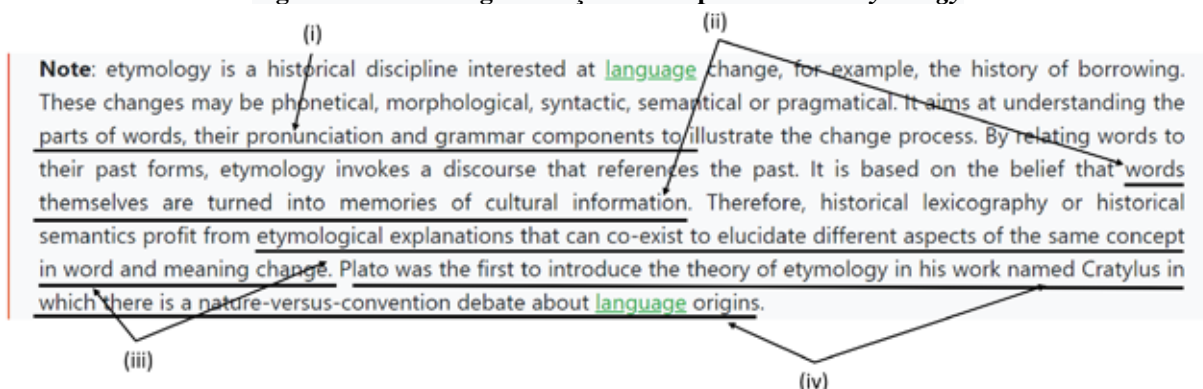
Fonte: Elaborada pelo autor.

De acordo com a definição do ST *etymology*, temos a descrição de um PD que corresponde à Proposta 2 em sua integralidade. Vejamos em detalhes: *etymology* is (i) *field of descriptive linguistics that*/subárea da Linguística Descritiva (ii) *estuda (o quê?) studies the origin of words/* a origem das palavras (iii) *(como?) by creating points of reference and orientation in the past and present/* criando pontos de referência e orientação no passado e no presente (iv) *(para que estuda?) to demonstrate cultural and ideological beliefs behind words/* a fim de demonstrar as crenças culturais e ideológicas por detrás das palavras.

A partir desses exemplos, observamos que, devido ao *corpus* robusto que utilizamos, é possível adaptar a Terminologia de *Frames* e a metodologia do VoTec para definir as subáreas da Linguística com poucas anomalias. As anomalias se manifestam mais frequentemente nos PD enciclopédicos, em LP e em LI, pois os dados disponíveis são bastante variáveis.

No capítulo seis desta tese, apresentamos quatro modelos diferentes de PD denominados de Propostas de PD enciclopédico um a quatro, com quatro ou seis descritores cada. O procedimento adotado, primeiramente, foi fazer a comparação da Proposta 4 com o PD enciclopédico de *etymology*, considerando os quatro descritores: (i) objeto de investigação, (ii) o conceito do objeto, (iii) como o estudo é conduzido e (iv) um breve histórico da disciplina. Para ilustrar esse dado, descrevemos a seguir dois modelos anômalos de PD enciclopédico do VoBLing.

Figura 53 – VoBLing: definição enciclopédica do ST *etymology*



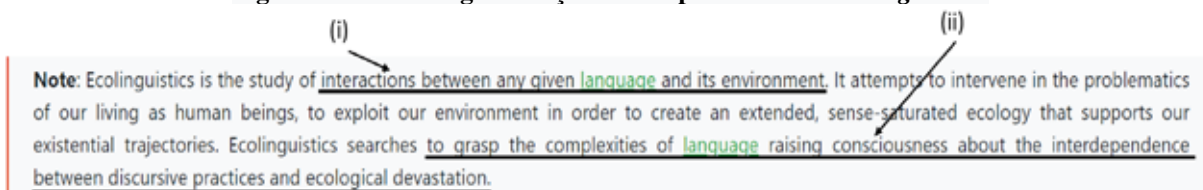
Fonte: Elaborada pelo autor.

Analisando os quatro descritores do PD enciclopédico de *etymology*, encontramos os seguintes descritores: (i) objeto de investigação da etimologia – *parts of words, their pronunciation and grammar components* – partes das palavras, sua pronúncia e componentes gramaticais; (ii) define o conceito de palavras para etimologia - *words themselves are turned into memories of cultural information* - palavras se transformam em memórias de informação cultural; (iii) descreve como o estudo é conduzido: *explanations can co-exist to elucidate different aspects of the same concept in word and meaning change* - explicações podem coexistir para esclarecer aspectos diferentes de um mesmo conceito na mudança lexical e deslocamento semântico; (iv) um breve histórico da disciplina - *Plato was the first to introduce the theory of etymology in his work named Cratylus in which there is a nature-versus-convention debate about language origins* - Platão foi o primeiro a lançar a discussão etimológica em sua obra chamada Crátilo, na qual há um debate entre o natural *versus* o convencional para explicar a origem da língua.

Ao analisar o item quatro, observamos que nele reside a anomalia desta definição, pois a menção a Platão e à sua obra Crátilo como marco fundador da etimologia ocidental é um evento pontual e não um breve histórico da disciplina. Ou seja, assim como ocorreu no PD terminológico aplicado à *biolinguística*, três descritores são análogos à proposta e o último é anômalo. As anomalias ocorrem pela ausência de traços distintivos suficientes para completar as lacunas necessárias para cada descritor.

Em segundo lugar, analisamos o PD enciclopédico do ST *ecolinguistics* e o comparamos à Proposta 3, a mais próxima daquela definição. A Proposta 3 compreende quatro descritores, quais sejam: (i) o objeto de estudo, (ii) a pergunta de pesquisa, (iii) a concepção do objeto da pesquisa e (iv) a abordagem ou perspectiva que embasa a subárea da linguística.

Figura 54 – VoBLing: definição enciclopédica do ST *ecolinguistics*.



Fonte: Elaborada pelo autor.

Ao analisar a definição enciclopédica criada para o ST *ecolinguistics*, notamos que é possível identificar somente dois descritores dos quatro presentes na Proposta 3 do PD enciclopédico baseado em *Frames*. Ou seja, encontramos o descritor presente no PD, isto é, (i) o objeto de estudo: *interactions between any given language and its environment*/ as interações entre uma língua e seu meio ambiente. Quanto ao descritor (ii) a pergunta de pesquisa, teríamos uma resposta aproximada que seria: ... *searches to grasp the complexities of language raising consciousness about the interdependence between discursive practices and ecological devastation* /procura apreender as complexidades da língua que conscientizam sobre a interdependência entre práticas discursivas e a devastação ecológica. Diante da ausência dos descritores três e quatro, entendemos que esta definição é parcialmente análoga e parcialmente anômala, matematicamente representada pela proporção 50/50.

Diante das análises relacionadas às quatro definições de PD enciclopédico, ressaltamos que o maior ou menor grau de analogia ou anomalia é determinado pela presença ou ausência dos traços distintivos ou descritores em um *corpus*. Quanto mais traços distintivos no *corpus*, mais análoga será a definição e vice-versa. Outro dado relevante é quanto ao objeto de estudo de cada subárea: cada uma delas tem um objeto de estudo e esse objeto é refletido no paradigma definicional (PD) e esse objeto de estudo implica na presença de concepções, teorias, abordagens, métodos, metodologias e aplicabilidades que, certamente, reverberam na definição do termo.

Um dos objetivos desta tese foi o de fazer a comparação entre o sub*corpus* de LD e de LA ao *corpus* de manuais de Linguística, a fim de constataremos a maior presença ou ausência de contextos definitórios úteis à construção das definições dos termos. Para executar essa comparação, realizamos os seguintes procedimentos: Primeiramente, buscamos equilibrar o *corpus* de manuais de Linguística (ML) com o *corpus* de LD e LA, considerando que os ML contavam aproximadamente um milhão de itens. Devido ao fato de que cada subárea contém 500 mil itens, adicionamos uma subárea da LD e uma da LA para serem comparadas ao *corpus* de ML.

Em seguida, efetuamos a pesquisa utilizando o concordanciador e constatamos que não há um fator preponderante para determinar em qual tipo de *corpus* os contextos definitórios são mais frequentes. Esta conclusão é proveniente da busca realizada nos *corpora* utilizando a expressão o ST ou a combinação ST + o verbo **ser**. Depois, quantificamos essas ocorrências no *corpus*, conforme descrito no Quadro 13 abaixo, onde CD indica contexto definitório e FPL representa Formação do Professor de Língua.

Quadro 13 – Contextos definitórios do *corpus* – amostra de análise.

Expressão de busca	Corpus de:	Subárea da	Nº CD
1. Linguagem é	LA	Terminografia	110
	LD	Terminologia	13
	ML		8
2. termo é	LA	Terminografia	57
	LD	Terminologia	48
	ML		8
3. hiponímia é	LA	Terminografia	0
	LD	Terminologia	1
	ML		6
4. sentido é	LA	Linguística matemática	7
	LD	Semântica	27
	ML		39
5. <i>corpus</i> é	LA	Linguística de contato	1
	LD	Linguística histórica	5
	ML		0
6. aprendizado	LA	FPL	16
	LD	Aquisição da linguagem	26
	ML		3

Fonte: Elaborada pelo autor.

Após o procedimento de equilibrar os *subcorpora*, obtivemos o número de contextos definitórios proveniente de um *corpus* de dois milhões de palavras. Os *subcorpora* nos quais os contextos definitórios foram mais frequentes estão marcados em negrito no quadro. Considerando as seis combinações, obtivemos contextos definitórios equilibrados para os três *subcorpora*: (i) na combinação *linguagem* e *termo* + *ser*, houve mais contextos definitórios na subárea da Terminografia (LA); (ii) na combinação *hiponímia* e *sentido* + *ser*, houve mais contextos definitórios nos ML; e (iii) na combinação *corpus* e *aprendizado* + *ser*, houve mais contextos definitórios nas subáreas da Linguística Histórica e Aquisição da linguagem, ambas da LD.

No que tange ao quesito de apresentar a obra ao consulente, a AJUDA/HELP do VoBLing foi concebida partindo da ideia das obras tradicionais. Dito de outra forma, era nosso objetivo apresentar o trabalho terminográfico de forma geral, instruir os consulentes sobre suas especificidades e ensiná-los a navegar na plataforma por meio do uso correto das ferramentas de busca e dos *hiperlinks*. O domínio da plataforma dever ser útil à construção do conhecimento, *easy-to-use*, e de caráter multimodal, ou seja, conjugando textos escritos, vídeos, áudio e imagens. Na próxima seção apresentamos as Considerações Finais dos quatro anos de concepção, elaboração e concretização do VoBling. Boa leitura!

8. CONSIDERAÇÕES FINAIS

Tecer as considerações finais de um trabalho que vem sendo conduzido há mais de uma década não é uma tarefa fácil. Há várias mãos que tocaram neste *corpus*, dando a ele várias formas, e com certeza a forma deste trabalho é somente mais uma delas. A fim de não fugir do caráter científico, teórico-metodológico e prático deste trabalho, buscamos apresentar ao leitor os passos que permearam a concepção, a condução e a concretização do que denominamos VoBLing, produto disponibilizado aos consulentes cibernéticos.

Tendo em vista o aperfeiçoamento e a continuidade deste projeto, o passo metodológico de balanceamento dos *corpora* quanto aos gêneros textuais será necessário, de forma que haja equilíbrio na quantidade de itens provenientes de artigos, dissertações, teses, livros, manuais e material instrucional. Apesar de termos em mãos um *corpus* de especialidade considerado grande, avaliamos que o trabalho rumo a um vocabulário bilíngue de Linguística completo está em seus passos iniciais, porém com futuro promissor que envolva universidades, discentes e docentes do Curso de Letras e áreas correlatas. Haja vista a constatação de que a produção de vocabulários bilíngues de Linguística no Brasil ainda é uma prática recente, vemos o VoBLing como um dos primeiros passos rumo ao “novo normal” da produção terminográfica on-line no Brasil e para o público de LP.

Neste trabalho, desenvolvemos a terminologia de três níveis da árvore de domínio da Linguística, quais sejam: a Linguística, a LD, a LA e suas 47 subáreas. É um trabalho minúsculo e modesto diante do imenso mundo de terminologias da Linguística, mas a exemplo de outros autores como Costa (2018), novas edições se fazem necessárias com o aumento de verbetes. A partir da perspectiva da LC, há que considerar o aumento do *corpus* e a possibilidade de expansão para outras línguas. Nossa intenção era de realizar um trabalho maior e mais abrangente, contudo, a chegada inesperada de uma pandemia que assolou o país e o mundo nos impediu de ter acesso à opinião do nosso público-alvo devido à suspensão das atividades acadêmicas. Ademais, o isolamento social nos impediu de termos acesso às bibliotecas, *locus* esse de fundamental importância para o desenvolvimento desta pesquisa, vinculada à consulta aos dicionários. A segunda fase da pesquisa permitiria que tivéssemos acesso ao *feedback* dos alunos, a fim de implementar melhorias na plataforma e promover a adequação do produto ao público-alvo, o que não nos foi possível fazer.

Como trabalhamos com os três níveis da árvore de domínio da Linguística, temos que considerar a expansão desta pesquisa para os próximos níveis. A título de exemplificação, temos as subáreas da Semântica, da Onomástica e de outras subáreas que não fizeram parte do

escopo desta pesquisa. Os parâmetros para definirmos o caráter deste quarto nível é outra questão epistemológica: (i) voltarmos nossa atenção para o objeto de pesquisa das subáreas como Antroponímia e Toponímia, subordinadas à Onomástica, ou (ii) voltarmos nossa atenção para as teorias, como acontece nas subáreas da Semântica?

Outro aspecto a ser considerado é a abordagem pedagógica, principalmente nas áreas da Terminografia Pedagógica e da Lexicografia Pedagógica. Preocupar-nos com a acessibilidade terminológica e com o aspecto pedagógico das terminologias é bastante pertinente devido à célere mudança de paradigmas do conhecimento e das tecnologias na atualidade, compondo o contexto do “novo normal” da humanidade. O modelo utilizado pelo *ThoughtCo - Lifelong Learning*⁸⁴ é um bom exemplo que une as tecnologias e a multimodalidade dos textos, além do aspecto de alcance mundial pela internet.

A acessibilidade textual e terminológica (ATT) proposta por Finatto (2020) visa trabalhar o texto escrito, no sentido de promover sua compreensão e interpretação de forma inteligível. Para atingir este objetivo, é necessário adotar uma metodologia que trabalhe o conteúdo, abrangendo todas as estruturas do texto: o léxico, a terminologia, a sintaxe e a semântica, e o torne familiar ao leitor.

A ATT é desenvolvida a partir dos conceitos de complexidade, simplificação e acessibilidade. Em primeiro lugar, a complexidade é definida a partir do questionamento de quão adequado um texto é ao leitor. Para compreender esta complexidade, é importante considerar características como “tema e estilo do texto, uso de terminologias, de vocabulário, tipo de organização sintática ou de tipos de frases, entre outros elementos” (FINATTO, 2020, p. 80). Em segundo lugar, a simplificação textual (ST) é o processo de reescrita ou de produção de um texto direcionado a um perfil específico de leitor e pode incluir imagens e outros recursos audiovisuais que promovem a maior compreensão do conteúdo. Finalmente, a acessibilidade é o resultado do processo de produção textual que se adequa ao seu público-alvo ou pode ser decorrente de um texto simplificado que visa atender aos princípios de comunicação de informação.

A aplicação da ATT na construção do VobLing pode ser observada no PD dos verbetes no uso de paráfrases, no uso de vídeos que explicam os conceitos da Linguística e nas imagens que ajudam o consulente a localizar os termos nas subáreas da Linguística. Futuramente, os desdobramentos do uso desses recursos podem gerar uma plataforma pedagógica multimodal

⁸⁴ Disponível em <https://www.thoughtco.com/>. Acesso em 30 maio de 2020.

com atividades de ensino-aprendizagem que utilize vários suportes e atenda a diversos perfis de usuários com atividades de compreensão e produção oral e escrita.

Quanto à abrangência da pesquisa, ela pode ser contextualizada em outras áreas do conhecimento, devido ao caráter interdisciplinar da terminologia e a da aplicabilidade da Terminologia de *Frames* para análise de textos científicos de outras áreas do conhecimento. O envolvimento das tecnologias e de especialistas de outras áreas permitem uma integração de várias habilidades num grupo de pesquisa e trabalho em prol do avanço da Terminologia. A Terminologia bilíngue pode promover cooperações internacionais como concretizado no protótipo deste projeto envolvendo profissionais e instituições voltadas ao avanço do conhecimento.

Há desafios tecnológicos a serem vencidos, como no caso do tratamento da homonímia *language* em inglês, que pode ter dupla significação no português, como em língua e linguagem. Como uni-los na plataforma, de forma que não causem confusão ao consulente e não deixem o sistema cibernético confuso quanto à administração de termos iguais que possuem significações diferentes?

Ao considerarmos o impacto científico e social da obra, devemos conceber a ideia de disponibilizar o *corpus* de várias formas, de Linguística como um todo ou segmentado em LD, LA, separado por subárea, de teses, de dissertações ou manuais. Um aspecto bastante recorrente na introdução das obras é a motivação que leva os autores a produzirem o dicionário: a busca da padronização ou da harmonização dos termos e conceitos da Linguística e suas subáreas. Este desafio perdura até os dias de hoje e perdurará por muito tempo, enquanto o conhecimento e a ciência forem virtudes crescentes da humanidade, o que sempre garantirá um espaço privilegiado para a Terminologia.

REFERÊNCIAS

- ALMEIDA, G. M. B.; ALUÍSIO, S. M.; OLIVEIRA, L. H. M. O método em Terminologia: revendo alguns procedimentos. *In*: ALVES, I. M.; ISQUERDO, A. N. (Org.). **As ciências do léxico**: lexicologia, lexicografia, terminologia. Vol. III. Campo Grande: Editora UFMS, 2007. p. 409-420.
- ALMEIDA, G. M. B.; CORREIA, M. Terminologia e *corpus*: relações, métodos e recursos. *In*: TAGNIN, S. E. O; VALE, O. A. **Avanços da Linguística de Corpus no Brasil**. São Paulo: Humanitas, 2008. p. 67-94.
- ALVES, D.; TAGNIN, S. E. O. Pela visibilidade da Linguística de *Corpus* em trabalhos acadêmicos. *In*: DUTRA, D. P.; MELLO, H. (Org.). **Anais do X Encontro de Linguística de Corpus**: aspectos metodológicos dos estudos de *corpora*. Belo Horizonte: Faculdade de Letras da UFMG, 2012, p. 13-37.
- ANDRADE, M. M. Lexicologia, terminologia: definições, finalidades, conceitos operacionais. *In*: ISQUERDO, A. N.; OLIVEIRA, M. P. P. **As ciências do léxico**: lexicologia, lexicografia, terminologia. Campo Grande: Editora UFMS, 2001. p. 191-200.
- AUBERT, F. H. **Introdução à metodologia da Pesquisa terminológica bilíngue**. São Paulo: FFLCH/CITRAT, 2001. Disponível em: http://citrat.fflch.usp.br/sites/citrat.fflch.usp.br/files/inline-files/Cad.%20Terminologia%202_0.pdf. Acesso em: 27 jun. 2019.
- BAGNO, M.; RANGEL, E. O. Tarefas da educação linguística no Brasil. **Revista Brasileira de Linguística Aplicada**, Belo Horizonte, v. 5, n. 1. 2005, p. 63-81. Disponível em: <http://www.scielo.br/pdf/rbla/v5n1/04.pdf>. Acesso em: 4 maio 2019. <https://doi.org/10.1590/S1984-63982005000100004>.
- 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 [...]**. Brasília: Cnpq/Ibict, 1990. p. 152-158.
- BARBOSA, M. A. Estrutura e formação do conceito nas línguas especializadas: tratamento terminológico e lexicográfico. *Revista Brasileira de Linguística Aplicada*, Belo Horizonte, v. 4, n. 1, p. 55-86. 2004. <https://doi.org/10.1590/S1984-63982004000100006>.
- BARROS, L. A. Aspectos epistemológicos e perspectivas científicas da terminologia. **Cienc. Cult.**, São Paulo, v. 58, n. 2, p. 22-26. 2006. Disponível em: http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200011&lng=en&nrm=iso. Acesso em: 26 set. 2018.
- BARROS, L. A. **Conhecimentos de terminologia geral para a prática tradutória**. São José do Rio Preto: NovaGraf, 2007.
- BERBER SARDINHA, T. Linguística de Corpus: histórico e problemática. **D.E.L.T.A.**, São Paulo, v. 16, n. 2, p. 323-367. 2000. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-

44502000000200005&lng=pt&tlng=pt Acesso em: 28 ago. 2018.
<http://dx.doi.org/10.1590/S0102-44502000000200005>.

BERBER SARDINHA, T. **Linguística de Corpus**. São Paulo: Manole, 2004.

BERBER SARDINHA, T. **Pesquisa em Linguística de Corpus com WordSmith Tools**. Campinas: Mercado de Letras, 2009.

BERTELS, A. The dynamics of terms and meaning in the domain of machining terminology. In: **Terminology** 17 (1). 2011, p. 94-112. DOI: 10.1075/term.17.1.06ber
<https://doi.org/10.1075/term.17.1.06ber>

BESSÉ, B. de, NKWENTI-AZEH, B. SAGER, J.C. Glossary of terms used in Terminology. In: **Terminology** 3, Amsterdam: John Benjamins, 1998. p. 119-156.
<https://doi.org/10.1075/term.4.1.08bes>

BEVILACQUA, C. R. Unidades fraseológicas especializadas: novas perspectivas para sua identificação e tratamento. In: **Temas da terminologia**. KRIEGER, M. G.; MACIEL, A. M. B. (Org.). Porto Alegre/SP: Editoria UFRGS/Humanitas, 2001. p. 106-118.

BIDERMAN, M. T. C. As Ciências do Léxico. In: OLIVEIRA, A. M. P. P.; ISQUERDO, A.N. **As ciências do léxico: Lexicologia. Lexicografia. Terminologia**. Campo Grande/MS: Editora UFMS, 2001. p. 13-22.

CABRÉ, M. T. Terminologie ou terminologies? Spécialité linguistique ou domaine interdisciplinaire? **Meta**, Montreal, v. 36, n. 1, p. 55-63. 1991. Disponível em: <https://www.erudit.org/fr/revues/meta/1991-v36-n1-meta331/002184ar/>. Acesso em: 5 maio 2019. <https://doi.org/10.7202/002184ar>

CABRÉ, M. T. **La terminología: teoría, metodología, aplicaciones**. Tradução de Carles Tebé. Barcelona: Editorial Antártida/Empúries, 1993.

CABRÉ, M. T. **Terminology: theory, methods, and applications**. Philadelphia/PA: John Benjamins, 1999. <https://doi.org/10.1075/tlrp.1>

CABRÉ, M. T. El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en terminología. **Organon**, Porto Alegre, v. 1, n. 1, p. 1-17. 2011. <https://doi.org/10.22456/2238-8915.28343>

CABRÉ et al. **Application-driven terminology engineering**. Benjamins Current Topics (v. 2) 2007

CARDOSO, S. A. F. **Termos Teo: a elaboração de vocabulários monolíngues de termos da Teologia em um estudo conduzido por corpus**. 2017. Tese (Doutorado em Estudos Linguísticos) - Instituto de Letras e Linguística, Universidade Federal de Uberlândia, Uberlândia, 2017. Disponível em: <https://repositorio.ufu.br/bitstream/123456789/21349/1/TermosTeoElaboracao.pdf>. Acesso em: 30 maio 2020.

CASTILHO, A. T. As letras no ensino e na pesquisa. In: **Veredas online - ensino**. PPG - Linguística / UFJF: V. 2, 2007. p.05-21. Disponível em: <http://www.ufjf.br/revistaveredas/files/2009/12/artigo011.pdf>. Acesso em: 04 maio 2019.

COROMINAS, Joan; con la colaboración de José A. PASCUAL. **Diccionario crítico etimológico castellano e hispánico**. 1ª ed. 1ª reimp. 5 v. Madrid: Gredos, 1984.

CHOMSKY, N. **Syntactic Structures**. Haia: Mouton, 1957.

COSERIU, E. **Teoria da linguagem e linguística geral**: cinco estudos. Rio de Janeiro: Presença, 1979.

COSTA, H. R. **O discurso historiográfico da linguística aplicada brasileira**. 2011. Dissertação (Mestrado em Letras e Linguística) - Faculdade de Letras, Universidade Federal de Goiás, Goiânia, 2011. Disponível em <https://files.cercomp.ufg.br/weby/up/26/o/hildarodriguesdacosta.pdf>. Acesso em: 4 jun. 2019. DAUZAT (1972)

DORNELLES, M. S. **Bases teórico-metodológicas para elaboração de um glossário bilíngue (português-inglês) de treinamento de força**: subsídios para o tradutor. 2015. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2015. Disponível em: <https://lume.ufrgs.br/handle/10183/117567>. Acesso em: 15 jun. 2017.

DUARTE NUNES, E. A pesquisa narrativa em saúde. **Interface**, Botucatu, v. 22, n. 64, p. 307-312. 2018. Disponível em: <http://dx.doi.org/10.1590/1807-57622017.0240>. Acesso em: 15 jun 2018.

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

DUCROT, O.; TODOROV, T. **Dictionnaire encyclopédique des sciences du langage**. Paris: Seuil, 1972.

FABER BENÍTEZ, P., et al. Framing Terminology: A Process-Oriented Approach. In: **Meta**, volume 50, número 4, décembre 2005, p. ?-?. DOI: <https://doi.org/10.7202/019916ar> Disponível em: <https://www.erudit.org/fr/revues/meta/2005-v50-n4-meta1024/019916ar.pdf> Acesso em 19 ab. 2020.

FADANELLI, S. B.; MONZÓN, A. J. B. Gêneros textuais datasheet e artigo científico em aulas de ESP. **Domínios de Linguagem**, Uberlândia, v. 11, n. 2, p. 351-378. 2017.

FAULSTICH, E. A socioterminologia na comunicação científica e técnica. **Cienc. Cult.**, São Paulo, v. 58, n. 2, p. 27-31. 2006.

FAULSTICH, E. Avaliação de dicionários: uma proposta metodológica. **Organon**, Porto Alegre, v. 25, n. 50, p. 1-23. 2011. Disponível em: <http://seer.ufrgs.br/index.php/organon/article/view/28346/16994>. Acesso em: 15 abr. 2020. DOI: <https://doi.org/10.22456/2238-8915.28346>.

FILLMORE, Charles J. An Alternative to Checklist Theories of Meaning. In: **Proceedings of the First Annual Meeting of the Berkeley Linguistics Society**, 1975, p. 123-131.

FILLMORE, Charles J. Frame semantics and the nature of language. In: **Annals of the New York Academy of Sciences**, 280 (1 Origins and E), 1976, p. 20 - 32. DOI:10.1111/j.1749-6632.1976.tb25467.x

FILLMORE, Charles J. Double-Decker Definitions: The Role of *Frames* in Meaning Explanations. In: **Sign Language Studies**, Volume: 3, Issues: 3, Gallaudet University Press, 2003, p. 263-295. DOI: 10.1353/sls.2003.0008

FILLMORE, Charles J. Frame Semantics. In: **Encyclopedia of Language & Linguistics**, Elsevier, 2006, p. 613 - 620. DOI:10.1016/b0-08-044854-2/00424-7

FINATTO, Maria José Bocorny (1994). Caracterização de paradigmas definicionais terminológicos. In: Simpósio Iberoamericano de Terminologia, 4, 1994, Buenos Aires. Actas... Buenos Aires: União Latina, 1994. p.55-59.

FINATTO, Maria José Bocorny. A definição terminológica do dicionário TERMISUL: expressões linguísticas de relações conceituais complexas. In: ISQUERDO, A. N.; OLIVEIRA, M. P. P. **As ciências do léxico**: lexicologia, lexicografia, terminologia. Campo Grande: Editora UFMS, 2001a. p. 191-200.

FINATTO, Maria José Bocorny. Terminologia e Ciência cognitiva. In: KRIEGER, M. G.; MACIEL, A. M. B. (Org.). **Temas da terminologia**. Porto Alegre/SP: Editora UFRGS/Humanitas, 2001b. p. 141-149.

FINATTO, Maria José Bocorny. Orientações para a Terminografia: das teorias às práticas em busca de amplitude da informação terminológica. In: ISQUERDO, A. N.; DAL CORNO, G. O. M. (Org.). **As Ciências do Léxico**: Lexicologia, Lexicografia e Terminologia. vol. VII. Campo Grande/MS: Editora UFMS, 2014. p. 439-459.

FINATTO, Maria José Bocorny. Acessibilidade textual e terminológica: promovendo a tradução intralinguística. In: **Estudos Linguísticos** (São Paulo. 1978), v. 49, p. 72-96, 2020. Disponível em: <http://https://revistas.gel.org.br/estudos-linguisticos/article/view/2775>. Acesso em: 30 jun. 2020. DOI: <https://doi.org/10.21165/el.v49i1.2775>

FINATTO, Maria José Bocorny.; ZILIO, Leonardo. (Org.). **Textos e termos por Lothar Hoffmann**. Porto Alegre: Palotti, 2015.

FIORIN, J. L. (Org.). **Introdução à Linguística I**: objetos teóricos. São Paulo: Contexto, 2002.

FIORIN, J. L. La création des cours de lettres au Brésil et les premières orientations de la recherche linguistique universitaire. In: GUIMARÃES, E.; BARROS, D. L. P. (Ed.) **History of linguistics 2002**: selected papers from the Ninth International Conference on the History of the Language Sciences. Amsterdam: John Benjamins B.V., 2007.

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

Disponível em: https://www.teses.usp.br/teses/disponiveis/8/8147/tde-08072008-150855/publico/TESE_GUILHERME_FROMM.pdf. Acesso em: 30 maio 2020.

FROMM, Guilherme. A construção e análise de *corpora* para alimentação de um banco de dados terminográfico: um exemplo. In: **Domínios de Linguagem** – Revista Eletrônica de Linguística, v. 2, 22 p., 2008. Disponível em: <http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/11483>. Acesso em: 25 jul. 2020.

FROMM, Guilherme. A questão da taxonomia num *corpus* colaborativo para construção de um vocabulário na área de Linguística. In: XIV Simpósio Nacional De Letras E Linguística E IV Simpósio Internacional De Letras E Linguística, 2013, Uberlândia. **Anais** [...]. Uberlândia: EDUFU, 2013. p. 1-9.

FROMM, Guilherme. Vocabulário de Linguística: treinamento em Terminografia Bilíngue, uso de *corpora* e ambiente de gestão terminológica. In: ISQUERDO, A. N.; DAL CORNO, G. O. M. (org.). **As ciências do léxico: lexicologia, lexicografia, terminologia**. vol. VIII. Campo Grande: Editora UFMS, 2018. p. 309-328.

FROMM, Guilherme. Por uma terminografia pedagógica. In: **Estudos Linguísticos** (São Paulo. 1978), v. 49, p. 761-776, 2020. Disponível em: <http://https://revistas.gel.org.br/estudos-linguisticos/article/view/2637>. Acesso em: 30 jun. 2020.

FROMM, Guilherme; YAMAMOTO, Márcio Issamu. Terminologia, terminografia, tradução e Linguística de *corpus*: a criação de um vocabulário bilíngue sobre Linguística. In: TAGNIN, S. E. O.; BEVILACQUA, C. R. **Corpora na Terminologia**. São Paulo: Hub Editorial, 2013, p. 129-152.

FROMM, Guilherme; YAMAMOTO, Márcio Issamu. A microestrutura em verbetes da área da Linguística. **Revista de Estudos da Linguagem**, Belo Horizonte, v. 28, n. 1, p. 205-234. 2020. Disponível em: <http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/15255>. Acesso em: 7 abr. 2020. DOI: <http://dx.doi.org/10.17851/2237-2083.28.1.205-234>.

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

GOMES, P. **Fruticultura Brasileira**. Nobel: São Paulo, 1989.

GONÇALVES, T. V. O.; NARDI, R. Aspectos epistemológicos da Pesquisa Narrativa presentes em teses e dissertações sobre formação de professores na área de Educação em Ciências e Matemática, no período de 2000 a 2012. Disponível em: <http://proceedings.ciaiq.org/index.php/ciaiq2016/article/download/704/692>. Acesso em: 15 jun. 2018.

HALLIDAY, M. A. K. **Spoken and written language**. Geelong Vict.: Deakin University, 1985.

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

ILARI, R. **Introdução à Semântica**. Brincando com a Gramática. São Paulo: Contexto, 2003.

ISO – INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **Norma 1087-1**: Terminology work – Vocabulary. Genebra, 2000.

JENKINS, J. English as a Lingua Franca from the classroom to the classroom. **ELT Journal**, Oxford, v. 66, n. 4, p. 486-494. 2012. Disponível em: <https://academic.oup.com/eltj/article/66/4/486/385078>. Acesso em: 24 set. 2019. DOI: <https://doi.org/10.1093/elt/ccs040>.

KAGEURA, K. Terminology and Lexicography. *In*: KOCKAERT, H. J.; STEURS, F. (Ed). **Handbook of Terminology**. Vol. 1. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2015.

KLEIN, Ernest. **Comprehensive Etymological Dictionary of the English Language**. Amsterdam: Elsevier Publishing, 1971.

KOCKAERT, H. J.; STEURS, F. (Ed). **Handbook of Terminology**. Vol. 1. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2015.

KRIEGER, M. G. A face linguística da Terminologia. *In*: **Temas da terminologia**. KRIEGER, M. G.; MACIEL, A. M. B. (Org.). Porto Alegre/SP: Editora UFRGS/Humanitas, 2001. p. 22-33.

KRIEGER, M. G. Sobre Terminologia e seus objetos. *In*: **Temas da terminologia**. KRIEGER, M. G.; MACIEL, A. M. B. (Org.). Porto Alegre/SP: Editora UFRGS/Humanitas, 2001. p. 34-38.

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

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

MACIEL, A. M. B. Terminologia, linguagem de especialidade e dicionários. *In*: KRIEGER, M. G.; MACIEL, A. M. B. (Org.). **Temas da terminologia**. Porto Alegre/SP: Editora UFRGS/Humanitas, 2001. p. 39-46.

MACIEL, A. M. B. Pressupostos sociocognitivos na descrição terminológica e produção terminográfica. *In*: ISQUERDO, A. N.; FINATTO, M. J. B. **As Ciências do Léxico**: Lexicologia. Lexicografia. Terminologia. Campo Grande/MS: Editora UFMS, 2010. p. 397-415.

MARTINS, S. C.; ZAVAGLIA, C. **A onomasiologia e seus dicionários**: o caso do dicionário onomasiológico de expressões cromáticas da fauna e flora. **Diacrítica**, Braga. v. 28, n. 1, p. 437-455. 2014. Disponível em http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S0807-89672014000100017&lng=pt&nrm=iso. Acesso em: 27 jun. 2018.

MCENERY, A.; XIAO, R. Character encoding in corpus construction. WYNNE, M. (Ed.). **Developing Linguistic Corpora**: a Guide to Good Practice. Oxbow Books, 2005. p. 58-72.

MCENERY, T.; WILSON, A. **Corpus linguistics**: an introduction. Edinburgh: Edinburgh University Press, 2001.

MCENERY, T.; HARDIE, A. **Corpus Linguistics**: Method, Theory and Practice. Cambridge: Cambridge University Press, 2012.

MOTTA-ROTH, D. The role of context in academic text production and writing pedagogy. In: BAZERMAN, C.; BONINI, A.; FIGUEIREDO, D. (Ed.). **Genre in a changing world**. Indiana: Parlor Press, 2009, p. 317-336.

NASCENTES, A. **Dicionário Etimológico da Língua Portuguesa**. Rio de Janeiro: Livraria Acadêmica, 1955.

OLIVEIRA, F. P. **ToGatherUp**: um protótipo de ferramenta para a construção de corpora a produção de vocabulários bilíngues direcionada por *corpus*. 2019. Dissertação (Mestrado em Estudos Linguísticos) – Instituto de Letras e Linguística, Universidade Federal de Uberlândia, Uberlândia, 2019. Disponível em: <https://repositorio.ufu.br/bitstream/123456789/25433/1/ToGatherUpProtótipoFerramenta>. Acesso em: 30 maio 2020.

PARÉ, A.; STARKE-MEYERRING, D.; MCALPINE, L. The dissertation as multi-genre: many readers, many readings. In: BAZERMAN, C.; BONINI, A.; FIGUEIREDO, D. (Ed.). **Genre in a changing world**. Indiana: Parlor Press, 2009, p. 179-193.

PARTRIDGE, E. **Origins**: A Short Etymological Dictionary of Modern English, 4th ed.; with numerous revisions and some substantial additions; Routledge and Kegan Paul: London, UK, 1966.

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

PERFOLL JUNIOR, A.; MODRO, N. R. Avaliação da qualidade em uso de um software educacional: um estudo aplicado ao SENAI/SC. In: **REAVI**, Alto Vale do Itajaí, v. 5, n. 7, p. 88-108. 2016. Disponível em: http://www.ceavi.udesc.br/arquivos/id_submenu/787/ademar_perfoll_junior.pdf. Acesso em: 30 abr. 2020. DOI: 10.5965/2316419005072016088.

POTTIER, B. **Sémantique générale**. Paris: PUF, 1992.

REY, A. **Essays on terminology**. Tradução e edição de Juan Sager. Philadelphia: John Benjamins B.V., 1995.

RIBEIRO, J. **Dicionário gramatical**. 3. ed. Rio de Janeiro: Francisco Alves, 1906.

RODRIGUES, C. R. S. A importância do léxico no processo de formação de palavras. In: ZAIDAN, J. C. S. M.; SOARES, L. E. (org.). **Letras por dentro III**. Vitória: PPGEL/MEL, 2006. Disponível em: https://www.academia.edu/3268456/Letras_por_Dentro_III. Acesso em: 14 nov. 2018.

ROBINS, R.H. **A Short History of Linguistics**. 4th ed. London: Longman, 1997.

RONDEAU, G. **Introduction à la terminologie**. 2. ed. Québec: Gaëtan Moran, 1984.

SAGER, J. C. **Curso práctico sobre el procesamiento de la terminología**. Traducción de Laura Chumillas Moya. Madrid: Pirámide, 1993.

SAGER, J. C. **A practical course in terminology processing**: with a bibliography by Blaise Nkwenti-Azeh. Philadelphia: John Benjamins Publishing Company, 1990.

SAUSSURE, F. de. **Curso de Linguística Geral**. Organização Charles Bally e Albert Sechehaye; com a colaboração de Albert Riedlinger. Tradução de Antônio Chelini, José Paulo Paes e Izidoro Blikstein. 28. ed. São Paulo: Cultrix, 2012.

SCHMITZ, K. Terminology and localization. *In*: KOCKAERT, H. J.; STEURS, F. (Ed.). **Handbook of Terminology**. Vol. 1. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2015, p. 453-456.

SCOTT, M. **WordSmith Tools**. Version 7. Stroud: Lexical Analysis Software, 2016.

SCOTT, M. **WordSmith Tools**. Version 8. Stroud: Lexical Analysis Software, 2020.

SOUZA, Diego Spader de. **Entre conceitos e conce(p)tos**: uma proposta teórico-metodológica na interface entre a onomasiologia, a lexicografia e a semântica lexical cognitiva. 2019. Tese (Doutorado) Programa de pós-graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos, São Leopoldo, RS, 2019. Disponível em: <http://www.repositorio.jesuita.org.br/handle/UNISINOS/7965>. Acesso em 13 de maio de 2020.

STEURS, F.; DE WATCHER, K.; DE MALSCHE, E. Terminology Tools. *In*: KOCKAERT, Hendrik J.; STEURS, Frieda (Ed.). **Handbook of Terminology**. Vol. 1. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2015, p. 222-227.

TAGNIN, S. E. O.; BEVILACQUA, C. R. (Org.). **Corpora na tradução**. São Paulo: Hub Editorial, 2015.

TAGNIN, S. E. O.; BEVILACQUA, C. R. Por que a linguística de *corpus* na terminologia. *In*: TAGNIN, S. E. O.; BEVILACQUA, C. R. (Org.). **Corpora na terminologia**. São Paulo: Hub Editorial, 2013, p. 11-28.

TALAVÁN, N. **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 (Doutorado em Estudos Linguísticos e Literários em Inglês) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2008. Disponível em: https://www.teses.usp.br/teses/disponiveis/8/8147/tde-16022009-141747/publico/TESE_ELISA_DUARTE_TEIXEIRA.pdf. Acesso em: 30 maio 2020.

TEMMERMAN, R. **Towards new ways of Terminology description**: the sociocognitive approach. Amsterdam; Philadelphia: John Benjamins, 2000.

TRAVAGLIA, L. C. **A caracterização de categorias de textos**: tipos, gêneros e espécies. **Alfa: Revista de Linguística**, v. 51, n. 1, p. 39-79. 2007. Disponível em: <https://periodicos.fclar.unesp.br/alfa/article/view/1426/1127>. Acesso em: 30 maio 2020.

VANDRESEN, P. **A Linguística no Brasil**. Disponível em <http://aldobizzocchi.com.br/texto9.html>. Acesso em: 27 jul. 2018.

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

WIERTLEWSKA, J. Ecolinguistic approach to foreign language teaching on the example of English. **Glottodidactica**, [s.l.], v. 37, p. 141-151. 2011. Disponível em: <https://pressto.amu.edu.pl/index.php/gl/article/view/320/227>. Acesso em: 24 set. 2019.

WÜSTER, E. **Introducción a la teoría general de la terminología y a la lexicografía terminológica**. Traducción de Anne-Cécile Nokerman. Barcelona: Institut Universitari de Lingüística Aplicada, 2003.

YAMAMOTO, Márcio Issamu. **Linguística histórica e linguística de corpus-Caminhos que se cruzam para desvelar a história da linguagem**: um vocabulário bilíngue português – inglês. 2015. Dissertação (Mestrado em Estudos Linguísticos) – Instituto de Letras e Linguística, Universidade Federal de Uberlândia, Uberlândia, 2015. Disponível em: <https://repositorio.ufu.br/bitstream/123456789/15483/1/LinguisticaHistoricaLinguistica.pdf>. Acesso em: 30 maio 2020.

YAMAMOTO, Márcio Issamu. Vocabulário português/inglês de Linguística Geral. **Revista Philologus**, Rio de Janeiro, v. 24, n. 70, p. 272-297. 2018. Disponível em: <http://www.filologia.org.br/rph/ANO24/70supl/023.pdf>. Acesso em: 30 maio 2020.

YAMAMOTO, Márcio Issamu; LISBOA, Joel Victor Reis. Corpus linguistics and phraseologies on TV series. **Itinerarius Reflectionis**, v. 15, n. 2, p. 1-19. 2019. Disponível em: <https://www.revistas.ufg.br/rir/article/view/58713/33207>. Acesso em: 30 maio 2020.

REFERÊNCIA DE MANUAIS DE LINGUÍSTICA

AARTS, B.; MCMAHON, A. (Ed.). **The handbook of English linguistics**. Hoboken (NJ): Wiley Blackwell, 2006.

ARONOFF, Mark; REES-MILLER, Janie (eds.) **The handbook of linguistics**. Oxford, UK: Blackwell, 2003.

BRASIL. **Manual de Linguística**: subsídios para a formação de professores indígenas na área de linguagem. SECADI. Brasília: LACED/Museu Nacional, 2006.

CÂMARA JUNIOR., J. M. **Manual de expressão oral e escrita**. Petrópolis: Vozes, 1991.

DAVIES, A.; ELDER, C. (Ed.) **The Handbook of Applied Linguistics**. Oxford: Blackwell, 2004.

FIORIN, J.L. **Introdução à linguística II: princípios de análise**. São Paulo: Contexto, 2003.

JAKOBSON, R. **Linguística e comunicação**. São Paulo: Cultrix, 1969.

LYONS, J. **Introdução à linguística teórica**. [Trad. de Rosa Virgínia Mattos e Silva e Hélio Pimentel do orig. inglês de 1968.] São Paulo: Nacional: Universidade de São Paulo, 1979.

MUSSALIM, F.; BENTES, A. C. (Org.). **Introdução à Linguística: Domínios e Fronteiras**. Volumes 1 e 2. São Paulo: Cortez Editora, 2004.

MUSSALIM, F.; BENTES, A. C. (Org.). **Introdução à linguística: fundamentos epistemológicos**. 5. ed. São Paulo: Cortez, 2011. v. 3.

PAIS, C. T. e RECTOR, M. (org.). **Manual de linguística**. Petrópolis: Vozes, 1979.

ROBINS, R. H. **General Linguistics**. An Introductory Survey. London: Longmans, Green and Co., 1980.

SCHWINDT, L. C. (org.). **Manual de linguística: fonologia, morfologia e sintaxe**. Petrópolis: Vozes, 2014.

VIDOS, B. E. **Manual de linguística românica**. Tradução de José Pereira da Silva. Apresentação de Evanildo Bechara. Rio de Janeiro: Eduerj, 1996.

WIDDOWSON, H. G. **Linguistics**. Oxford: Oxford University Press, 1996.

REFERÊNCIA DE DICIONÁRIOS DE LINGÜÍSTICA ANALISADOS

GRUPO 1: DICIONÁRIOS IMPRESSOS E ELETRÔNICOS

CÂMARA JUNIOR., J. M. **Dicionário de linguística e gramática**. 13. ed. Petrópolis: Vozes, 1986.

COSTA, S. R. **Dicionário de gêneros textuais**. 3. ed. rev. e ampl., 1. reimpr. Belo Horizonte: Autêntica, 2018.

CRYSTAL, D. **Dicionário de linguística e fonética**. Trad. e adapt.: Maria Carmelita Pádua Dias. Rio de Janeiro: Jorge Zahar, 2000.

DUBOIS, J. et al. **Dicionário de linguística**. Trad.: Izidoro Blikstein et. al. São Paulo: Cultrix, 1973.

DUCROT, O.; TODOROV, T. **Dicionário enciclopédico das ciências da linguagem**. São Paulo: Perspectiva, 1977.

FLORES, V. do N. et al. **Dicionário de linguística da enunciação**. Prefácio de José Luiz Fiorin. São Paulo: Contexto, 2009.

GREIMAS, A. J.; COURTÉS, J. **Dicionário de semiótica**. 2. ed. 1ª reimpr. São Paulo: Contexto, 2012.

MOISÉS, M. **Dicionário de termos literários**. São Paulo: Cultrix, 2001.

MORAIS, O. M. de. **Dicionário de gramática**. 7. ed. aum. e atual. Rio de Janeiro: Científica, 1965.

SILVA, J. P. Bases para um dicionário linguístico-gramatical. In: Revista Philologus, v. 25, p. 237-247, 2019.

TRASK, R. L. **Dicionário de linguagem e linguística**. Trad.: Rodolfo Ilari; revisão técnica: Ingedore Villaça Koch e Thaís Cristófaros Silva. 3. ed., 1. reimpr. São Paulo: Contexto, 2015.

GRUPO 2: DICIONÁRIOS DISPONÍVEIS ON-LINE

ACKEMA, Peter et al. **Lexicon of linguistics**. Disponível em: <https://lexicon.hum.uu.nl/>. Acesso em 01 de ago. de 2020.

APPLEYARD, David V.: **Bilingual Grammar Glossary for Japan**. Disponível em: <https://davidappleyard.com/english/jgrammar.htm>. Acesso em 01 de ago. de 2020.

GRUPO DE CIENCIAS DEL LENGUAGE. **Lingüística-TÉRMINOS**. Disponível em: <http://www.diccionariosdigitales.net/glosarios%20y%20vocabularios/ciencias%20del%20enguaje-4-linguistica-terminos.htm>. Acesso em 01 de ago. de 2020.

JOHANSON, Bill. **Glossary of Grammar Terms**. Disponível em <http://www.dailygrammar.com/glossary.html>. Acesso em 01 de ago. de 2020.

LITHUANIAN TERMINOLOGICAL SYSTEM OF CORPUS LINGUISTICS. **Systematic Dictionary of Corpus Linguistics**. Disponível em: <http://donelaitis.vdu.lt/publikacijos/SDoCL.htm>. Acesso em 01 de ago. de 2020.

LOOS, Eugene E. **SIL Glossary of Linguistic Terms**. Disponível em: <https://glossary.sil.org/>. Acesso em 01 de ago. de 2020.

MEINERS, Jocelly. **Index of French Grammar and Pronunciation Glossary Terms**. Disponível em: <http://www.dailygrammar.com/glossary.html>. Acesso em 01 de ago. de 2020.

APÊNDICES

APÊNDICE A - Questionário de pesquisa para escolha do paradigma definicional.

Pesquisa com os alunos do curso de Letras Inglês – UFU & UFJ

Esta pesquisa diz respeito a diferentes tipos de elaboração de definições. A resposta será utilizada como dado para a pesquisa que está desenvolvendo o Vocabulário bilíngue português-inglês de Linguística

Agradeço sua colaboração.

Assinale, entre as alternativas a seguir, aquela que melhor define cada termo em destaque:

B. HAPAX LEGOMENA

5. () **Hapax legomena** – Sabemos que algumas das palavras correm o risco de terem sido empregadas uma só vez e por uma só pessoa. Contudo, optamos por apresentá-las com sua frequência única - as hapax legomena.
6. () **Hapax legomena** – Somente uma variante foi encontrada na busca: prepina. Tal variante foi encontrada apenas em um contexto, se constituindo, dessa forma, hapax legomena.
7. () **Hapax legomena** – Dentre as palavras de uma obra escrita, literária ou não, é aquela que aparece uma única vez.
8. () **Hapax legomena** – Neologismo genuíno.
9. () **Hapax legomena** – Do grego hápax, "uma única vez", e legomenon, "dito", considerada a ocorrência única de uma palavra em uma obra escrita; concebido como neologismo, a partir da perspectiva da produtividade morfológica, ou formações analógicas como em "enxadachim" (Guimarães Rosa), e *bebemorar* e *trêbados*, comuns atualmente.

C. TRAÇOS DISTINTIVOS

1. () **Traços distintivos** – Traços que podem ser tanto articulatórios quanto acústicos.
2. () **Traços distintivos** – O que difere [p] de [b] são os seus traços distintivos. Qual a diferença em faca e vaca?
3. () **Traços distintivos** – Elemento composto de traços diferentes, que serão os menores elementos da cadeia falada como único som.
4. () **Traços distintivos** – Traços específicos.
5. () **Traços distintivos** – Traços peculiares de som ou dos grupos fraseológicos que determinam as unidades mínimas formadoras dos sinais; podem ser acústicos ou articulatórios para vogais ou consoantes; a idiosincrasia e a fixação são os principais traços distintivos dos grupos fraseológicos.

D. LÉXICO

1. () **Léxico** – Organizado separadamente em palavras que denominam entidades concretas [...]
2. () **Léxico** - No nível do léxico há o uso de vocabulário popular ou gírio, comum na oralidade; no nível da sintaxe, os diálogos podem ser marcados por repetições, paráfrases, cortes e correções.

3. () **Léxico** – Conjunto de signos linguísticos que representa as experiências de uma sociedade, cristalizando os conceitos em palavras e socializando-as para a comunicação e a interação social.
4. () **Léxico** – Tesouro vocabular.
5. () **Léxico** – Constituído por unidades lexicais, palavras e morfemas; unidades representadas por uma associação entre som e significado, uma realidade psicológica [...]

E. TOPÔNIMO

1. () **Topônimo** – Um pequeno texto, veículo de ideologia.
2. () **Topônimo**– Dauzat (1972) fornece exemplos de cidades gaulesas que tiveram seus nomes celtas substituídos por nomes latinos. As estratégias de romanização - a arquitetura, os fatos culturais como anfiteatros, arenas e termas, as estradas, [...] - incluíam também, portanto, o nome.
3. () **Topônimo**– Resultado da ação do nomeador ao praticar um papel de registro de uma porção de um espaço representado numa palavra.
4. () **Topônimo** – Nome próprio.
5. () **Topônimo**– O termo topônimo é formado pelos radicais gregos: topos + ónimo, significando lugar e nome respectivamente, ou seja, literalmente o nome de um lugar; resultado da ação do nomeador ao realizar um recorte no plano das significações, das representações; signo duplamente motivado, pois possui o motivo semântico e o motivo intencionalidade para compor aquele enunciado.

F. ESTILÍSTICA

1. () **Estilística** – Sua essência está no romper limites, permitindo-se brincar com as palavras e expressões.
2. () **Estilística** – O discurso indireto livre é um bom exemplo de como se entrosam o elemento individual e o coletivo em matéria de Estilística.
3. () **Estilística** – é todo o aparato afetivo, emocional e linguístico que caracteriza a expressividade do autor.
4. () **Estilística** – complemento da gramática.
5. () **Estilística** – No final do século XIX, a Linguística é dependente da gramática, sendo regulada por esta, no século XX, ela se torna uma disciplina da língua portuguesa com seu próprio objeto; a representação política da relação constitutiva (imaginária) entre sujeito e língua...

APÊNDICE B - Resultado da pesquisa com os alunos de Letras da UFU e da UFJ.

Pesquisa com os alunos do curso de Letras Inglês – UFU & UFJ

110 respostas

Eu aceito participar do projeto citado acima, voluntariamente, após ter sido devidamente esclarecido?

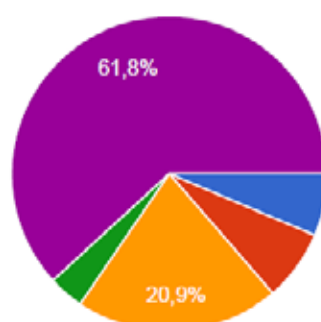
110 respostas



Questões 1 e 2

A. HAPAX LEGOMENA

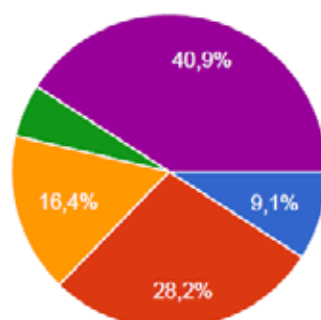
110 respostas



- Opção 1: Hapax legomena – Sabemos que algumas das...
- Opção 2: Hapax legomena – Somente uma variante foi...
- Opção 3: Hapax legomena – Dentre as palavras de uma...
- Opção 4: Hapax legomena – Neologismo genuíno.
- Opção 5: Hapax legomena – Do grego hápax, "uma únic...

B. TRAÇOS DISTINTIVOS

110 respostas

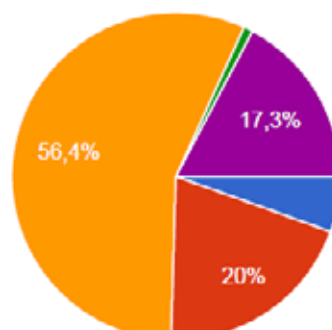


- Opção 1: Traços distintivos – Traços que podem ser tant...
- Opção 2: Traços distintivos – O que difere [p] de [b] são...
- Opção 3: Traços distintivos – Elemento composto de traç...
- Opção 4: Traços distintivos – Traços específicos.
- Opção 5: Traços distintivos – Traços peculiares de som...

Questões 3 e 4

C. LÉXICO

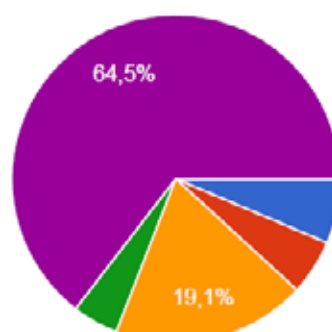
110 respostas



- Opção 1: Léxico – Organizado separadamente...
- Opção 2: Léxico - No nível do léxico há o uso de vocabul...
- Opção 3: Léxico – Conjunto de signos linguísticos que r...
- Opção 4: Léxico – Tesouro vocabular.
- Opção 5: Léxico – Constituído por unidades le...

D. TOPÔNIMO

110 respostas

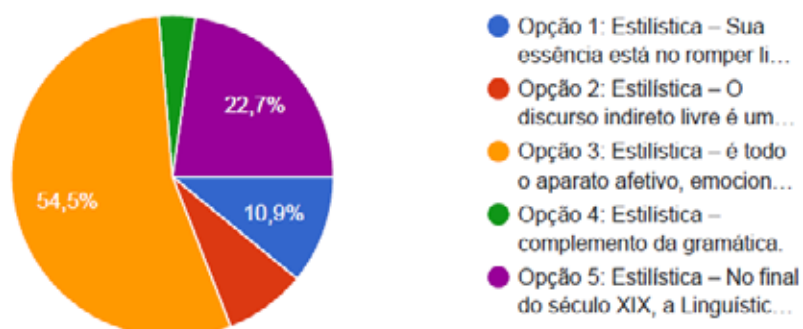


- Opção 1: Topônimo – Um pequeno texto, veículo de i...
- Opção 2: Topônimo – Dauzat (1972) fornece exemplos d...
- Opção 3: Topônimo – Resultado da ação do nom...
- Opção 4: Topônimo – Nome próprio.
- Opção 5: Topônimo – O termo topônimo é formado...

Questão 5

E. ESTILÍSTICA

110 respostas

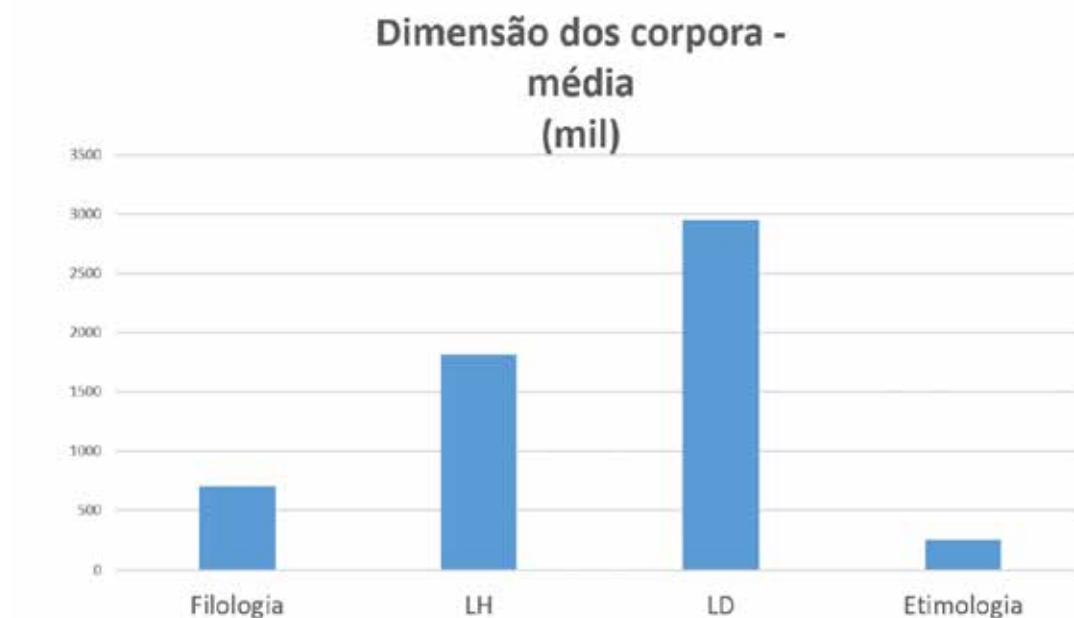


Este conteúdo não foi criado nem aprovado pelo Google. [Denunciar abuso](#) - [Termos de Serviço](#)

Google Formulários

APÊNDICE C - Dimensionamento de *corpora* de filologia, linguística histórica, linguística diacrônica e etimologia (2018).

Dados da pesquisa de textos acadêmicos nas áreas de Filologia, Linguística Histórica (LH), Linguística Diacrônica (LD) e Etimologia (2018).



Fonte: Elaborada pelo autor (2018).

APÊNDICE D - Dimensão e codificação inicial do *corpus* da subárea da Linguística Descritiva (LD) em português.

SUBÁREA	DIMENSÃO ⁸⁵	CODIFICAÇÃO
1. Análise da Conversação	502.929	L_AC_PT_G_TXTS#
2. Análise do Discurso	1.117.999	L_AD_PT_G_TXTS#
3. Aquisição da linguagem	538.012	L_AL_PT_G_TXTS#
4. Dialetoлогия	561.348	L_DL_PT_G_TXTS#
5. Ecolinguística	169.027	L_EL_PT_G_TXTS#
6. Estilística	536.277	L_EST_PT_G_TXTS#
7. Etimologia	Não há	L_ET_PT_G_TXTS#
8. Filologia	Não há	L_FL_PT_G_TXTS#
9. Fonética	673.197	L_FNT_PT_G_TXTS#
10. Fonologia	461.227	L_FNL_PT_G_TXTS#
11. Historiografia da Linguística	398.449	L_HL_PT_G_TXTS#
12. Lexicologia	504.171	L_LXL_PT_G_TXTS#
13. Linguística Cognitiva	579.759	L_LCG_PT_G_TXTS#
14. Linguística Histórica	935.113	L_LH_PT_G_TXTS#
15. Linguística Textual	1.642.673	L_LT_PT_G_TXTS#
16. Morfologia	764.058	L_M_PT_G_TXTS#
17. Neurolinguística	692.036	L_NL_PT_G_TXTS#
18. Onomástica	508.455	L_ON_PT_G_TXTS#
19. Pragmática	618.798	L_PRG_PT_G_TXTS#
20. Prosódia	516.062	L_PRS_PT_G_TXTS#
21. Psicolinguística	725.630	L_PL_PT_G_TXTS#
22. O tamanho desses corpos Semântica	492.198	L_SM_PT_G_TXTS#
23. Sintaxe	486.992	L_SN_PT_G_TXTS#
24. Sociolinguística	56.625	L_SL_PT_G_TXTS#
25. Teoria da Interpretação	615.404	L_TI_PT_G_TXTS#
26. Teoria da Tradução	542.209	L_TT_PT_G_TXTS#
27. Terminologia	192.084	L_TM_PT_G_TXTS#

Fonte: Elaborada pelo autor.

⁸⁵*Tokens* ou itens usados na lista de palavras; unidade: milhar. Esta forma de quantificação dos itens foi a mesma adotada para os dados dos Apêndices D ao K.

APÊNDICE E - Dimensão e codificação inicial do *corpus* da subárea da Linguística Aplicada (LA) em português.

SUBÁREA	DIMENSÃO	CODIFICAÇÃO
1. Bilinguismo e Multilinguismo	503.228	LA_BL&ML_PT_G_TXTS#
2. EaD Ensino de Línguas	623.264	LA_EaD_EL_PT_G_TXTS#
3. Ensino e Aprendizagem de Língua estrangeira	1.065.959	LA_E&A LE_PT_G_TXTS#
4. Ensino e Aprendizagem de Língua Instrumental	1.372.203	LA_E&A_LI_PT_G_TXTS#
5. Ensino e Aprendizagem de Língua materna	1.311.608	LA_E&A_LM_PT_ G_TXTS#
6. Formação de Professor de Línguas	468.933	LA_FPL_PT_G_TXTS#
7. Interpretação	Não há	LA_I_PT_G_TXTS#
8. Letramento	515.272	LA_L_PT_G_TXTS#
9. Lexicografia	1.032.826	LA_LXF_PT_G_TXTS#
10. Linguagem e Identidade	508.834	LA_L&I_PT_G_TXTS#
11. Linguagem e Tecnologia	1.091.729	LA_L&T_PT_G_TXTS#
12. Linguística de Contato	Não há	LA_LDC_PT_G_TXTS#
13. Linguística Forense	612.278	LA_LF_PT_G_TXTS#
14. Linguística Matemática	172.138	LA_LM_PT_G_TXTS#
15. Nivelamento e Avaliação	510.789	LA_N&A_PT_G_TXTS#
16. Políticas Linguísticas	95.695	LA_PL_PT_G_TXTS#
17. Terminografia	1.177.356	LA_TG_PT_G_TXTS#
18. Tradução	480.897	LA_TD_PT_G_TXTS#

Fonte: (Elaborada pelo autor).

APÊNDICE F - Dimensão e codificação inicial do *corpus* da subárea da LD em inglês.

	SUBÁREA	DIMENSÃO	CODIFICAÇÃO
1.	Análise da Conversação	461.072	L_AC_EN_G_TXTS#
2.	Análise do Discurso	1.709.141	L_AD_EN_G_TXTS#
3.	Aquisição da linguagem	829.948	L_AL_EN_G_TXTS#
4.	Dialetologia	639.285	L_DL_EN_G_TXTS#
5.	Ecolinguística	470.217	L_EL_EN_G_TXTS#
6.	Estilística	787.853	L_EST_EN_G_TXTS#
7.	Etimologia	Não há	L_ET_EN_G_TXTS#
8.	Filologia	Não há	L_FL_EN_G_TXTS#
9.	Fonética	497.535	L_FNT_EN_G_TXTS#
10.	Fonologia	487.806	L_FNL_EN_G_TXTS#
11.	Historiografia da Linguística	559.005	L_HL_EN_G_TXTS#
12.	Lexicologia	569.308	L_LXL_EN_G_TXTS#
13.	Linguística Cognitiva	651.179	L_LCG_EN_G_TXTS#
14.	Linguística Histórica	688.818	L_LH_EN_G_TXTS#
15.	Linguística Textual	536.594	L_LT_EN_G_TXTS#
16.	Morfologia	1.135.907	L_M_PORT_G_TXTS#
17.	Neurolinguística	784.446	L_NL_EN_G_TXTS#
18.	Onomástica	674.629	L_ON_EN_G_TXTS#
19.	Pragmática	1.328.929	L_PRG_EN_G_TXTS#
20.	Prosódia	522.147	L_PRS_EN_G_TXTS#
21.	Psicolinguística	1.033.362	L_PL_EN_G_TXTS#
22.	Semântica	534.194	L_SM_EN_G_TXTS#
23.	Sintaxe	1.020.136	L_SN_EN_G_TXTS#
24.	Sociolinguística	1.085.311	L_SL_EN_G_TXTS#
25.	Teoria da Interpretação	764.952	L_TI_EN_G_TXTS#
26.	Teoria da Tradução	742.539	L_TT_EN_G_TXTS#
27.	Terminologia	651.331	L_TM_EN_G_TXTS#

APÊNDICE G - Dimensão e codificação inicial do *corpus* da subárea da LA em inglês.

SUBÁREA		DIMENSÃO	CODIFICAÇÃO
1.	Análise da Conversação	461.072	L_AC_EN_G_TXTS#
2.	Análise do Discurso	1.709.141	L_AD_EN_G_TXTS#
3.	Aquisição da linguagem	829.948	L_AL_EN_G_TXTS#
4.	Dialetologia	639.285	L_DL_EN_G_TXTS#
5.	Ecolinguística	470.217	L_EL_EN_G_TXTS#
6.	Estilística	787.853	L_EST_EN_G_TXTS#
7.	Etimologia	Não há	L_ET_EN_G_TXTS#
8.	Filologia	Não há	L_FL_EN_G_TXTS#
9.	Fonética	497.535	L_FNT_EN_G_TXTS#
10.	Fonologia	487.806	L_FNL_EN_G_TXTS#
11.	Historiografia da Linguística	559.005	L_HL_EN_G_TXTS#
12.	Lexicologia	569.308	L_LXL_EN_G_TXTS#
13.	Linguística Cognitiva	651.179	L_LCG_EN_G_TXTS#
14.	Linguística Histórica	688.818	L_LH_EN_G_TXTS#
15.	Linguística Textual	536.594	L_LT_EN_G_TXTS#
16.	Morfologia	1.135.907	L_M_PORT_G_TXTS#
17.	Neurolinguística	784.446	L_NL_EN_G_TXTS#
18.	Onomástica	674.629	L_ON_EN_G_TXTS#
19.	Pragmática	1.328.929	L_PRG_EN_G_TXTS#
20.	Prosódia	522.147	L_PRS_EN_G_TXTS#
21.	Psicolinguística	1.033.362	L_PL_EN_G_TXTS#
22.	Semântica	534.194	L_SM_EN_G_TXTS#
23.	Sintaxe	1.020.136	L_SN_EN_G_TXTS#
24.	Sociolinguística	1.085.311	L_SL_EN_G_TXTS#
25.	Teoria da Interpretação	764.952	L_TI_EN_G_TXTS#
26.	Teoria da Tradução	742.539	L_TT_EN_G_TXTS#
27.	Terminologia	651.331	L_TM_EN_G_TXTS#

APÊNDICE H - Redimensionamento e codificação do *corpus* final da subárea da LD em português.

SUBÁREA	DIMENSÃO	CODIFICAÇÃO
1. Análise da Conversação	512.430	L_AC_PT_G_TXTS#
2. Análise do Discurso	503.114	L_AD_PT_G_TXTS#
3. Aquisição da linguagem	535.134	L_AL_PT_G_TXTS#
4. Biolinguística	532.561	L_BL_PT_G_TXTS#
5. Dialetologia	500.513	L_DL_PT_G_TXTS#
6. Ecolinguística	518.049	L_EL_PT_G_TXTS#
7. Estilística	505.768	L_EST_PT_G_TXTS#
8. Etimologia	544.563	L_ET_PT_G_TXTS#
9. Filologia	543.634	L_FL_PT_G_TXTS#
10. Fonética	510.445	L_FNT_PT_G_TXTS#
11. Fonologia	506.883	L_FNL_PT_G_TXTS#
12. Historiografia da Linguística	510.261	L_HL_PT_G_TXTS#
13. Lexicologia	527.645	L_LXL_PT_G_TXTS#
14. Linguística Cognitiva	517.088	L_LCG_PT_G_TXTS#
15. Linguística Diacrônica	503.293	L_LD_PT_G_TXTS#
16. Linguística Histórica	540.720	L_LH_PT_G_TXTS#
17. Linguística Textual	513.410	L_LT_PT_G_TXTS#
18. Morfologia	521.757	L_M_PT_TXTS#
19. Neurolinguística	503.131	L_NL_PT_G_TXTS#
20. Onomástica	513.016	L_ON_PT_G_TXTS#
21. Pragmática	515.898	L_PRG_PT_G_TXTS#
22. Prosódia	516.062	L_PRS_PT_G_TXTS#
23. Psicolinguística	504.190	L_PL_PT_G_TXTS#
24. Semântica	515.631	L_SM_PT_G_TXTS#
25. Sintaxe	531.013	L_SN_PT_G_TXTS#
26. Sociolinguística	501.904	L_SL_PT_G_TXTS#
27. Teoria da Interpretação	532.757	L_TI_PT_G_TXTS#
28. Teoria da Tradução	526.550	L_TT_PT_G_TXTS#
29. Terminologia	501.380	L_TM_PT_G_TXTS#

APÊNDICE I - Redimensionamento e codificação do corpus final da subárea da LA em português.

SUBÁREA	DIMENSÃO	CODIFICAÇÃO
1. Bilinguismo e Multilinguismo	503.228	LA_BL&ML_PT_G_TXTS#
2. EaD Ensino de Línguas	514.776	LA_EaD_EL_PT_G_TXTS#
3. Ensino e Aprendizagem de Língua estrangeira	504.823	LA_E&A_LE_PT_G_TXTS#
4. Ensino e Aprendizagem de Língua Instrumental	515.546	LA_E&A_LI_PT_G_TXTS#
5. Ensino e Aprendizagem de Língua materna	530.522	LA_E&A_LM_PT_G_TXTS#
6. Formação de Professor de Línguas	515.982	LA_FPL_PT_G_TXTS#
7. Interpretação	507.675	LA_I_PT_G_TXTS#
8. Letramento	521.843	LA_L_PT_G_TXTS#
9. Lexicografia	504.245	LA_LXF_PT_G_TXTS#
10. Linguagem e Identidade	506.556	LA_L&I_PT_G_TXTS#
11. Linguagem e Tecnologia	502.050	LA_L&T_PT_G_TXTS#
12. Linguística de Contato	506.660	LA_LDC_PT_G_TXTS#
13. Linguística Forense	517.832	LA_LF_PT_G_TXTS#
14. Linguística Matemática	220.245	LA_LM_PT_G_TXTS#
15. Nivelamento e Avaliação	512.635	LA_N&A_PT_G_TXTS#
16. Políticas Linguísticas	533.040	LA_PL_PT_G_TXTS#
17. Tradução	508.665	LA_TR_PT_G_TXTS#
18. Terminografia	535.301	LA_TG_PT_G_TXTS#

APÊNDICE J - Redimensionamento e codificação do *corpus* final da subárea da LD em inglês.

SUBÁREA	DIMENSÃO	CODIFICAÇÃO
1. Análise da Conversação	523.611	L_AC_EN_G_TXTS#
2. Análise do Discurso	516.697	L_AD_EN_G_TXTS#
3. Aquisição da linguagem	532.407	L_AL_EN_G_TXTS#
4. Biolinguística	534.591	L_BL_EN_G_TXTS#
5. Dialetologia	501.778	L_DL_EN_G_TXTS#
6. Ecolinguística	521.776	L_EL_EN_G_TXTS#
7. Estilística	537.566	L_EST_EN_G_TXTS#
8. Etimologia	504.332	L_ET_EN_G_TXTS#
9. Filologia	501.019	L_FL_EN_G_TXTS#
10. Fonética	547.482	L_FNT_EN_G_TXTS#
11. Fonologia	555.185	L_FNL_EN_G_TXTS#
12. Historiografia da Linguística	502.884	L_HL_EN_G_TXTS#
13. Lexicologia	522.024	L_LXL_EN_G_TXTS#
14. Linguística Cognitiva	526.429	L_LCG_EN_G_TXTS#
15. Linguística Diacrônica	508.094	L_LD_EN_G_TXTS#
16. Linguística Histórica	518.534	L_LH_EN_G_TXTS#
17. Linguística Textual	518.955	L_LT_EN_G_TXTS#
18. Morfologia	543.852	L_M_EN_G_TXTS#
19. Neurolinguística	534.849	L_NL_EN_G_TXTS#
20. Onomástica	528.356	L_ON_EN_G_TXTS#
21. Pragmática	506.631	L_PRG_EN_G_TXTS#
22. Prosódia	523.302	L_PRS_EN_G_TXTS#
23. Psicolinguística	534.269	L_PL_EN_G_TXTS#
24. Semântica	509.149	L_SM_EN_G_TXTS#
25. Sintaxe	560.800	L_SN_EN_G_TXTS#
26. Sociolinguística	551.769	L_SL_EN_G_TXTS#
27. Teoria da Interpretação	526.904	L_TI_EN_G_TXTS#
28. Teoria da Tradução	535.211	L_TT_EN_G_TXTS#
29. Terminologia	519.006	L_TM_EN_G_TXTS#

APÊNDICE K - Redimensionamento e codificação do *corpus* final da subárea da LA em inglês.

SUBÁREA	DIMENSÃO	CODIFICAÇÃO
1. Bilinguismo e Multilinguismo	504.866	LA_BL&ML_EN_G_TXTS#
2. EaD Ensino de Línguas	517.073	LA_EaDEL_EN_G_TXTS#
3. Ensino e Aprendizagem de Língua estrangeira	524.186	LA_E&A LE_EN_G_TXTS#
4. Ensino e Aprendizagem de Língua Instrumental	515.013	LA_E&A LI_EN_G_TXTS#
5. Ensino e Aprendizagem de Língua materna	503.562	LA_E&A LM_EN_G_TXTS#
6. Formação de Professor de Línguas	539.812	LA_FPL_EN_G_TXTS#
7. Interpretação	532.691	LA_I_EN_G_TXTS#
8. Letramento	506.061	LA_L_EN_G_TXTS#
9. Lexicografia	510.271	LA_LXF_EN_G_TXTS#
10. Linguagem e Identidade	509.757	LA_L&I_EN_G_TXTS#
11. Linguagem e Tecnologia	522.849	LA_L&T_EN_G_TXTS#
12. Linguística de Contato	551.623	LA_LdC_EN_G_TXTS#
13. Linguística Forense	522.918	LA_LF_EN_G_TXTS#
14. Linguística Matemática	223.056	LA_LM_EN_G_TXTS#
15. Nivelamento e Avaliação	532.862	LA_N&A_EN_G_TXTS#
16. Políticas Linguísticas	520.784	LA_PL_EN_G_TXTS#
17. Terminografia	524.301	LA_TG_EN_G_TXTS#
18. Tradução	514.857	LA_TD_EN_G_TXTS#

APÊNDICE L - VoBLing - lista de verbetes definidos

Nº	Área/Subárea	português	inglês
1.	Linguística	linguística	<i>linguistics</i>
2.		língua	<i>language</i>
1.	Linguística Descritiva (LD)	linguística descritiva	<i>descriptive linguistics</i>
2.		sílaba	<i>syllable</i>
1.	Linguística Aplicada (LA)	linguística aplicada	<i>applied linguistics</i>
2.		ensino	<i>teaching</i>
LINGUÍSTICA DESCRITIVA			
1.	Análise da Conversação	análise da conversação	<i>conversation analysis</i>
2.		conversação	<i>conversation</i>
3.	Análise do Discurso	análise do discurso	<i>discourse analysis</i>
4.		discurso	<i>discourse</i>
5.	Aquisição da Linguagem	aquisição da linguagem	<i>first language acquisition</i>
6.		criança	<i>child</i>
7.	Biolinguística	biolinguística	<i>biolinguistics</i>
8.		merge	<i>merge</i>
9.	Dialetologia	dialetologia	<i>dialectology</i>
10.		atlas linguístico	<i>linguistic atlas</i>
11.	Ecolinguística	ecolinguística	<i>ecolinguistics</i>
12.		ecossistema	<i>ecosystem</i>
13.	Estilística	estilística	<i>stylistics</i>
14.		estilo	<i>style</i>
15.	Etimologia	etimologia	<i>etymology</i>
16.		atestações	<i>attestations</i>
17.	Filologia	filologia	<i>philology</i>
18.		textos	<i>texts</i>
19.	Fonética	fonética	<i>phonetics</i>
20.		sons	<i>sounds</i>
21.	Fonologia	fonologia	<i>phonology</i>
22.		tepe	<i>tap</i>
23.	Historiografia da Linguística	historiografia da linguística	<i>historiography of linguistics</i>
24.		historiógrafo da linguística	<i>historiographers of linguistics</i>
25.	Lexicologia	lexicologia	<i>lexicology</i>
26.		dicionários	<i>dictionaries</i>
27.	Linguística Cognitiva	linguística cognitiva	<i>cognitive linguistics</i>
28.		metáfora	<i>metaphor</i>
29.	Linguística Diacrônica	linguística diacrônica	<i>diachronic linguistics</i>

30.		construções	<i>constructions</i>
31.	Linguística Histórica	linguística histórica	<i>historical linguistics</i>
32.		derivação	<i>derivation</i>
33.	Linguística Textual	linguística textual	<i>text linguistics</i>
34.		gênero	<i>genre</i>
35.	Morfologia	morfologia	<i>morphology</i>
36.		sufixos	<i>suffixes</i>
37.	Neurolinguística	neurolinguística	<i>neurolinguistics</i>
38.		afasia	<i>aphasia</i>
39.	Onomástica	onomástica	<i>onomastics</i>
40.		sobrenomes	<i>surnames</i>
41.	Pragmática	pragmática	<i>pragmatics</i>
42.		negação	<i>negation</i>
43.	Prosódia	prosódia	<i>prosody</i>
44.		acentos	<i>stress</i>
45.	Psicolinguística	psicolinguística	<i>psycholinguistics</i>
46.		processamento linguístico	<i>language processing</i>
47.	Semântica	semântica	<i>semantics</i>
48.		significado	<i>meaning</i>
49.	Sintaxe	sintaxe	<i>syntax</i>
50.		orações	<i>sentences</i>
51.	Sociolinguística	sociolinguística	<i>sociolinguistics</i>
52.		variação	<i>variation</i>
53.	Teoria da Interpretação	teoria da interpretação	<i>interpretation theory</i>
54.		interpretação	<i>interpretation</i>
55.	Teoria da Tradução	teoria da tradução	<i>translation theory</i>
56.		tradutor	<i>translator</i>
57.	Terminologia	terminologia	<i>terminology</i>
58.		termo	<i>term</i>
LINGUÍSTICA APLICADA (LA)			
1.	Bilinguismo e Multilinguismo	bilinguismo e multilinguismo	<i>bilingualism & multilingualism</i>
2.		surdos	<i>deaf children</i>
3.	Ead Ensino de Línguas	ead ensino de línguas	<i>distance language learning</i>
4.		teletandem	<i>tandem language learning</i>
5.	Ensino e Aprendizagem de Língua Estrangeira	ensino e aprendizagem de língua estrangeira	<i>foreign language teaching and learning</i>
6.		moodle	<i>moodle</i>
7.	Ensino e Aprendizagem de Língua Instrumental	ensino e aprendizagem de língua instrumental	<i>language for specific purposes</i>
8.		abordagem	<i>approach</i>
9.	Ensino e Aprendizagem de Língua Materna	ensino e aprendizagem de língua materna	<i>mother tongue teaching and learning</i>

10.		ensino-aprendizagem	<i>teaching and learning</i>
11.	Formação de Professor de Línguas	formação de professor de línguas	<i>teaching and teacher education</i>
12.		formação contínua	<i>continuous training</i>
13.	Interpretação	interpretação	<i>interpretation</i>
14.		intérprete	<i>interpreter</i>
15.	Letramento	letramento	<i>literacy</i>
16.		gráficos	<i>graphics</i>
17.	Lexicografia	lexicografia	<i>lexicography</i>
18.		frames	<i>frames</i>
19.	Linguagem e Identidade	linguagem e identidade	<i>language and identity</i>
20.		identidade	<i>identity</i>
21.	Linguagem e Tecnologia	linguagem e tecnologia	<i>language and technology</i>
22.		tecnologias	<i>technologies</i>
23.	Linguística de Contato	linguística de contato	<i>contact linguistics</i>
24.		contato linguístico	<i>language contact</i>
25.	Linguística Forense	linguística forense	<i>forensic linguistics</i>
26.		corpus	<i>corpus</i>
27.	Linguística Matemática	linguística matemática	<i>mathematical linguistics</i>
28.		anotação	<i>annotation</i>
29.	Nivelamento e Avaliação	nivelamento e avaliação	<i>placement & evaluation</i>
30.		avaliação	<i>evaluation</i>
31.	Políticas Linguísticas	políticas linguísticas	<i>language policy</i>
32.		práticas linguísticas	<i>language practices</i>
33.	Tradução	tradução	<i>translation</i>
34.		equivalência	<i>equivalence</i>
35.	Terminografia	terminografia	<i>terminography</i>
36.		linguagem de especialidade	<i>specialized languages</i>

APÊNDICE M - Roteiro para avaliação de dicionários ou glossários de Linguística impressos ou eletrônicos segundo Faulstich (2011).

Título:

Autor:

Editora:

Edição:

Data:

Local de publicação:

Volume(s):

1. Sobre o autor

1.1. Trata-se de pessoa reconhecida na área de dicionarística ou de terminologia?

1.2. Qual a formação acadêmica do autor principal e dos participantes do grupo de pesquisa?

2. Sobre a apresentação da obra pelo autor

2.1. Há introdução na qual apareçam claramente:

a) os objetivos da obra?

b) o público para o qual o conteúdo se dirige?

c) as informações sobre como consultar o dicionário ou vocabulário?

d) referências à bibliografia de onde foi extraído o *corpus*?

2.2. Há bibliografia de consulta justificada pelo autor?

3. Sobre a apresentação do material da obra

3.1. Há prefácio redigido por personalidade reconhecida na área de dicionarística? Científica, técnica?

3.2. A família tipográfica empregada é adequada à faixa etária do usuário?

3.3. As ilustrações, se houver, estão adequadas à microestrutura informacional?

3.4. A utilização de negrito, de itálico e de outros recursos gráficos está de acordo com o equilíbrio visual da obra?

3.5. Os verbetes são apresentados em ordem alfabética ou em ordem sistemática?

3.6. A obra contempla uma só língua ou mais de uma?

3.7. O formato do dicionário ou vocabulário permite manuseio prático e fácil?

3.8. A obra está editada em suporte informatizado?

3.9. A qualidade do acabamento garante a sua durabilidade?

3.10. O sistema de abreviações e de símbolos aparece corretamente no corpo do texto?

3.11. A obra possui ampla divulgação?

4. Sobre o conteúdo

4.1. Os verbetes apresentam:

a) categoria gramatical?

b) gênero?

c) sinonímia?

d) variante(s) da entrada?

e) indicação de área ou subárea de especialidade?

f) contexto? (exemplo ou abonação?)

g) equivalente(s)?

h) indicação de pronúncia?

i) remissivas úteis entre conceitos?

j) fontes?

k) notas?

4.4. A definição é constituída por um enunciado de uma só frase?

4.5. A definição leva em conta o nível de discurso do usuário?

APÊNDICE N - Resultado da análise dos dicionários ou glossários **de Linguística** impressos e eletrônicos segundo Faulstich (2011).

1. Título: Dicionário enciclopédico das ciências da linguagem

Autor: Oswald Ducrot; Tzvetan Todorov

Editora: Perspectiva

Edição: 1ª

Data: 1977

Local de publicação: SP

Volume(s): 1

1. Sobre o autor

1.1. Sim.

1.2. Doutorado em Letras (1970).

2. Sobre a apresentação da obra pelo autor

2.1.

a) Sim, servir para leitura como dicionário ou enciclopédia.

b) Sim, de especialistas a neófitos.

c) Sim, estrutura em cortes conceituais, com verbetes por temas e remissivas numeradas por número das páginas.

d) Sim, permeia toda a obra no interior e no fim de cada tema.

2.2. Sim, de forma sucinta e objetiva, como por exemplo: “alguns textos essenciais dessa divertida história pré-linguística das línguas, grandes tratados de gramática comparada, exemplo de estudos etnolinguísticos, para uma discussão pormenorizada da transformação de nominalização, além das propriedades sintáticas e semânticas do nome”.

3. Sobre a apresentação material da obra

3.1. Não há.

3.2. Sim.

3.3. Não há.

3.4. Sim, uso de letras maiúsculas para indicação de termos linguísticos e uso de itálico para exemplificações e nome de obra seguindo a ABNT.

3.5. Sim, em ordem temática.

3.6. Só uma, o português, traduzido do francês.

3.7. Sim.

3.8. Não.

3.9. Não, obra impressa em capa comum.

3.10. Sim, abreviaturas utilizadas somente para detalhes de bibliografia e paginação.

3.11. Sim.

4. Sobre o conteúdo

4.1.

a) Não.

b) Não.

c) Não.

d) Não.

e) Sim.

f) Sim, por exemplos criados pelos autores.

g) Sim, às vezes, como em *clusters* para aglomerados semânticos (ver p. 258).

h) Não.

i) Sim, com indicação da remissiva entre parênteses e do número da página entre colchetes.

- j) Sim, com indicação de bibliografia e do autor (índice de autores ao final da obra) e suas teorias.
- k) Não há.
- 4.4. Sim, seguida por parágrafos explicativos em formato enciclopédico.
- 4.5. Sim.

2. Título: **Dicionário de Linguística e Gramática: referente à língua portuguesa**

Autor: Joaquim Mattoso Câmara Jr.

Editora: Vozes

Edição: 16^a

Data: 1992

Local de publicação: Petrópolis

Volume(s): 1

1. Sobre o autor

1.1. Não.

1.2. Filologia Latina e Neolatina (antiga Universidade do Distrito Federal, 1949).

2. Sobre a apresentação da obra pelo autor

2.1. Há introdução na qual apareçam claramente:

a) Sim.

b) Parcialmente. Não há a especificação da idade, nem do nível acadêmico do consulente.

c) Não. Menciona que a estrutura de verbetes está em ordem alfabética e as remissivas, mas não informa como utilizá-las.

d) Sim.

2.2. A obra disponibiliza referências bibliográficas, siglas (obras coletivas), textos que são referenciados na microestrutura da obra.

3. Sobre a apresentação material da obra

3.1. Não, a obra é apresentada pelos editores e pelo autor. Hamilton Elia faz um histórico da disciplina no Brasil e descreve o perfil acadêmico do autor.

3.2. Sim.

3.3. Não há.

3.4. Sim, as letras maiúsculas indicam as remissivas circulares, sem definição (ex.: simples vs. complexo); o (v.) indica a remissiva de um termo com definição.

3.5. Em ordem alfabética.

3.6. Uma língua: o português. O autor lança mão do latim e do grego para explicações etimológicas.

3.7. Sim.

3.8. Não.

3.9. Não, obra impressa em capa comum.

3.10. Sim.

3.11. Sim, pois esta foi uma das obras basilares da linguística do Brasil.

4. Sobre o conteúdo

4.1.

a. Não.

b. Não.

c. Há uma rede de remissivas que busca interligar termos semelhantes ou de significação semelhante, como forma, significado e significação.

d. Não.

e. Não há.

f. Alguns exemplos providenciados pelo autor, como em complemento.

g. Não há.

- h. Algumas, com “sinais de transcrição fonética usados no livro” na *frontmatter*.
- i. Parcialmente, outras circulares como exposto no item 3.4.
- j. Referências bibliográficas disponibilizadas ao final da obra.
- k. Não há.
- 4.4. Sim.
- 4.5. Não.

3. Título: **Dicionário de Termos Literários**

Autor: Massaud Moisés⁸⁶

Editora: Cultrix

Edição: 14^a

Data: 1995

Local de publicação: São Paulo

Volume(s): 1

1. Sobre o autor

- 1.1. Não.
- 1.2. Doutorado em Literatura, membro da Academia Paulista de Letras, ocupante da 17^a cadeira.

2. Sobre a apresentação da obra pelo autor

- 2.1.
 - a) Não.
 - b) Não.
 - c) Não.
 - d) Sim.
- 2.2. Sim, na microestrutura do dicionário.

3. Sobre a apresentação material da obra

- 3.1. Não, a obra é apresentada pelo autor.
- 3.2. Sim.
- 3.3. Não há.
- 3.4. Sim, os termos em letras maiúsculas indicam as remissivas.
- 3.5. Em ordem alfabética.
- 3.6. Uma língua: o português. O autor lança mão do latim, grego, francês, alemão, japonês e outras línguas estrangeiras para explicações etimológicas.
- 3.7. Sim.
- 3.8. Não.
- 3.9. Não, obra impressa em capa comum.
- 3.10. Sim.
- 3.11. Sim, pois esta é uma das obras basilares da Literatura no Brasil.

4. Sobre o conteúdo

- 4.1.
 - a) Não.
 - b) Não.
 - c) O autor menciona a sinonímia na microestrutura como em *dicção*, sinônimo de *arte ou maneira de dizer*, e *elocução* e *lexis* entre gregos e romanos.
 - d) Não.
 - e) Não há.
 - f) Alguns exemplos providenciados pelo autor, como em *complemento*.
 - g) Não há.

⁸⁶ Breve biografia do autor: <https://www.fflch.usp.br/518>. Acesso em 30 jul. 2020.

- h) Algumas, com “sinais de transcrição fonética usados no livro” na *frontmatter*.
- i) Parcialmente, outras circulares como exposto no item 3.4.
- j) Referências bibliográficas disponibilizadas ao final da obra.
- k) Não há.

4.4. Sim.

4.5. Não.

4. Título: **Dicionário de Linguística e Fonética**

Autor: David Crystal

Editora: Jorge Zahar

Edição: 2^a

Data: 2000

Local de publicação: Rio de Janeiro

Volume(s): 1

1. Sobre o autor

1.1. Sim⁸⁷.

1.2. Doutorado em Linguística (University College de Londres).

2. Sobre a apresentação da obra pelo autor

2.1.

a) Sim.

b) Sim, profissionais da área de ensino e da terapia da língua/fala e aqueles do meio acadêmico em que o estudo da língua é seu objeto principal: psicologia, sociologia, crítica literária, filosofia e os alunos de linguística

c) Sim.

d) Sim.

2.2. Sim, na microestrutura do dicionário, ao final de cada verbete.

3. Sobre a apresentação material da obra

3.1. Não, a obra é apresentada pelo autor e pela tradutora.

3.2. Sim.

3.3. Não há.

3.4. Sim, os termos em versalete indicam as remissivas.

3.5. Em ordem alfabética.

3.6. Uma língua: o português, traduzida do inglês.

3.7. Sim.

3.8. Não.

3.9. Não, obra impressa em capa comum.

3.10. Não há.

3.11. Sim.

4. Sobre o conteúdo

4.1.

a) Não.

b) Não.

c) O autor menciona a sinonímia na microestrutura como em *dicção*, sinônimo de *arte ou maneira de dizer*, e *elocução* e *lexis* entre gregos e romanos.

d) Não.

e) Não há.

f) Alguns exemplos providenciados pelo autor, como em *complemento*.

⁸⁷

Produção

bibliográfica

disponível

em

https://web.archive.org/web/20071026144547/http://www.crystalreference.com/David_Crystal/books.htm.

Acesso em 22 jul. 2020.

- g) Não há.
- h) Algumas, com “sinais de transcrição fonética usados no livro” na *frontmatter*.
- i) Parcialmente, outras circulares como exposto no item 3.4.
- j) Referências bibliográficas disponibilizadas ao final da obra.
- k) Não há.

4.4. Sim.

4.5. Sim.

5. Título: **Dicionário de Linguística**

Autor: Jean Dubois⁸⁸ et al.

Editora: Cultrix

Edição: 14^a

Data: 2004

Local de publicação: Rio de Janeiro

Volume(s): 1

1. Sobre o autor

1.1. Sim.

1.2. Linguista, gramático e lexicógrafo.

2. Sobre a apresentação da obra pelo autor

2.1.

a) Sim.

b) Sim, alunos e professores, além daqueles que se identificam com as ciências humanas.

c) Não.

d) Sim, p. 621-653 (32 p).

2.2. Sim, na microestrutura do dicionário, os autores explicam os ST associando-os aos seus autores. Ex.: *translação* - segundo Charles Bally e L. Tesnière, com notas em *translativo*.

3. Sobre a apresentação material da obra

3.1. Não, a obra é apresentada pelos autores: Jean Dubois, Mathêe Giacomo, Louis Guespin, Christiane Marcellesi, Jean-Baptiste Marcellesi e Jean-Pierre Mevel.

3.2. Sim.

3.3. Não há.

3.4. Sim, os verbetes, sinônimos e exemplos são impressos em itálico; as remissivas e os nomes de autores são escritas em letras maiúsculas.

3.5. Em ordem alfabética.

3.6. Uma língua: o português, traduzida do francês.

3.7. Sim.

3.8. Não.

3.9. Não, obra impressa em capa comum.

3.10. O uso de V. para ver, indicando as remissivas.

3.11. Sim.

4. Sobre o conteúdo

4.1.

a) Não.

b) Não.

c) Os autores marcam a sinonímia com verbetes em itálico como em *pleonismo* e *transformação pleonástica*.

d) Não.

e) Não há, senão a indicação de autores e suas correntes linguísticas. Ex.: *performativo*.

⁸⁸ Perfil do autor disponível na Biblioteca Nacional da França: <https://catalogue.bnf.fr/ark:/12148/cb120029592>. Acesso em: 22 jul. 2020.

- f) Sim, apresentados em itálico.
 - g) Não há.
 - h) Não há.
 - i) Sim, marcadas pela letra V. e verbete escrito em letras maiúsculas.
 - j) Referências bibliográficas disponibilizadas ao final da obra.
 - k) Não há.
- 4.4. Sim, com a repetição dos verbetes no início do PD.
- 4.5. Sim, considera o perfil de universitários, mencionando frequentemente autores referência da Linguística.

6. Título: Dicionário de semiótica

Autor: Algirdas Julien Greimas e Joseph Courtés⁸⁹

Editora: Cultrix

Edição: 1ª

Data: 2008

Local de publicação: Rio de Janeiro

Volume(s): 1

1. Sobre o autor

1.1. Sim.

1.2. Greimas: Doutorado em Filologia Francesa (1948) e Courtés: Doutorado em Letras.

2. Sobre a apresentação da obra pelo autor

2.1.

a) Sim: homogeneizar os conceitos da teoria semiótica.

b) Não, mas preveem que alguns leitores vão considerar a “terminologia exageradamente sofisticada e até mesmo repulsiva” (p. 2).

c) Sim.

d) De forma vaga, no prólogo: "dicionários são feitos a partir de outros dicionários: tal foi, também, o que fizemos" (p. 7).

2.2. Não há.

3. Sobre a apresentação material da obra

3.1. Não, a obra é apresentada supostamente pelos autores: Greimas e Courtés.

3.2. Sim.

3.3. Não há.

3.4. Só há o uso de asterisco para marcação de remissivas de uma seta, ao final dos verbetes que indicam os antônimos ou hipônimos.

3.5. Em ordem alfabética.

3.6. Verbetes em 3 línguas: o português, o francês ou o inglês. A obra é originalmente traduzida do francês.

3.7. Sim.

3.8. Não.

3.9. Não, obra impressa em capa comum.

3.10. Sim, o asterisco.

3.11. Sim.

4. Sobre o conteúdo

4.1.

a) Sim.

b) Sim.

⁸⁹ Perfil do autor disponível na Biblioteca Nacional da França: Greimas Courtés
<https://catalogue.bnf.fr/ark:/12148/cb11905841v>
<https://catalogue.bnf.fr/ark:/12148/cb11897976z>.

- c) Não há.
- d) Não.
- e) Sim, indicam a qual escola o conceito do verbete pertence. Ex.: história, índice (Pierce, Prieto, Saussure, Barthes).
- f) Não há.
- g) Sim, em francês e em inglês.
- h) Não há.
- i) Sim, marcadas por asterisco.
- j) Não há.
- k) Não há.

4.4. Parcialmente, os autores apresentam um conceito de forma explicativa em um parágrafo, normalmente indicando o autor e a escola que o adota.

4.5. Não, pois o perfil do usuário não é identificado na obra.

7. Título: Dicionário de linguagem e linguística

Autor: Robert Lawrence Trask

Editora: Contexto

Edição: 3ª, 1ª reimpressão

Data: 2015

Local de publicação: São Paulo

Volume(s): 1

1. Sobre o autor

1.1. Sim.

1.2. Doutorado em Linguística.

2. Sobre a apresentação da obra pelo autor

2.1.

a) Não.

b) Iniciantes no estudo linguístico curso de graduação ou de pós-graduação.

c) Sim.

d) Não há.

2.2. Sim, ao final dos verbetes.

3. Sobre a apresentação material da obra

3.1. Sim, a obra é apresentada por Ilari e pelo autor.

3.2. Sim.

3.3. Não há.

3.4. Sim, o uso é racional, o negrito marca os verbetes e as remissivas e o itálico marca as traduções e as *leituras suplementares*. Há o uso do itálico para marcar termos da linguística, no interior da microestrutura do PD, apesar de os mesmos não serem definidos na obra. Ex.: *tradução automática* dentro da microestrutura do verbete **linguística computacional**.

3.5. Em ordem alfabética.

3.6. Verbetes em 2 línguas: o português e o inglês. A obra é originalmente traduzida do inglês e os PD escritos em LP.

3.7. Sim.

3.8. Não.

3.9. Não, obra impressa em capa comum.

3.10. Sim, o uso de abreviações depois da definição como em *Systemic Linguistics* (SL) no PD Linguística sistêmica.

3.11. Sim.

4. Sobre o conteúdo

4.1.

- a) Não.
 - b) Não.
 - c) Não há.
 - d) Não.
 - e) Às vezes, como em **competência**, contudo não é este o foco da obra. No prefácio, o autor informa que buscou catalogar os conceitos mais importantes da Linguística em geral “os conceitos incluídos estão entre os mais importantes desse campo e entre aqueles que todo iniciante nos estudos linguísticos está sujeito a encontrar.” (p. 13).
 - f) Sim. Ex.: no PD de **conjunção**.
 - g) Sim, em inglês.
 - h) Não há.
 - i) Sim, marcadas por *Ver* e fonte em negrito.
 - j) Sim, ao final da obra, na bibliografia original e no suplemento bibliográfico.
 - k) Não há.
- 4.4. Parcialmente, os autores apresentam um PD terminológico e um texto explicativo com detalhes do conceito, podendo indicar o teórico e a subárea da Linguística que o adota.
- 4.5. Sim, aos alunos de graduação e pós-graduação.

8. Título: Dicionário de Linguística da Enunciação

Autor: Flores et al. (Leci Borges Barbisan, Marlene Teixeira, Maria José Bocorny Finatto, Valdir do Nascimento Flores).

Editora: Contexto

Edição: 1^a

Data: 2009

Local de publicação: São Paulo

Volume(s): 1

1. Sobre o autor

1.1. Sim.

1.2. (i) Leci Borges Barbisan: doutora em linguística e fonética pela Université de Grenoble III (1983), na França; (ii) Marlene Teixeira, doutora em Letras pela Pontifícia Universidade Católica do Rio Grande do Sul (1998); (iii) Maria José Bocorny Finatto (Pós-Doutorada em Ciência da Computação junto o Núcleo Interinstitucional de Linguística Computacional – NILC - do ICMC-USP em 2011); (iv) Valdir do Nascimento Flores (Pós-doutorado, na Université de Paris XII-Val-de-Marne; Pós-doutorado, sob a direção de Claudine Normand, na Université de Paris X⁹⁰

2. Sobre a apresentação da obra pelo autor

2.1.

a) Sim.

b) Alunos dos anos finais de Letras e pós-graduandos.

c) Sim.

d) Sim, na seção Códigos das equipes, Fontes bibliográficas em bibliografia recomendada (p. 257-267).

2.2. Sim, ao final dos verbetes. Idem item 2.1 d.

3. Sobre a apresentação material da obra

3.1 Sim, a obra é apresentada por Fiorin.

3.2. Sim.

3.3. Há um guia rápido dos verbetes para o usuário.

90 O currículo dos autores encontra-se disponível na plataforma Lattes, do CNPq, disponível em <http://lattes.cnpq.br/>

3.4. Sim, os verbetes são escritos em tamanho de fonte maiores, seguidos da classificação gramatical em fonte menor, o nome do autor em fonte intermediária entre a do verbete e a da classificação gramatical. O negrito introduz a definição, a fonte da definição, a nota explicativa, a fonte da nota, a leitura recomendada e os termos relacionados.

3.5. Em ordem alfabética, segundo a perspectiva de cada autor.

3.6. Verbetes em português.

3.7. Sim.

3.8. Sim, na versão Kindle (o único).

3.9. Não, obra impressa em capa comum.

3.10. Sim, os códigos das equipes que elaboraram os verbetes, das fontes bibliográficas e da bibliografia recomendada são sistematizados na parte final da obra.

3.11. Sim.

4. Sobre o conteúdo

4.1.

a) Sim.

b) Sim.

c) Sim, identificada segundo a perspectiva dos teóricos

d) Não.

e) Não se aplica, pois a obra é específica da área da linguística da enunciação.

f) Sim, no interior das notas explicativas.

g) Sim, entre verbetes que indicam noções do mesmo autor.

h) Não há.

i) Sim, a letra maiúscula V. marca a indicação de remissivas, entre os mesmos autores.

Há outra forma de remissivas que indica os verbetes integrantes da mesma rede de noções sob um mesmo autor. Ex. ele (1) Benveniste, V. não-pessoa.

a) Sim, ao final da obra, nas fontes bibliográficas e na bibliografia recomendada.

l) Sim, campo denominado **Nota explicativa**.

4.4. Sim.

4.5. Sim, aos alunos de graduação e pós-graduação.

9. Título: 'Letras' entre aspas⁹¹

Autor: Elzio Damasceno Silva⁹²

Editora: CRV

Edição: 1^a

Data: 2015

Local de publicação: Curitiba

Volume(s): 1

1. Sobre o autor

1.1. Não.

1.2. Graduação em Letras - Língua Portuguesa (Universidade Estadual do Sudoeste da Bahia, UESB, 2000)

2. Sobre a apresentação da obra pelo autor

2.1.

a) Sim.

b) Sim.

c) Não, é uma obra impressa e digital.

⁹¹

Obra

disponível

em

<https://editoracrv.com.br/produtos/busca?assunto=%E2%80%98Letras%E2%80%99+entre+aspas>. Acesso em 15 maio 2020.

⁹² Perfil do autor disponível em <http://lattes.cnpq.br/7965176353375571>

d) Parcialmente, o autor referencia as obras impressas consultadas, mas não referencia os *sites* da internet de onde as definições foram extraídas *ipsis literis* como em **sílaba**, **zetacismo** disponíveis em <https://www.osdicionarios.com/>.

2.2. Não há, o autor cita excertos de obras literárias, referencia o autor, mas não disponibiliza sua bibliografia nas referências. Ex.: Zen – a definição é proveniente do dicionário da língua portuguesa de Aurélio.

3. Sobre a apresentação material da obra

3.1. Não há.

3.2. Sim.

3.3. Não há.

3.4. Não há diferenciação. A fonte utilizada para a escrita dos verbetes é a mesma utilizada no paradigma definicional.

3.5. Em ordem alfabética, com barra lateral retrátil no app. Não houve acesso à obra impressa.

3.6. Verbetes em português.

3.7. Sim.

3.8. Sim, no app. Recentemente, a editora lançou a versão da internet, via acesso à área do cliente.

3.9. Não houve acesso à obra impressa.

3.10. Não há.

3.11. Em partes, disponível no site da editora, na *Amazon* e na *FNAC.pt*.

4. Sobre o conteúdo

4.1.

a) Não.

b) Não.

c) Sim, não identificada. Normalmente, no início do paradigma definicional. Por exemplo: significado - acepção.

d) Não. O verbo datilografar é grafado com a variação do português europeu–Dactilografar.

e) Não.

f) Sim, no PD.

g) Não há.

h) Não há.

i) Sim, por meio de *hyperlinks* para *site* externo como Wikipedia.

j) Sim, ao final da obra, nas referências

k) Não há.

4.4. Variável. Pode iniciar com um sinônimo e, em seguida, disponibilizar uma definição.

4.5. Em partes, aos alunos de graduação.

10. Título: Dicionário de gêneros textuais

Autor: Sérgio Roberto Costa⁹³

Editora: Autêntica

Edição: 3ª

Data: 2014

Local de publicação: Belo Horizonte

Volume(s): 1

1. Sobre o autor

1.1. Sim.

⁹³ Perfil do autor disponível em <http://lattes.cnpq.br/3929428427165558>.

1.2. Doutorado em Linguística Aplicada pelo LAEL/PUC de São Paulo (1997); pós-doutorado pela Universidade de Paris X/Nanterre (2002).

2. Sobre a apresentação da obra pelo autor

2.1.

a) Sim, obra de pesquisa e consulta.

b) Não.

c) Não. Obra impressa e digital.

d) Sim, o autor usa citações na microestrutura dos verbetes e disponibiliza a bibliografia nas referências ao final da obra.

2.2. Sim, como no verbete *frase* e a referência ao autor Pedrosa.

3. Sobre a apresentação material da obra

3.1. Sim, por Magda Soares.

3.2. Sim.

3.3. Sim, há quadros e figura que ilustram o conteúdo e exemplificam os conceitos na obra.

3.4. Sim, o negrito é utilizado para indicar os verbetes e fazer referência aos autores, o itálico é utilizado para exemplos e citações.

3.5. Em ordem alfabética. Na versão Kindle não há sumário alfabético. Os verbetes são acessados pelo número da página ou pela posição no Kindle.

3.6. Verbetes em português. Ao final da obra, há um adendo com palavras estrangeiras e uma proposta de ortografia aportuguesada.

3.7. Não houve acesso à obra impressa. A versão Kindle é de fácil manuseio.

3.8. Sim, na versão Kindle.

3.9. Não houve acesso à obra impressa.

3.10. Sim, o autor utiliza o V. para indicar as remissivas e marca as notas de rodapé com numeração progressiva.

3.11. Sim, encontra-se na 3ª edição.

4. Sobre o conteúdo

4.1.

a) Não.

b) Não.

c) Sim, marcados pela expressão *o mesmo que*. Ex.: Artigo de fundo - o mesmo que editorial.

d) Sim, como em *lead* e *lide*. Somente a lista de palavras estrangeiras (Adendo, ao final da obra) e sua proposta de aportuguesamento.

e) Não.

f) Sim, na microestrutura do PD.

g) Caso o termo seja usado em LP, o autor traz seus equivalentes como em *bate-papo virtual*, *ciberconversa*, *chat*, *conversa/conversação*, *conversação oral digital*, *papo*.

h) Não há.

i) Sim, na macroestrutura, as remissivas vêm precedidas de V., enquanto na microestrutura o V. vem após a remissiva.

j) Sim, ao final da obra, nas referências

k) Sim, na forma de nota de rodapé (NR.).

4.4. Sim. O autor define os verbetes e explica os conceitos na microestrutura, se necessário.

4.5. Sim, de linguagem acessível com informações bem organizadas.

11. Título: Bases para um dicionário linguístico-gramatical⁹⁴

⁹⁴ Disponível em http://www.filologia.org.br/xi_sinefil/completos/bases_PEREIRA.pdf. Acesso em 15 maio 2020.

Autor: José Pereira da Silva ⁹⁵

Editora: Não há, pois está disponível on-line.

Edição: Não há.

Data: 2019

Local de publicação: <http://www.josepereira.com.br/DL-G/DL-G-sumario.pdf>

Volume(s): 1

1. Sobre o autor

1.1. Sim.

1.2. Doutorado em Linguística e Filologia- UFRJ (1991).

2. Sobre a apresentação da obra pelo autor

2.1.

a) Sim, obra de referência.

b) Sim, estudantes e profissionais das áreas de linguística e letras usuários da língua portuguesa.

c) Não. Obra em suporte eletrônico.

d) Sim, o autor cita o teórico, o ano e a obra que embasou os dados. A bibliografia nas referências ao final da obra (398 p.).

2.2. Sim, o autor define os verbetes e registra como os teóricos concebem os conceitos de cada termo.

3. Sobre a apresentação material da obra

3.1. Não há.

3.2. Sim.

3.3. Sim, são apresentadas de forma racional. No início cada letra do alfabeto, o autor disponibiliza uma imagem da origem da letra a partir do fenício, grego ou latim.

3.4. Sim, o negrito é utilizado para indicar os verbetes, o itálico para indicar as remissivas.

3.5. Em ordem alfabética. Na versão disponibilizada on-line, cada arquivo corresponde a uma letra do alfabeto, em um total de 6869 páginas, disponibilizadas por meio de um *hyperlink*.

3.6. Verbetes em português.

3.7. Não houve acesso à obra impressa, lançada em julho/2020. A versão on-line é de fácil manuseio.

3.8. Sim, na versão on-line.

3.9. Não houve acesso à obra impressa. A obra digital está disponível na página do autor ⁹⁶.

3.10. Sim, o autor utiliza a abreviação s.v. para indicar que a referência se encontra no arquivo bibliografia ⁹⁷.

3.11. Parcialmente, o primeiro produto impresso encontra-se disponível à venda em livrarias on-line (Vocabulário Terminológico Geral de Linguística e Letras - *Amazon*, Editora Autografia, Americanas entre outras).

4. Sobre o conteúdo

4.1.

a) Sim.

b) Não.

c) Sim, marcados pela expressão *o mesmo que*. Ex.: Semicólon é o mesmo que ponto e vírgula.

d) Sim, como em *lead* e *lide* ⁹⁸.

e) Não.

f) Sim, na microestrutura do PD.

⁹⁵ Perfil do autor disponível em <http://lattes.cnpq.br/4567157761169588>.

⁹⁶ Obra disponível em <http://www.josepereira.com.br/DL-G/DL-G-sumario.pdf>. Acesso em 15 maio 2020.

⁹⁷ Disponível em http://www.josepereira.com.br/DL-G/DL-G_bibliografia.pdf. Acesso em 15 maio 2020.

⁹⁸ Disponível em http://www.josepereira.com.br/DL-G/DL-G_L1.pdf. Acesso em 15 maio 2020.

- g) Não, somente em casos em que o termo é estrangeirismo no Brasil.
 - h) Não há.
 - i) Sim, na microestrutura, ao final dos verbetes, há o ver também ou o ver.
 - a) Sim, as fontes permeiam toda a microestrutura, quer seja como exemplificação, como referência ou como leituras complementares.
 - l) Não há.
- 4.4. Sim. O autor define os verbetes e explica os conceitos na microestrutura, conforme os teóricos ou as escolas linguísticas.
- 4.5. Sim, de linguagem acessível com informações bem organizadas e bem explicadas.

APÊNDICE O - Roteiro para avaliação de dicionários ou glossários de Linguística disponíveis on-line segundo Faulstich (2011).

Título:

Autor:

Data:

Disponível em:

1. Sobre o autor

- 1.1. Trata-se de pessoa reconhecida na área de dicionarística ou de terminologia?
- 1.2. Qual a formação acadêmica do autor principal e dos participantes do grupo de pesquisa?

2. Sobre a apresentação da obra pelo autor

2.1. Há introdução na qual apareçam claramente:

- a) os objetivos da obra?
- b) o público para o qual o conteúdo se dirige?
- c) as informações sobre como consultar o dicionário ou vocabulário?
- d) referências à bibliografia de onde foi extraído o corpus?

2.2. Há bibliografia de consulta justificada pelo autor?

3. Sobre a apresentação virtual da obra

3.1. Há prefácio redigido por personalidade reconhecida na área de dicionarística? Científica, técnica?

3.2. A família tipográfica empregada é adequada à faixa etária do usuário?

3.3. As ilustrações, se houver, estão adequadas à microestrutura informacional?

3.4. A utilização de negrito, de itálico e de outros recursos gráficos está de acordo com o equilíbrio visual da obra?

3.5. Os verbetes são apresentados em ordem alfabética? Em ordem sistemática?

3.6. A obra contempla uma só língua? Mais de uma?

3.7. O leiaute do dicionário ou vocabulário permite navegação prática e fácil (*easy-to-use*)?

3.8. O sistema de abreviações e de símbolos aparece corretamente no corpo do texto?

3.9. A obra possui ampla divulgação?

4. Sobre o conteúdo

4.1. Os verbetes apresentam:

- a) categoria gramatical?
- b) gênero?
- c) sinonímia?
- d) variante(s) da entrada?
- e) indicação de área ou subárea de especialidade?
- f) contexto? (exemplo ou abonação?)
- g) equivalente(s)?
- h) indicação de pronúncia?
- i) remissivas úteis entre conceitos?
- j) fontes?
- k) notas?

4.2. A definição é constituída de um enunciado de uma só frase?

4.3. A definição leva em conta o nível de discurso do usuário?

APÊNDICE P - Resultado da análise dos dicionários ou glossários de Linguística disponíveis on-line segundo Faulstich (2011).

1. Título: Glossary of Grammar Terms

Autor: Bill Johanson

Data: 1996-2020 Word Place, Inc.

Disponível em: <http://www.dailygrammar.com/glossary.html>

1. Sobre o autor

1.1. Não. Professor de ensino fundamental e médio durante 30 anos e professor de cursinho.

1.2. Não disponível.

2. Sobre a apresentação da obra pelo autor

2.1. Há introdução na qual apareçam claramente:

- a) Servir ao ensino facilitado da gramática.
- b) Professores da escola pública, alunos do ensino fundamental final.
- c) Informa que os verbetes são relacionados às aulas de gramática.
- d) Não há.

2.2. Não há.

3. Sobre a apresentação virtual da obra

3.1. Não há. Obra apresentada por Pete Peterson, ex vice-presidente executivo da Word Perfect, um dos editores de *software* de computadores da Microsoft.

3.2. Sim. O consulente é livre para aumentar ou diminuir a fonte.

3.3. Não há.

3.4. O negrito marca os verbetes. As remissivas são marcadas com número e um *hiperlink*.

3.5. Em ordem alfabética.

3.6. Uma língua: inglês.

3.7. Sim, todos os verbetes são disponibilizados na mesma página HTML com *hiperlink* numerado para as lições.

3.8. Não há.

3.9. Obra disponível na internet.

4. Sobre o conteúdo

4.1. Os verbetes apresentam:

- a) Não há.
- b) Não há.
- c) Não há.
- d) Não há.
- e) Não há.
- f) Disponível na microestrutura do verbete e na aula disponível via *hiperlink*.
- g) Não há.
- h) Não há.
- i) Sim, disponível no *hiperlink* da aula.
- j) Não há.
- k) Não há.

4.2. Sim.

4.3. Não, definição muito técnica para aprendizes leigos.

2. Título: Index of French Grammar and Pronunciation Glossary Terms⁹⁹

⁹⁹ ThoughtCo. Index of French Grammar and Pronunciation Glossary Terms. Disponível em: thoughtco.com/french-grammar-and-pronunciation-glossary-1368855. Acesso em 01 ago. 2020.

Autor: Jocelly Meiners¹⁰⁰

Data: 1996-2020

Disponível em: <http://www.dailygrammar.com/glossary.html>

1. Sobre o autor

1.1. Não. Professora do Departamento de espanhol e português da Universidade do Texas em Austin; atualmente, ela é especializada em cursos para alunos de espanhol da Heritage e codirige o *Texas Consortium for Heritage Spanish*.

1.2. Doutorado em Linguística Hispânica e mestrado em Linguística Francesa pela Universidade do Texas em Austin.

2. Sobre a apresentação da obra pelo autor

2.1. Há introdução na qual apareçam claramente:

- a) Servir como material de apoio para o aprendizado de língua francesa
- b) Usuários da internet, aprendizes de língua francesa.
- c) Busca por tópicos ou por ordem alfabética.
- d) Não há.

2.2. Não há.

3. Sobre a apresentação virtual da obra

3.1. Não há.

3.2. Sim. O consulente é livre para aumentar ou diminuir a fonte.

3.3. Não há.

3.4. Os termos gramaticais com *hiperlink* são realçados na página.

3.5. Em ordem alfabética.

3.6. Bilíngue inglês e francês.

3.7. Parcialmente, a maioria dos verbetes são disponibilizados na mesma página HTML com *hiperlink*. No canto superior esquerdo da página há *hiperlinks* para pronúncia, conversação, vocabulário e recursos didáticos.

3.8. Não há.

3.9. Obra disponível na internet.

4. Sobre o conteúdo

4.1. Os verbetes apresentam:

- a) Não há.
- b) Não há.
- c) Não há.
- d) Não há.
- e) Não há.
- f) Sim, disponível na página acessível pelo *hiperlink* originado no verbete. Há também informações disponíveis em vídeo e acesso a recursos didáticos como exercícios
- g) Sim.
- h) Sim, mas não especificamente do termo¹⁰¹.
- i) Sim, disponível na parte superior da página¹⁰².
- j) Não há.
- k) Não há.

4.2. Sim.

4.3. O usuário falante de LI.

¹⁰⁰ Perfil disponível em <https://www.thoughtco.com/about-us>

¹⁰¹ French Words Starting With A, B and C, disponível em <https://www.thoughtco.com/audio-dictionary-a-b-and-c-4085195>. Acesso em 01 ago. 2020.

¹⁰² Understanding 'Si' Clauses in French, disponível em <https://www.thoughtco.com/french-si-clauses-1368944>. Acesso em 01 ago. 2020.

3. Título: Ciencias del Lenguaje - 4 - Lingüística-TÉRMINOS ¹⁰³

Autor: Grupo de Ciencias del Lenguaje

Data: 2016

Disponível

em:

<http://www.diccionariosdigitales.net/glosarios%20y%20vocabularios/ciencias%20del%20lenguaje-4-linguistica-terminos.htm>

1. Sobre o autor

1.1. Não disponível.

1.2. Não disponível.

2. Sobre a apresentação da obra pelo autor

2.1. Há introdução na qual apareçam claramente:

a) Não há.

b) Não há.

c) Não há.

d) Não há.

2.2. Não há.

3. Sobre a apresentação virtual da obra

3.1. Não há.

3.2. Sim. O consulente é livre para aumentar ou diminuir a fonte.

3.3. Não há.

3.4. Os termos são digitados em letras maiúsculas.

3.5. Em ordem alfabética, da letra A à letra L.

3.6. Monolíngue em espanhol.

3.7. Sim, todos os verbetes são disponibilizados em uma única página.

3.8. Não há.

3.9. Lista disponível na internet.

4. Sobre o conteúdo

4.1. Os verbetes apresentam:

a) Não há.

b) Não há.

c) Não há.

d) Não há.

e) Termos da linguística.

f) Sim, disponibilizado entre parênteses pelo desenvolvedor da página.

g) Não há.

h) Não há.

i) Não há.

j) Não há.

k) Não há.

4.2. Sim.

4.3. O usuário não é especificado.

4. Título: Systematic Dictionary of Corpus Linguistics ¹⁰⁴

¹⁰³

Lingüística-TÉRMINOS.

Disponível

em:

<http://www.diccionariosdigitales.net/glosarios%20y%20vocabularios/ciencias%20del%20lenguaje-4-linguistica-terminos.htm>. Acesso em 01 ago. 2020.

¹⁰⁴ Systematic Dictionary of Corpus Linguistics, Systematic Dictionary of Corpus Linguistics. Disponível em: <http://donelaitis.vdu.lt/publikacijos/SDoCL2.htm>. Acesso em 01 ago. 2020.

Autor: *Lithuanian terminological system of Corpus Linguistics*¹⁰⁵

Data: s/d

Disponível em: <http://donelaitis.vdu.lt/publikacijos/SDoCL.htm>

1. Sobre o autor

1.1. Pesquisadores da LC.

1.2. Doutorado.

2. Sobre a apresentação da obra pelo autor

2.1. Há introdução na qual apareçam claramente:

a) É uma tentativa de agrupar, sistematizar, definir e explicar os termos básicos da LC em inglês; pode ser usado como dicionário e como um livro introdutório de LC.

b) Não há.

c) Não há.

d) Não há.

2.2. Sim, disponível no link *References*¹⁰⁶.

3. Sobre a apresentação virtual da obra

3.1. Não há.

3.2. Em partes, a fonte é muito pequena. O consulente é livre para aumentar ou diminuir a fonte.

3.3. Não há.

3.4. Os termos são digitados em negrito e itálico.

3.5. Em ordem sistemática e alfabética, sob cada termo superordenado com *hiperlink*.

3.6. Monolíngue em inglês.

3.7. Não, os verbetes são disponibilizados em uma única página com fonte muito pequena, menor que 10.

3.8. Não há.

3.9. Não é possível identificar, mas a obra encontra-se disponível na internet.

4. Sobre o conteúdo

4.1. Os verbetes apresentam:

a) Não há.

b) Não há.

c) Não há.

d) Não há.

e) Termos da LC.

f) Sim, após a abreviação e.g.

g) Não há.

h) Não há.

i) Sim, escritos em negrito.

j) Sim, ao final da definição do verbete.

k) Não há.

4.2. Sim.

4.3. O usuário não é especificado, mas presumimos que seja uma obra destinada a alunos universitários pelo perfil da instituição.

5. Título: Bilingual Grammar Glossary for Japan

Autor: David V. Appleyard

Data: 1998-2020

¹⁰⁵ Perfil do grupo de pesquisa disponível em <https://klc.vdu.lt/en/about/ccl-team/>. Acesso em 01 ago. 2020.

¹⁰⁶ *References*, disponível em: <http://donelaitis.vdu.lt/publikacijos/SDoCL3.htm>. Acesso em 01 ago. 2020.

Disponível em: <https://davidapppleyard.com/english/jgrammar.htm>

1. Sobre o autor

1.1. Não disponível.

1.2. Não disponível.

2. Sobre a apresentação da obra pelo autor

2.1. Há introdução na qual apareçam claramente:

- a) Disponibilizar informações linguísticas de forma clara e objetiva para exames.
- b) Falantes de língua japonesa aprendendo a língua inglesa.
- c) Não há. De acordo com o autor, o *site* pode ser acessado por celulares, mas o ideal é o acesso a partir de computadores pessoais ou *notebook*.
- d) Não há, porém o *site* é como uma plataforma pedagógica para aprendizagem de língua inglesa.

2.2. Não há.

3. Sobre a apresentação virtual da obra

3.1. Não há.

3.2. Sim, as páginas têm perfil equilibrado.

3.3. Não há.

3.4. Sim.

3.5. Em ordem sistemática, como *parts of speech, syntax, verb tenses*.

3.6. Obra bilíngue em português e japonês.

3.7. Sim.

3.8. Não, disponibilizadas no rodapé quando necessárias.

3.9. Sim, disponível em *websites* de vários países.

4. Sobre o conteúdo

4.1. Os verbetes apresentam:

- a) Não há.
- b) Não há.
- c) Não há.
- d) Não há.
- e) Subárea como sintaxe, modalidade e estilo.
- f) Sim.
- g) Sim. Em partes, o site disponibiliza arquivos de áudio¹⁰⁷ para a prática de pronúncia em LI.
- h) Não há.
- i) Não há.
- j) Não há.

4.2. Sim.

4.3. Sim, alunos em processo preparatório para exames.

6. Lexicon of linguistics

Autor: Peter Ackema, Nine Elenbaas, Esther Janse, Iris Mulders, Jean Rutten, Ellis Visch.

Data: 1996-2001

Disponível em: <https://lexicon.hum.uu.nl/>

1. Sobre o autor

1.1. Professor de Linguística na Universidade de Edimburgo.

1.2. Não disponível.

2. Sobre a apresentação da obra pelo autor

2.1. Há introdução na qual apareçam claramente:

- a) Não há.
- b) Não há.
- c) Sim, inserindo o termo na busca ou clicando na letra do alfabeto.
- d) Sim, referências disponíveis no site.

2.2. Sim, o verbete é ligado à biografia pelo ano de publicação da obra como *hiperlink*.

3. Sobre a apresentação virtual da obra

3.1. Não há.

3.2. Sim. O consulente é livre para aumentar ou diminuir a fonte.

3.3. As árvores linguísticas em termos do gerativismo. Ex.: *syntax*.

3.4. Os verbetes são escritos em letras maiúsculas e negrito. As remissivas são marcadas em negrito com letra minúscula e *hyperlinks*.

3.5. Em ordem alfabética ou pela busca personalizada pela digitação na caixa de texto.

3.6. Uma língua: inglês.

3.7. Sim.

3.8. Sim, *also see* para remissivas; *opposed to* para antônimos e *i.e.* para indicar exemplos.

3.9. Site disponível na internet.

4. Sobre o conteúdo

4.1. Os verbetes apresentam:

- a) Não há.
- b) Não há.
- c) Não há.
- d) Não há.
- e) Sim, logo após o verbete.
- f) Sim, como em *clitic*.
- g) Não há.
- h) Não há.
- i) Sim, disponível na definição como *hiperlink*.
- j) Sim, ao final de cada verbete, com o *hiperlink* da bibliografia.
- k) Não há.

4.2. Sim.

4.3. Perfil de usuário não identificado.

7. Título: SIL Glossary of Linguistic Terms

Autor: Eugene E. Loos (Editor-chefe), Susan Anderson, Dwight H. Day, Jr., Paul C. Jordan e J. Douglas Wingate (editors).

Data: 2003

Disponível em: <https://glossary.sil.org/> e <https://feglossary.sil.org/fr?language=en>

1. Sobre o autor

1.1. Não.

1.2. Linguista.

2. Sobre a apresentação da obra pelo autor

2.1. Há introdução na qual apareçam claramente:

- a) Fornecer informações em inglês e em francês na forma de glossários e bibliografias projetadas para apoiar pesquisas linguísticas.
- b) Pesquisadores na área de linguística.
- c) Não há. Site de perfil intuitivo.

d) Sim, bibliografia completa (de A-Z) disponibilizada no site.¹⁰⁸

2.2. Sim.

3. Sobre a apresentação virtual da obra

3.1. Não há.

3.2. Sim.

3.3. Disponibilizam árvore de domínio, como em *Directive Modality*.

3.4. Sim.

3.5. Em ordem alfabética.

3.6. Obra bilíngue em inglês e francês.

3.7. Sim, com ícones de acesso ao glossário de termos e à bibliografia.

3.8. Sim.

3.9. A obra é disponível on-line para acesso gratuito.

4. Sobre o conteúdo

4.1. Os verbetes apresentam:

a) Não há.

b) Sim, no francês.

c) Não há.

d) Não há.

e) Sim, abaixo do verbete final.

f) Sim.

g) Sim.

h) Não há.

i) Sim, por meio de *hyperlinks* e numeração.

j) Sim.

k) Não há.

4.2. Sim.

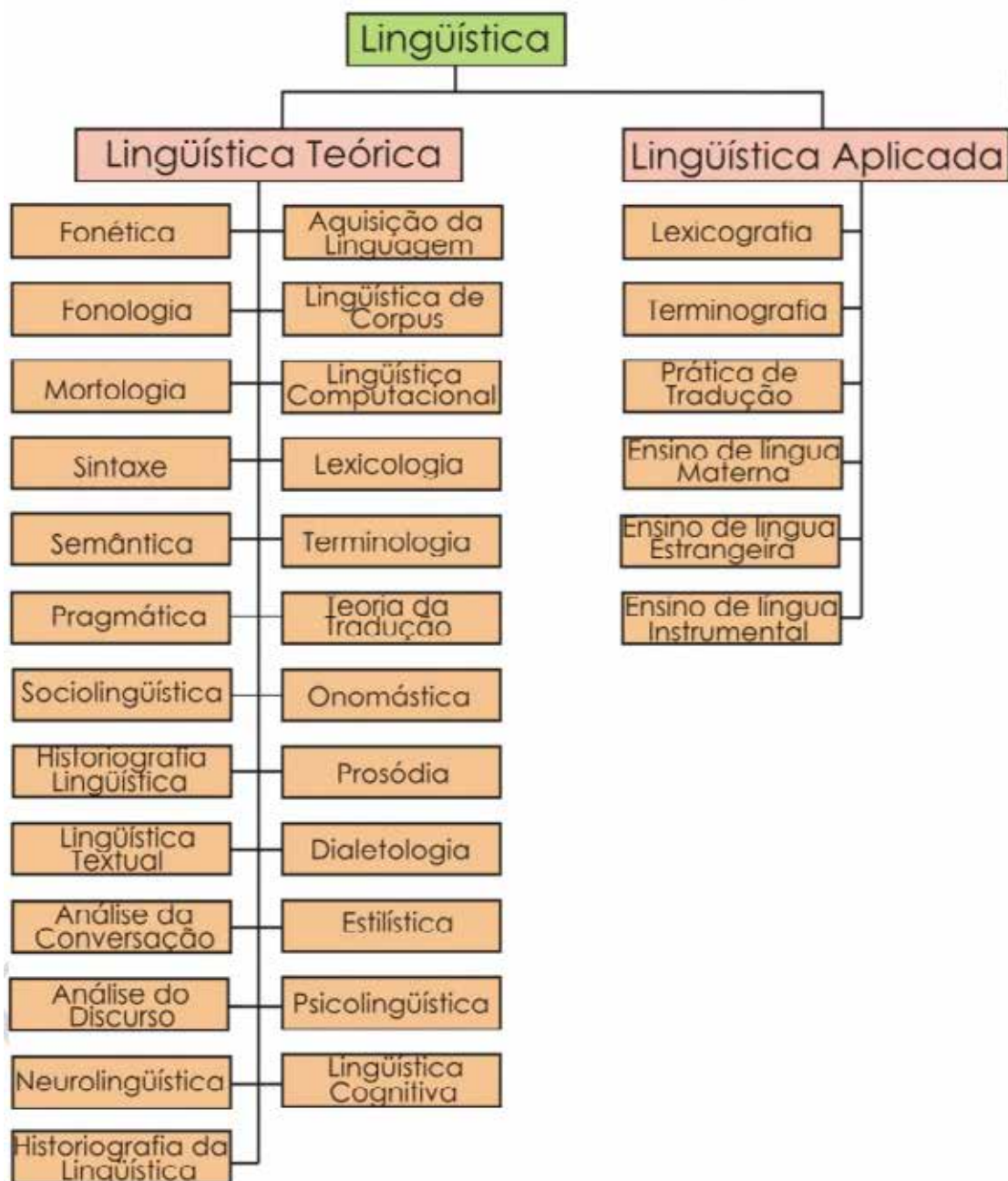
4.3. Sim, alunos em processo preparatório para exames.

¹⁰⁸ Bibliography (Linguistics), disponível em: <https://glossary.sil.org/bibliography>. Acesso em 01 de ago. de 2020.

ANEXOS: Árvores de domínio da Linguística propostas por Fromm

ANEXO A - Árvore de campo da Linguística - Fromm (2008).

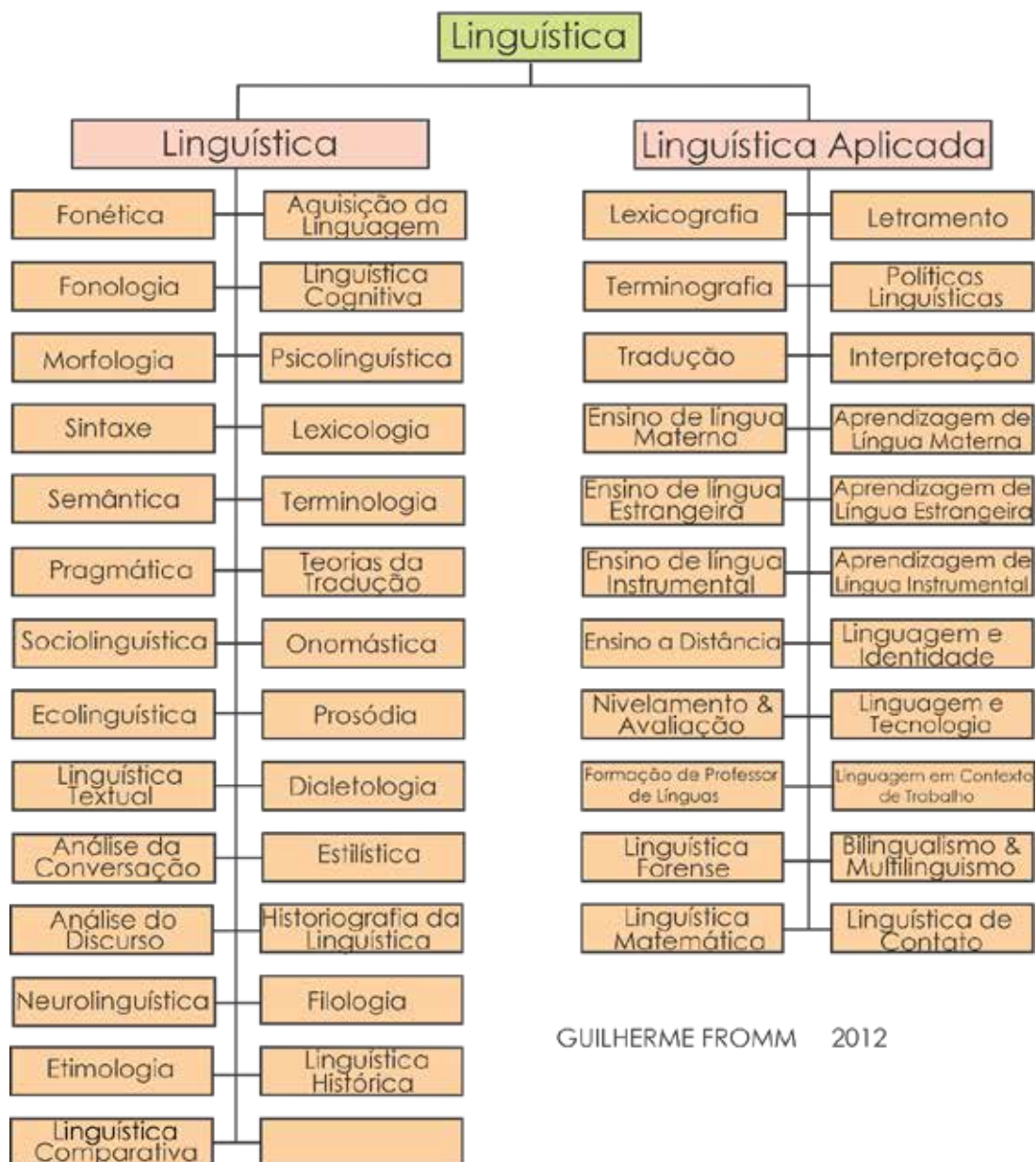
Árvore do Campo da Lingüística



Fonte: Fromm (2008).

ANEXO B - Árvore de campo da Linguística em discussão (2013).

Árvore do Campo da Linguística

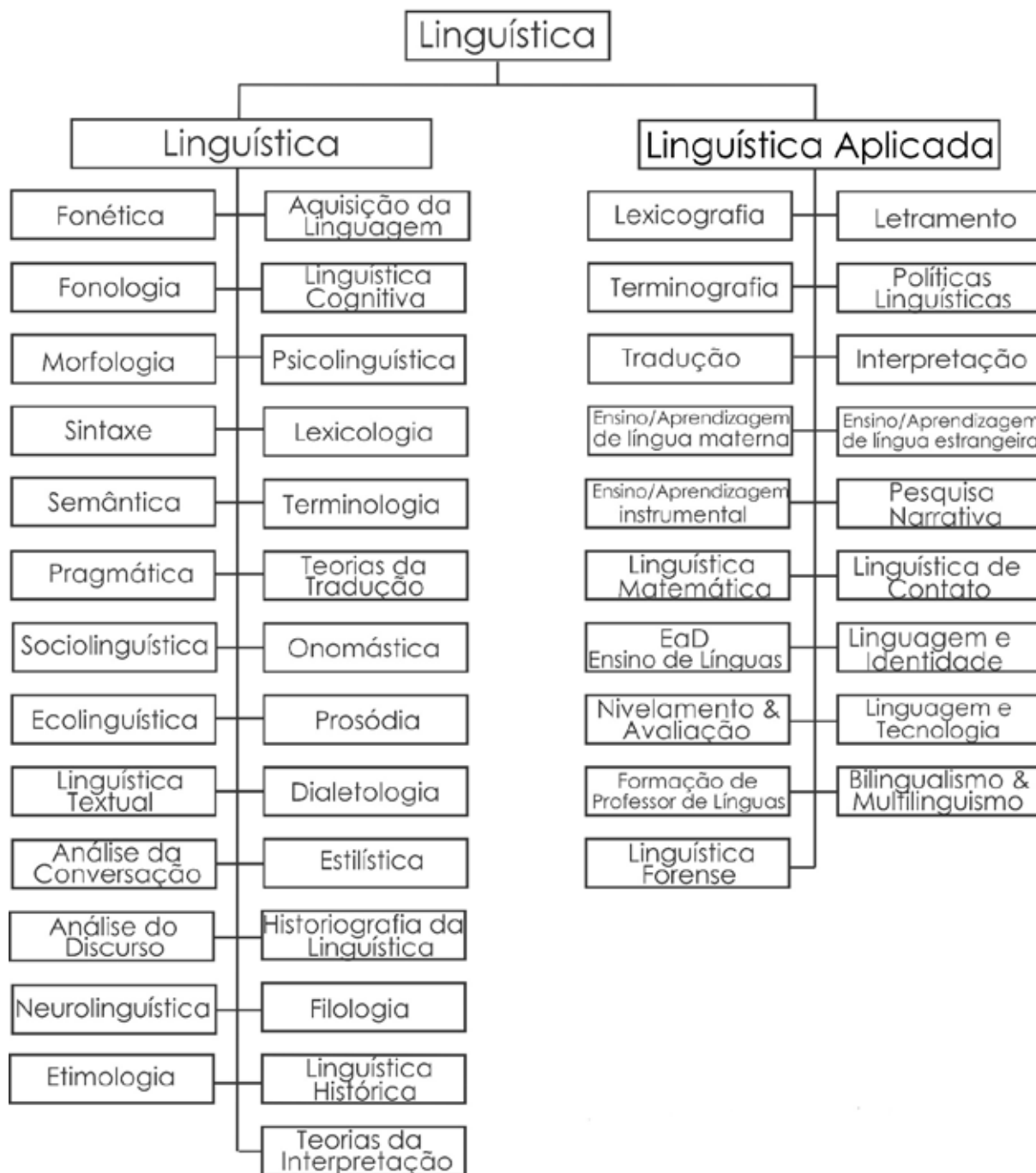


GUILHERME FROMM 2012

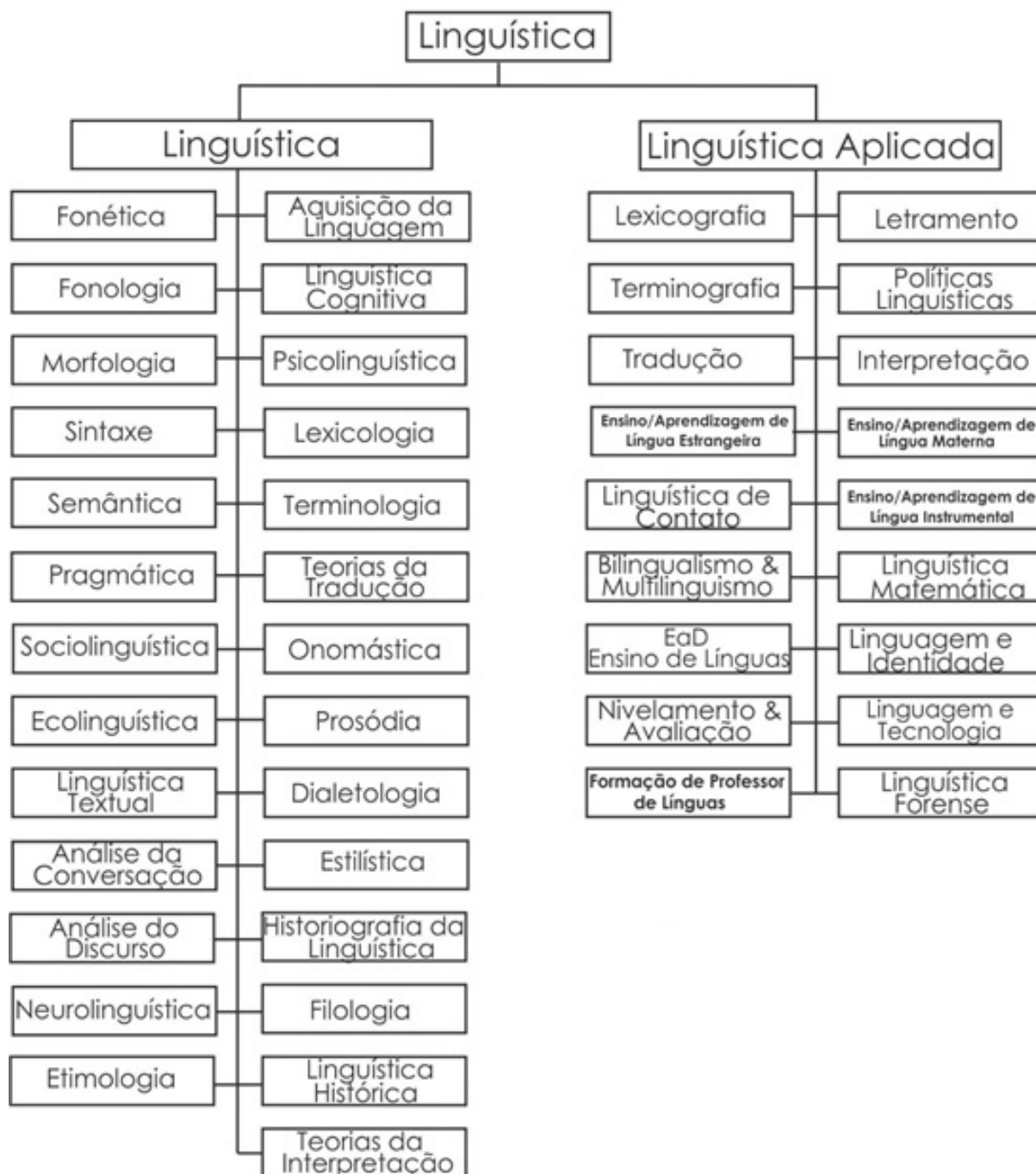
Fonte: Fromm & Yamamoto (2013).

ANEXO C - Árvore de campo da Linguística, segundo Fromm, atualizada até 2014.

Árvore do Campo da Linguística



Fonte: Yamamoto (2015).

ANEXO D - Árvore de domínio da **Linguística** (2018).*Árvore do Campo da Linguística*

Fonte: Fromm (2018).