Círculo Fluminense de Estudos Filológicos e Linguísticos

#### LINGUÍSTICA HISTÓRICA E LINGUÍSTICA DE CORPUS: UMA PROPOSTA PARA DICIONÁRIO BILÍNGUE PORTUGUÊS-INGLÊS

Márcio Issamu Yamamoto (UFU) issamu2009@gmail.com

#### 1. Introdução

O objetivo deste trabalho é apresentar, explicitar e detalhar os passos que serão tomados para confeccão de um dicionário bilíngue português-inglês, na área de linguística histórica e, metodologicamente embasado na linguística de corpus, denominado VoTec - Vocabulário Técnico – disponível na rede mundial de computadores<sup>97</sup> e de acesso gratuito. Esse dicionário tem o intuito de servir a profissionais da Linguística, tradutores e quaisquer outros profissionais interessados na área de traducão e obras bilíngues. Para a confecção deste dicionário serão levantados corpora em português e inglês de linguística histórica, etimologia e filologia. Posteriormente estes corpora serão tratados com a ferramenta Wordsmith Tools (SCOTT, 2009)98 para construção de uma lista de palavras (Wordlist) e de palavras-chave (Keywords). A partir dessas listas serão selecionados os termos que serão dicionarizados nessas duas línguas, cujas definições serão provenientes dos contextos nos quais eles estiverem inseridos. Esses contextos serão identificados pelo concordanciador (*Concord*), de onde extrairemos os conceitos para a formação das definições. As definições serão inseridas na plataforma do VoTec e posteriormente disponibilizadas na Internet.

#### 2. Linguística histórica e linguística de corpus

A concepção desta obra terminográfica será beneficiada pela linguística histórica e pela metodologia da linguística de *corpus*. Aquela responsável por estudar a história das línguas e descrevê-las em seus contextos histórico-sociais (COSERIU, 1979, p. 236; SILVA, 2008, p. 30), e esta proveniente da evolução tecnológica e usada para o tratamento de *corpora* como modo de investigação empírica da linguagem (VIANA,

<sup>&</sup>lt;sup>97</sup> Dicionário disponível em <<u>www.pos.voteconline.com.br</u>>. Acesso em: 23-03-2013.

<sup>&</sup>lt;sup>98</sup> Programa disponível em http://lexically.net/wordsmith/version5. Acesso em: 23-03-2013.

Círculo Fluminense de Estudos Filológicos e Linguísticos

2010, p. 27, 34).

Como definição de *corpus*, apoiamo-nos em Viana (2010, p. 27) que o define como uma compilação de textos de ocorrência natural que representa uma certa língua ou seus aspectos mais específicos para que possibilite uma análise linguística pré-estabelecida. Os corpora que serão utilizados para esta obra terminográfica serão compilados na área da linguística teórica, subárea da filologia, etimologia e linguística histórica. A linguística histórica foi escolhida pois é uma área que pode ser útil ao ensino de língua portuguesa e para o ensino de línguas estrangeiras, mais especificamente com sua contribuição com os metaplasmos. Além dessa vantagem, há aspectos das áreas da fonologia e ortografia que permeiam o ensino da língua portuguesa e que podem ser elucidados com o auxílio da linguística histórica. Esses corpora são de caráter acadêmico, tais quais artigos científicos, dissertações e teses. Essa variedade de textos contribui para o enriquecimento da "gama de dados e aproxima um pouco mais o linguista da ideal onisciência dos dados da huge chart bloomfieldiana." (VIARO, 2011, p. 104).

#### 3. Árvore de domínio

Para iniciar, tomaremos a árvore de domínio da linguística com o objetivo de delimitar o campo de pesquisa terminográfica na área da linguística histórica. Para construção da obra terminográfica teremos uma macro e uma microestrutura que servem à elaboração das fichas terminológicas, nas quais os termos serão selecionados e inseridos por ordem de maior frequência nos *corpora*.



Figura 1. Árvore de domínio da linguística em construção segundo Fromm (2012).

#### 4. Coleta e análise de corpora

Os *corpora* que estão sendo usados para extração de termos e contextos para criação do banco de dados são arquivos da área acadêmica, sendo eles artigos científicos, dissertações, teses ou livros nas subáreas de etimologia, filologia e linguística histórica. Os arquivos são de acesso público, baixados em formato pdf e salvos em formato txt a partir de *sites* específicos da internet. Dentre eles, citamos os *sites* da *Revista Philologus*, disponível em <http://filologia.org.br/revista>, e *Filologia e Linguística Portuguesa*, disponível em

<http://www.fflch.usp.br/dlcv/lport/flp> para o *corpus* de língua portuguesa e a base de dados Jstor, disponível em: <http://www.jstor.org> para o *corpus* de língua inglesa. Ressalvamos que este último disponibiliza

um maior número de artigos ao ser acessado de um servidor público federal devido a acordos interinstitucionais. Os arquivos são salvos em formato txt para possibilitar a leitura pelo programa WordSmith Tool doravante WST -, ferramenta usada para análises lexicais. O tamanho dos corpora usados é de 536.330 palavras para o português, com 33 textos, e de 521.794 para o inglês, com 8 textos. De acordo com a classificação proposta por Viana (2011, p. 30), esses corpora classificam-se como: especializado, escrito, sincrônico, contemporâneo, estático, bilíngue não paralelo e de primeira língua. O corpus especializado, que objetiva a linguagem de especialidade contrasta com o de língua geral. O escrito contrasta com o oral; o sincrônico é aquele que traz o registro de uma dada língua num momento específico da história, enquanto o diacrônico cobre períodos diferentes dessa língua. É contemporâneo pois reflete a producão do tempo presente: estático significa que o corpus não permite ser alterado, ou seja, a inclusão ou exclusão de textos não pode ser realizada. Bilíngue, nesse caso, por abarcar a língua portuguesa e inglesa, não paralelo pois o corpus do português não é uma tradução do texto em inglês ou vice-versa. Um corpus de primeira língua significa que seus autores têm essa língua como língua nativa.

UHLPO	ortugues_)	wordlist.lst	Sector Sector			- 1997
File	Edit	View	Compute	Settings	Windows	Help
N						text fie
1	C:\Use	rs'MarcioV	Dropbox'MAS	TERIVOTEC	LH_bd'Portu	igues'io vocabulário têxtil e a história da língua portuguesa_18_02_2013 txt
- 4	Users'	Aarcio Dro	pbox/MASTE	RIVOTECILH	tod Portugue	IS/O LATIM NA FORMAÇÃO DO PROFESSOR DE LETRAS_17_02_2013
3			C.\Users\M	larcio/Dropbo	or MASTERIN	/OTEC/LH_txt/Portugues/ORTOGRAFIA concurso público_17_02_2013.bxt
- 4				C:\Users	Marcio/Droph	xx/MASTER/VOTEC/LH_txt/Portugues/O-PRAZER_freud_16_02_2013.txt
5				C:\Users\/	Aarcio/Dropbo	x/MASTER/VOTEC/LH_btliPortugues/LingHist_FARACO_19_11_2011.bt
6	C	Users Wa	arcio/Dropbox	MASTERIVI	DTEC/LH_txt	Portugues Laboratório de História do Português Brasileiro_19_11_2011 txt
7		C:\Users	Marcio/Dropi	box MASTER	RIVOTECILH_	.txt/Portugues/IO futuro da lingua portuguesa ontern e hoje_16_02_2013.txt
8		C:V	Users'Marcio	Dropbox MA	STER/VOTER	CILH_bd\Portugues/METAPLASMOS POR SUPRESSÃO_17_02_2013.bd
9	0	:\UsersW	farcio/Dropbo	xMASTER()	OTECLH_bo	t/Portugues/Teoria da mudanca linguistica_Virginia Matos_04_04_2012.txt
10				C:\Users\Ma	ecio/Dropbex	MASTER/VOTEC/LH_txt/Portugues/Revista Philologus N.50_5-8_2011.txt
11		C	Users/Marcie	o\Dropbox\M	ASTER/VOTE	EC/LH_txt\Portugues/Uma perspectiva linguistica historica_04_04_2012.txt
12	C:VU	ers'Marci	o/Dropbox/MA	ASTERIVOTE	CLH_bd/Por	tugues/TESE_ERICA_SANTOS_SOARES_DE_FREITAS_18_02_2013.bd
13	6	: WsersW	farcio/Dropbo	KMASTER W	OTEC/LH_ba	t/Portugues/Paul Teyssier - História da Língua Portuguesa_17_02-2013.txt
14			C:\Users	Marcio/Drop	box'MASTER	WOTECILH_bt/Portugues/PAIXAODESOUZA_MC-2006a_20_01_2012.bt
15				C:\Users\Ma	rcio/Dropbex/	MASTER/VOTEC/LH_bd/Portugues/Revista Philologus N.49_1-4_2011 bd
16	Users'Ma	arcio/Dropi	box MASTER	VOTECILH	bit/Portugues	VPREFIXOS E PREFIXOIDES LATINOS E VERNÁCULOS_17_02-2013 ba
17				C:\Users\M	arcio\Dropbox	MASTER/VOTEC/LH_txt/Portugues/historiografia_jarbas_04_04_2012.txt
18		C:Ws	ers'Marcio\Dri	opbox/MAST	ERIVOTECIL	H_txt/Portugues/A IMPORTÂNCIA DA LÍNGUA ROMENA_17_02_2013.txt
19	M/xodqo	ASTERIV	OTEC/LH_txt/	Portugues\5	9555697-08-8	STRANGEIRISMOS-NO-LEXICO-PORTUGUES-DIACRONIA_10_03_2012
20			C	Users'Marci	o/Dropbox/M/	ASTER/VOTEC/LH_txt/Portugues/Alexandrino_Linguistica_19_03_2011.txt
21	Marcio/D	W/xodqor	ASTERWOTE	C/LH_txt/Por	tugues A IMF	ORTÂNCIA DO LATIM PARA O ENSINO DO PORTUGUÊS_17_02_2013
22			C:10	Isers'Marcio	Dropbox/MAS	STER/VOTEC/LH_txt/Portugues/2004_verbos_abundantes_16_02_2013.txt
23			C:\UsersW	farcio/Dropbe	x MASTER!	/OTEC/LH_txt/Portugues/1996_Respostas_da_morfologia_16_02_2013.txt
24	io/Dropb	ox MASTE	RWOTECILH	txt'Portugu	es\58509871-	OS-SUFIXOS-PORTUGUESES-NUMA-VISAO-DIACRONICA_10_03_2012
25	pbox'MA	STERIVO	TEC/LH_txt/F	ortugues\52	4664-ORIENT	ACOES-ATUAIS-DA-LINGUISTICA-HISTORICA-BRASILEIRA_12_03_2012
26	Isers/Mar	cio'Dropbe	ox MASTERN	OTECUH b	xt'Portugues'	escrita colonial brasileira em uma carta oficial do século XVII_18_02_2013
27			c	Wsers Marc	io/Dropbox/M	ASTER/VOTEC/LH_bd/Portugues/Do Latim ao português_18_02_2013.bd
28		C:Ws	ers'Marcio\Dr	opbox/MAS1	ERIVOTECIL	H_txt\Portugues\Filologia Textual e Lingüística Românica_17_02_2013 txt
29			C:\Users	Marcio/Drop	box'MASTER	WOTECILH_txt\Portugues\Filologia Portuguesa no Brasil_17_02_2012.txt
30			C:\UsersW	larcio/Dropbo	MASTER N	/OTECILH_txt/Portugues/Arqueologia das Linguas de Sinais_4_4_2012 txt
31					Cillues	Marrin/Dennhov/MASTER/VOTEC/LH sytiPortumias/Almaida Andrada tyt
equenc	y alphab	etical stati	slics filenames	notes		

LICO	gion_wor	diacias				
File	Edit	View	Compute	Settings	Windows	Help
N						text file
1	Users'Ma	rcio/Drop	box/MASTE	RIVOTECU	_txt/English/(0	Current Issues in Linguistic Theory 237) Barry J. Blake, Kate Burridge, Jo
2	C:\\	Jsers'Mar	cio/Dropbo	MASTERIN	DTEC/LH_bat/E	inglish/HOW LONG WAS THE NINETEENTH CENTURY_04_04_2012.txt
3pr	cio/Dropb	ox MAST	ERIVOTEC	LH_txt\Englis	NEnk jens Mo	ogensen, INTERNATIONAL CONFERENCE ON HISTORICAL L-Historical
4	C	Users'M	arcio/Dropbe	WMASTER!	OTECILH_bd	English\degraff2009language_acquisition_in_creolization_14_02_2013.txt
5		C:Users	s'Marcio'Dri	opbox/MASTE	RWOTECILH	txt/English/degraff2005fallacy of creole exceptionalism_14_02_2013.txt
60	rs'Marcio	Dropbax	MASTERW	OTECILH_bd	English/David	Potter, Richard Talbert, editors-American Journal of Philology-The Johns
7			C:WH	ers/Marcio/D	ropbox/MASTE	R/VOTEC/LH_txt/English/1992_Portuguese compounds_16_02_2013.txt
0.0	relliarris	Durchave	MASTERA	OTECUH M	English Jame	a B. Greensuch, Frederic D. Allen, John Williams White-Hanard Studies

Figura 2. Visão parcial dos arquivos que fazem parte dos corpora de Linguística Histórica.

O primeiro passo tomado para o tratamento dos *corpora* é a utilização de três ferramentas do WST: o gerador de listas de palavras, o extrator de palavras-chave e o concordanciador (VIANA, 2011, p. 43). O gerador de lista de palavras levanta as formas diferentes de palavras no *corpus* com suas frequências a partir de uma seleção de textos estabelecida pelo pesquisador.

File	Edit View	Compute S	ettings	Window	s F	elp	File	Edit	View C	ompute Se	ttings	Windows	н	elp
N		Word	Freq.	%	Texts	%.emmat Set	N	1		Word	Freq	%	Texts	% emmas Se
1		DE	22,242	4.15	33	100.00	1			THE	29,241	5.60	8	100.00
2		A	14,583	2.72	33	100.00	2			OF	17.846	3.42	8	100.00
3		E	11,448	2.13	33	100.00	3			IN	13.829	2.65	8	100.00
4		0	11,186	2.09	33	100.00	4			AND	11.046	2.12	8	100.00
5		QUE	10,692	1.99	33	100.00	5			A	9.159	1.76	8	100.00
6		DO	8,056	1.50	33	100.00	6			то	8,410	1.61	8	100.00
7		EM	7,017	1.31	33	100.00	7			IS	5.950	1.14	8	100.00
8		DA	6,486	1.21	33	100.00	8			THAT	5,001	0.96	8	100.00
9		SE	6,151	1.15	33	100.00	9			AS	4,456	0.85	8	100.00
10		NO	4,038	0.75	33	100.00	10			1	3,207	0.61	8	100.00
11		OU	3,852	0.72	33	100.00	11			FOR	3.042	0.58	8	100.00
12		COM	3,494	0.65	33	100.00	12			WITH	2,910	0.56	8	100.00
13		UM	3,446	0.64	33	100.00	13			BE	2,856	0.55	8	100.00
14		OS	3,426	0.64	33	100.00	14			п	2,825	0.54	8	100.00
15		UMA	3,355	0.63	33	100.00	15			THIS	2.809	0.54	8	100.00
16		É	3,337	0.62	33	100.00	16			BY	2,551	0.49	8	100.00
17		COMO	3,208	0.60	33	100.00	17			ON	2,324	0.45	8	100.00
18		POR	3,148	0.59	33	100.00	18			ARE	2,301	0.44	8	100.00
19		PARA	3,101	0.58	33	100.00	19			FROM	2,220	0.43	8	100.00
20	1	NA	2,853	0.53	33	100.00	20			NOT	2,128	0.41	8	100.00
equency	alphabetical s	tatistics filenames no	tes				frequency	alphabe	tical statistics	filenames no	tes			

Figura 3. Lista de palavras em ordem de frequência do *corpus* de Linguística Histórica (visão parcial).

Além do gerador de lista de palavras, usamos também o extrator de palavras-chave para levantar os termos que existem no *corpus* de estudo e compará-los com palavras de um *corpus* de referência. Nesse caso, usamos o *corpus* de referência BNC e ANC, de 122 milhões de palavras, para a língua inglesa, e o Banco do Português, 689 milhões de palavras, para o português. Após essa comparação, o WST exibe as palavraschaves positivas, aquelas que servem para identificar o *corpus* a ser descrito. Isto significa que as palavras que aparecem na lista de palavraschave são aquelas que são mais recorrentes no *corpus* de estudo do que no *corpus* de referência, o que é chamado de chavicidade. "A chavicidade reporta o resultado de um procedimento estatístico pelo qual a ferra-

menta levanta o quão importante cada palavra-chave positiva é para o *corpus* de pesquisa em relação ao de referência" (VIANA, 2011, p. 64).

File	Edit View	Compute S	ettings	Window	9 Hel	Y	File	Edit Vie	w Compute	Settings	Windows	Help	11 10-11
N		Key word	Freq.	%	Fred	C. %/nest Phmas3e	N		Key word	Freq	%D. Fred	RC. %eynest F	pmma: Se
1		LANGUAGES	3.176	0.25	3.919	8.489000d 3141	1	5	LINGU/	1,964	0.3750,049	11,5100000	
2	1	VERE	1 192	0.14	198	5 443000dr7321	2		PORTUGUES	1,494	0.2843,447	379.5 00000	
3		INGUISTICS	489	0.09	895	3.550000d	3		LATIN	751	0.14 3,370	859.090000	
4	AMANAT	CALIZATION	201	0.06		3 12 0000	4		SUFIXC	513	0.10 510	928.8700000	
-	Common I	TRANSITIVE	240	0.00		3,12,00000	0		VERBC	728	0.14 6,761	670.9900000	
D	-	TRANSITIVE	242	0.05	05	2,2690000	0		FOTORC	1,200	0.2484,000	0.01207.8900000	
6	CH	EOLIZATION	194	0.04	2	2,09ep000d	0		LINGOAS	662	0.10.0.055	643.1400000 604.7000000	
7		GENITIVE	182	0.03	30	1,8100000	0		DALAVRAS	1 020	0.1003.540	0.01372 3.00000	
8	CONS	STRUCTIONS	233	0.04	376	1,7370000	10		FILOLOGIA	206	0.06 540	157.6(00000	
9		VOWEL	235	0.05	465	1,6770000	11		SUBSTANTIVO	320	0.06 1.79	886 900000	
10		CHANGE	887	0.177	7.709	0.031.5620000	12		EFEITC	1.096	0 2056 276	0.02882.5300000	
11		LATIN	368	0.07	904	1.5440000	13		PERIFRASE	224	0.04 233	578.2300000	
12		CLITIC	130	0.02	4	1 38 0000	14		INFINITIVO	245	0.05 564	520.9 00000	
13		CDEAKEDS	201	0.06	0.74	1 3500000	15		VERBOS	300	0.06 2,673	3 360.1300000	
13		SPEAKERS	501	0.00.	070	1,358/0000	16		SÉCULO	868	0.1626,727	0.02246.3000000	
14		FORMS	512	0.10.	2,971	0.011,3480000	17		GALEGO	220	0.04 533	244.900000	
15		PRONOUNS	175	0.03	278	1,3090000	18		OCORRÊNCIAS	361	0.07 9,730	076.3400000	
16		DATIVE	127	0.02	14	1,2950000	19		ETIMOLOGIA	227	0.04 1,167	017.1900000	
17	CON	ISTRUCTION	404	0.087	497	1,2800000	20	·	GRAMATICA	272	0.05 5,401	724.400000	
18	1 /	ACCUSATIVE	124	0.02	12	1,2730000	21	GRAM	MATICALIZAÇÃO	153	0.03 300	614.4 00000	
19		NOMINATIVE	118	0.02	14	1.199000d	22		VOGAL	174	0.03 1,155	464.5200000	
20	1 3	REANAL VSIS	114	0.02	27	1 10700000	23		VERBAL	295	0.0612,765	429.2 00000	
24		MEANING	200	0.086	tron a	1.0700000	24		DERIVAÇÃO	195	0.04 2,527	395.6000000	
21		IVIE/ANTIACI	300	0.001	000	1.07.00000	20		VARIANTES	192	0.04 4,040	195.8000000	
22		SYNTAX	1/6	0.03	668	1,0630000	20		LIETÓDICA	223	0.04 7,812	170.5500000	
23		DIALECTS	149	0.03	349	1,02200000	20		DICIONÁRIO	340	0.0030,014	004 7 00000	
Na 100	t linka chi	sters flenames note	s source	text			KWA F	ot finks	chusters filenames n	sten nour	ce text	094.7 000000	
Ty	0.0.10			-			443 7	vpe-in					

Figura 4. Lista de palavras-chave do corpus de linguística histórica (visão parcial).

A terceira ferramenta utilizada é o concordanciador, ferramenta que permite a análise de uma palavra em seu cotexto<sup>99</sup>. Os resultados dessa ferramenta são dispostos em linhas de concordância, fragmentos dos textos que são exibidos, a partir de uma palavra selecionada pelo usuário.

N	Concordance	Sel ad Vord #	ent, #	t. Post	ParaPara. ad	Sec	Sect. File	%
1	é outra coisa. Para ela, a língua é uma expressão da	3,573	265	33%	0 3%	0	3%4S_18_02_2013	3%
2	Neto (1986:18) declara: A língua é um produto social, é	3,283	104	78%	0 37%	0	37% as_04_04_2012	37%
3	Nossa língua, como toda língua, 'é um rastro de velhos	1,165	51	16%	0 43%	0	43%NA FORMAÇÃO	439
4	estática de língua – a língua é em si mesma um	3,152	99	34%	0 36%	0	36% as_04_04_2012	35%
5	em que a aquisição da nova língua é feita por indivíduos	1,566	54	65%	0 20%	0	20%IA_10_03_2012	209
6	(2005: 24) que diz que "a língua é uma realidade	4,911	155	73%	0 77%	0	77% olonial brasileira	769
7	preceito saussuriano de que a língua é forma não substância	19,763	985	35%	0 33%	0	33%evista Philologus	339
8	da gênese e evolução da língua é um "vôo cego", se tais	368	11	32%	0 9%	0	9%Arqueologia das	99
9	O léxico de uma língua é constituído por	204	7	38%	0 7%	0	7%jia_16_02_2013	69
10	perspectiva da crítica textual, a língua é apenas um dos	1,268	65	14%	0 34%	0	34%lologia Textual e	329
11	língua e que o estudo da língua é objeto da Lingüística,	247	11	62%	0 7%	0	7%lologia Textual e	79
12	tanto, sabendo-se que cada língua é fruto da conjunção	51	1	10%	0 3%	0	3% IMPORTÂNCIA	39
13	lingüística. E, nesse sentido, a língua é mais do que uma	3,119	98	30%	0 35%	0	35%as_04_04_2012	35%
14	, acha que a sintaxe de uma língua é atin- gida primeiro que	43,031	2,395	32%	0 73%	0	73%avista Philologus	739
15	de regras sintáticas de uma língua é também relativamente	4,013	277	95%	0 3%	0	3%45_18_02_2013	39
16	a coletividade as repete. A língua é eminentemente	8,867	409	11%	78 33%	0	38%es_18_02_2013	389
17	particularidades locais de sua língua: é a conseqüência dos	8,947	414	60%	78 34%	0	38%es_18_02_2013	399
18	dos fonemas – os sons da língua – é impressionante	39,873	1,595	56%	0 73%	0	73%avista Philologus	739
19	com o historiógrafo da língua é um esquecimento de	8,352	283	27%	0 94%	0	94%as_04_04_2012	93%
20	períodos antigos de qualquer língua é algo difícil e que	401	24	23%	0 5%	0	5%Jma perspectiva	69
21	possibilita visualizar como a língua é usada por uma	471	26	74%	0 6%	0	6%Jma perspectiva	79
22	autoritária de uma língua é um instrumento que	12,961	826	50%	78 80%	0	55%es_18_02_2013	55%
23	partindo do princípio de que a língua é expressão de cultura,	311	20	65%	0 10%	0	10%ca_19_03_2011	109

Figura 5. Linha de concordância para o termo língua seguida do colocado é (vista parcial).

<sup>&</sup>lt;sup>99</sup> Cotexto em Linguística de Corpus refere-se ao ambiente linguístico, isto é, os itens lexicais que estão à direita e à esquerda de uma dada palavra. (VIANA, 2011, pg. 71).

Círculo Fluminense de Estudos Filológicos e Linguísticos

#### 5. VoTec

Feitos os levantamentos de corpora, as listas de palavras e de palavras-chave, os dados serão inseridos na plataforma do VoTec para construção da macro e microestrutura do dicionário. Para definirmos essa plataforma, "VoTec é uma ferramenta que se vale de corpora técnicos para a construção de seus verbetes e de um banco de dados (ambos exaustivamente descritos) para o seu funcionamento". (FROMM, 2007, pg. 8). É uma ferramenta que oferece várias formas de visualização (normal e descritiva), e as consultas podem ser nos módulos total, tradutor e modular. A visualização normal é o formato que segue o padrão dos dicionários impressos. Diferentemente dessa, a descritiva apresenta os dados de forma hierárquica e detalhadamente. Nas opções de consulta, a de módulo total disponibiliza todos os campos do banco de dados, sendo de caráter lexicográfico; em seguida, a de módulo tradutor permite ao consulente acessar as informações mais frequentes usadas por tradutores, tais quais: área de especialidade, tradução, sinônimos, definição, entre outros. Finalmente, a consulta modular exibe a microestrutura de acordo com a busca do consulente, acelerando sua busca pela seleção mais específica de dados. Além dessas opções de visualização, o site disponibiliza ao usuário acesso para consultas externas.

Firefox Vocabulário Técr	sico Online +	
🔶 🔶 🔐 😁 pos.voteconlin	ne.com.br	合 マ C 🛛 🔀 - Google
CONFCTE: • Saisir une forme:	✓ D Chercher catégorie: Synonymie	
Vocabulário Técnico Online	•	Tela Cheia   English  Ajuda
Linguística • Linguística Busci	a Teórica 🔹 Lexicología 🔹	
Tipos de Exibição Normal Descritiva	▼ Português Por Favor entre com o termo de Busca. ▼ English	
Tipos de Consulta Total Tradutor Modular	Please Enter a search term.	
Consultas Externas Corpus NILC Google Answers.com Wikipedia CORTEC		
	01/04/2013 05:07 © 20	07 Guilberme Fromm - ICMC Ir

Figura 6. Visualização da página inicial do VoTec.

A abordagem do banco de dados adotada para o *VoTec* é a lexicográfica, na qual os campos usados para a criação do verbete do dicionário constam na ficha terminológica. O acesso do pesquisador ao banco de dados se faz por meio do cadastro de um usuário e de uma senha pessoal. Na próxima tela ele pode clicar na opção Novo Termo e iniciar o processo de registro de dados na plataforma.

oulário Técnico Online +		
s.voteconline.com.br/admin/termo_1.php	 	😫 - Google 🔎 🦆 🏫
me:	D Chercher catégorie: Synonymie	
Vocabulário Técnico O	nline	Tela Cheia   English
Novo Termo		
	Voltar ao painel	
Passo 1		
Termo		
Língua Escolha uma língua:	Português -	
Ontologia Grande Área:	Escolha uma Área 👻	
	Próximo Passo (Contextos)	
	01	/04/2013 06:10 © 2007 FFLCH - ICMC Jr.

Figura 7. Visualização do passo 1 para cadastro de termos no VoTec.

Nessa primeira página, o pesquisador deve inserir o termo a ser registrado no item Termo, após deve escolher a língua do referido termo (português ou inglês) e, finalmente, escolher sua ontologia desde a grande área, nesse caso linguística, até a subárea: linguística histórica. Selecionados esses dados, o próximo passo será o cadastro de contextos.

Vocabulário Técnico O	nline	Tela Cheia   Engl
	Voltar ao Painel Próximo Passo	
Novo Contexto		
Passo 2		
Dados do Contexto		
Exemplo*:		
Conseitet		
Fonte*:	OPDF	- Cadastrar Nova
Fonte		
Data de Coleta*:	(dia/mês/ano ex.: 18/03/2007)	
	Salvar	

Figura 8. Visualização da página Cadastro de contextos para língua (visão parcial).

Nesta etapa do procedimento, no espaço chamado Exemplo\*, registramos o contexto no qual o termo se insere, dados esses provenientes do concordanciador do WST. A partir desse exemplo, faz-se um recorte de possíveis contextos definitórios ou explicativos<sup>100</sup> e o transpomos para

<sup>&</sup>lt;sup>100</sup> Segundo Aubert (1996, p. 66-67), "os contextos explicativos apresentam alguns traços conceptuais pertinentes específicos do termo sob observação, frequentemente relativos à materialidade, finalidade, funcionamento e similares. [...] Talvez mais desejáveis, mas certamente menos encontradiços, os contextos definitórios proporcionam um conjunto completo dos traços conceptuais distintivos do termo.

o campo Conceito\*. Abaixo desse campo, há o campo Data de coleta\* onde inserimos a data em que o texto do *corpus* foi coletado, registrado no documento em formato txt. Na parte inferior dessa página, há o registro dos contextos cadastrados. À direita desse campo há as opções editar e excluir que podem ser usadas para alteração ou exclusão de dados, de acordo com a avaliação do pesquisador.

.....

Exemplo	Conceito	Fonte	Ações
Para Mattos e Silva, um estudo que trabalhe com a sincronia contemporânea não exige tanto quanto o trabalho que objetive merguihar, por exemplo, no português arcaico. Por essa razão, os aspectos sociais são tão importantes, visto que possibilita visualizar como a língua é usada por uma determinada comunidade.	aspectos sociais possibilitam visualizar como a língua é usada por uma determinada comunidade.	PDF 04/04/2012	editar - excluir
Embora encontremos em Ferdinand de Saussure esclarecimentos para aqueles que postulam uma concepção estática de língua - a língua é em si mesma um sistema imutável -, podemos observar que, por essa perspectiva, a língua que não muda não é a mesma exposta aos fatores externos, ou seja, aquela constituída pela historicidade e concreitzada pela interação humana.	constituída pela historicidade e concretizada pela interação humana.	PDF 04/04/2012	editar - excluir
Nossa língua, como toda língua, 'é um rastro de velhos mistérios', no dizer de Guimarães Rosa, é um produto histórico, cujas marcas hão de ser perquendias através dos tempos. O túnel do tempo, expressão consagrada por Tarallo, levar-nos-á aso mistérios tão indevassíveis, às raízes - qual a diferença entre radical e raíz? - às respostas a tantos porquês grotescamente dosados como regras e/ ou exceções gramaticais.	é um produto histórico	PDF 17/02/2013	editar - excluir
A língua é eminentemente mutável no tempo e o seu movimento de mudanca tem o caráter de uma evolucão.	mutável no tempo e o seu movimento de mudanca tem o caráter de uma	PDF 18/02/2013	editar - exclui

Figura 9. Visualização da página Cadastro de contextos para o termo língua (visão parcial).

O acesso ao próximo procedimento é possível ao clicar-se sobre o ícone Próximo passo no canto superior direito da página.

<u></u>	oca	bulário Técnico Online		Tela Cheia   English
l	Ге	rmo: língua		
Γ		Passo Anter	ior Salvar Sair sem salvar	
6	Con Cor	textos Conceito Final / Definição Itextos		
H		Exemplo	Conceito	Fonte
	1	A língua é um produto social, é uma atividade do espírito humano. Não é, assim Independente da vontade do homem, porque o homem não é uma folha seca ao sabor dos ventos veementes de uma fatuldade desconhecida e cega. Não está obrigado a proseguir na su trajetória, de acordo com leis determinadas, porque as línguas seguem o destino dos que a falam, são o que delas fazem as sociedades que as empregam.	um produto social, é uma atlvidade do espírito humano	PDF 04/04/0
ļ	Dad Dad	os / Traços Distintivos / Semântica / Tern	no Equivalente 🖌 Termos Remissivos 🧹 Infor	mações Enciclopédicas
		Ontologia: Linguística > Linguística T	eórica > Linguística Histórica	Â

Figura 10. Visualização da página de construção de microestrutura da página do VoTec.

Nesse momento do preenchimento da ficha terminográfica, o pesquisador preencherá as abas (i) Dados, (ii) Traços distintivos, (iii) Semântica, (iv) Termo equivalente, (v) Termos remissivos, (vi) Informações enciclopédicas e (vii) Conceito Final/Definição. Abaixo segue uma descrição de cada um deles e o que deve ser inserido.

(i) **Dados:** na parte superior desta aba temos os dados ontológicos do termo. Nesse caso: Linguística> Linguística Teórica> Linguística Histórica. Logo abaixo, há o campo categoria gramatical, onde selecionamos a opcão substantivo para o termo *língua*, no canto esquerdo. À direita há o campo Número onde selecionamos a opção singular, plural ou dual (como *lápis*, por exemplo). À esquerda, abaixo da categoria gramatical, há o campo Gênero, onde selecionamos as opções masculino, feminino ou neutro. À direita há o campo Sigla/Acrônimo, muito usados para nomes de instituições ou termos da informática, tais quais CPU, CD-Rom, etc. Em seguida abaixo, há a Entrada por extenso - o significado de uma abreviação ou acrônimo, Variações morfossintáticas como center e centre no inglês, e Acepção nº- campo usado para termos polissêmicos, como memória em informática (distinção feita para software e hardware). Logo abaixo desses itens há o registro de dados em relação ao corpus: dispostos lado a lado, eles são (a) Posição na ordem de frequência e (b) Nº de ocorrências do termo, dados provenientes da lista de palavras do WST.

(	Dados Traços Distintiv	os Semântica	Termo Equivalente	Termos Rer	missivos	Informações Encido	opédicas
[	Dados						
	Ontologia:	Linguística > Lingu	ística Teórica > Linguíst	tica Histórica			
	Categoria Gramatical:	Substantivo	•	Número:	Singular 👻	]	
	Gênero:	Masculino 👻	Sig	a/Acrônimo:			
	Entrada por Extenso:	Língua					
	Var. Morfossintáticas:						
J	Acepção Nº:						
	Córpus						
	Posição na Ordem de	26	N° de	Ocorrências d	0 1964		
	rrequencia			termo			

Figura 11. Visualização da aba Dados da página do VoTec.

(ii) **Traços distintivos:** nesse campo da ficha terminográfica são inseridos as informações extraídas dos contextos/exemplos que servirão à construção da definição final. Os dados serão organizados em colunas, nas quais os diferentes traços distintivos serão listados. Cada linha repre-

700 Revista Philologus, Ano 19, N° 55. Rio de Janeiro: CiFEFiL, jan./abr. 2013 - Suplemento.

senta os dados provenientes de cada contexto/exemplo. Todas as vezes que os traços pertencerem ao mesmo campo semântico, eles serão listados nas mesmas colunas.

lad Tra No	os Traços Dis ços Distintivos — ava Coluna — N	āo use aspas du	mântica (Te	ermo Equivale	nte Termos	Remissivos	Informações End	ticlopé
	A	В	с	D	E	F	G	
1	produto so	atividade						
2			produto hi					
3	concretiza		constituíd					
4		necessidad		realidade	mudança in	níveis ort		
5			processo d		mutável no			
6	aspectos s							
7	comunidade				variações		organismo	

Figura 12. Visualização da aba Traços Distintivos da página do VoTec.

(iii) Semântica: indicamos se o termo é dicionarizado e se a definição dicionarizada é correspondente total ou parcialmente à definição em construção - o que geralmente será parcial por questões de direito autoral e só serve como norte para o pesquisador, não sendo disponibilizadas na página de visualização do programa. Em seguida seleciona-se o dicioná-rio a partir do qual provém a definição. No campo abaixo, insere-se a definição de acordo com a fonte mencionada. Isto feito, classificamos os exemplos dentro dos campos da hiperonímia, hiponímia, co-hiponímia, sinonímia e antonímia, se possível for. Se necessário, o pesquisador pode deixar alguma nota que será verificada pelo administrador do sistema.

Termo Dicionarizado?	9	Definições Coincidentes? Parcial -		1
Fonte:	Dicionário eletrônio	co Houaiss da língua portuguesa	•	
efinição Dicionarizada:	sistema de represer combinam em frase como principal mei	ntação constituído por palavras e por regras que es que os indivíduos de uma comunidade linguísti o de comunicação e de expressão, falado ou escr	as ca usam ito.	B
Hiperônimo de:	palavras, regras, fra	ases, nível ortográfico, fonético-fonológico, sinta	itico, lexical, r	
Hipônimo de:	produto social, ativ	idade humana, produto histórico, interação hum	ana, nesistema	
Co-hipônimo de:				
Sinônimo(s):		Antônimo(s):		
Notas:				

Figura 13. Visualização da aba Semântica da página do VoTec.

Círculo Fluminense de Estudos Filológicos e Linguísticos

(iv) Termo equivalente: nessa aba, o sistema busca o termo equivalente na outra língua (inglês) para associá-lo ao termo em português. Esse processo só é possível se o termo foi registrado dentro das mesmas características ontológicas do termo que está sendo registrado.



Figura 14. Visualização da aba Termo Equivalente da página do VoTec.

(v) Termos remissivos: nessa aba, podem ser resgatados os termos que se relacionam dentro do mesmo campo semântico tais quais: sinônimos, antônimos, hipônimos, co-hipônimos e hiperônimos (FROMM, 2007, p. 105). Essa relação e ligação só é possível se os termos já tiverem sido cadastrados e aprovados pelo administrador do sistema.

ermos Remissivos		
	Os termos remissivos devem ser termos cadastrado em português e com os respec já aprovados pelo administrador.	tivos cadastro
Buscar Termo Remissivo:		Buscar

Figura 15. Visualização da aba Termos remissivos da página do VoTec.

(vi) Informações enciclopédicas: adiciona-se as definições provenientes de uma fonte enciclopédica, o artigo, a fonte, o *link* da informação e o tipo de fonte, não disponível na versão atual. Normalmente as informações desta aba são fornecidas, novamente por uma questão de direitos autorais, pela Wikipédia – as mesmas podem ser visualizadas na página de consulta do programa.

Definição:						^
Artigo:						L
Franker						L
ronte:						н
Link:		Tipo: Ir	magem 👻			L
(	Adicionar					Ŀ
(	Adicionar					
(	Adicionar					
Definição	Adicionar	Artigo	Fonte	Link	Apagar	
Definição Língua natural (língua hur	Adicionar nana ou somente língua) é qualquer linguagem desenvolvida	Artigo língua	Fonte Wikipédia	Link	Apagar	E
Definição .íngua natural (língua hur aturalmente pelo ser hui calidade inata para a lin	Adicionar nana ou somente língua) é qualquer linguagem desenvolvida nano, de forma não premeditada, como resultado da guagem possuida pelo intelecto humano. Vários exemplos	Artigo língua	Fonte Wikipédia	Link	Apagar	

Figura 16. Visualização da aba Informações Enciclopédicas da página do VoTec.

(vii) Conceito Final/Definição: nesse campo, o pesquisador formula as definições para o termo selecionado em duas janelas denominadas Conceito Final e Definição. Na primeira, o pesquisador insere os dados para uma definição prévia do termo, construída a partir dos traços distintivos. A segunda janela, a Definição, já com um texto mais encaixado dentro de parâmetros do projeto de pesquisa em questão, é a definição final que será visualizada pelos consulentes ao acessar a página do dicionário na Internet.



Figura 17. Visualização da aba Conceito Final da página do VoTec.

Preenchidos todos os campos, o pesquisador deve salvar as informações antes de sair do sistema. Essas informações serão enviadas ao administrador do sistema que, após avaliação, as disponibilizará para o acesso público. Do contrário elas poderão ser reeditadas, a pedido do administrador, se necessário for.

Concluída a análise e aprovação do administrador, o termo será visualizado em português e inglês na página do Votec. Acessível ao público em geral e com visualizações alternativas de acordo com a necessidade do consulente.

Tinos do Evibição	▼ Português		
Normal	Voltar ao resultado da busca		
Corpus NILC Corpus NILC Corpus NILC Corpus NILC Coglia Corpus Co	Ingua. Linguistica Histórica. Lingua. s.1.s. atividade do espírito humano, que responde às necessidades socio- interativas de uma comunidade linguistica; produto socio-histórico, mutável no tempo caracterizado por um processo dinàmico, gradual es coerente de mutações nos niveis ortográfico, honetito-fonológico, sintatico, lexical, mortológico e discursivo. Ex.: à lingua é um produto social, é uma atividade do espírito humano. Não é, assim, independente da vontade do homem, porque o homem não é uma folha seca ao sabor. Não e, assim, nicependente ta vontade do homem, porque o homem não é uma folha seca ao sabor. Não e, assim, nicependente da vontade do homem, porque o homem não é uma folha seca ao sabor com leis determinadas, porque as linguas seguem o destino dos que à falam, são o que delsa fazem as a reacio nom testo determinadas, porque as linguas seguem o destino dos que a falam, são o que delsa fazem as a reacio humana, necessidades socio-interativas, organismo vivo. <i>Hiperialmo de</i> : palavras, regras, frases, nivel ortográfico, nordetico. Fonológico, sintático, texical, morfológico, discursivo, variações lingüíscas, córpus: Posição na Ordem de Freqüência: (20): Nº de Ocorrências do termo: (1964). Informações Enciclopódicas: Lingua natura (língua humano au osmente lingua) e qualquer linguagem desenvolvida naturalmente pelo ser humano, de forma não premeditada, como resultado da facilidade intat para a linguagem possuída pelo intelecto humano. Vários exemplos podem ser dados Exer. língua - <u>Mikoréal</u>		
	The second s		
	▼ English		
	Go back to search results		
	language. Historical Linguistics. n.m.Mf.s. a system that bears human speech which goes under progression and modifications creating new linguistic forms; grouped in families understood by analogy of factors that work together; can be divided in two main classes: monosyllabic, that is isolating or radical, and polysyllabic or inflectional. Ex: Labovian linguistics is thus of crucial importance for understanding some of the ways in which languages, at both the I- and E-language levels, are constantly and gradually evolving, and why 'fully convergent grammars' (cf. Hale 1998: 1) are an unlikely (i.e., 'abornary') outcome of language acquisition (see 52.3 below). Hyponym of; factors, families. Hypernym of; grammars, linguistics forms, human speech, monosyllabic, polysyllabic. Corpus: Frequency order position: (1); Ferm number of occurrences: (3919). Encyclopedic information: Language is the human capacity for acquiring and using complex systems of communication, and a language is any specific example of such asystem. The scientific study of language is called linguistics. Any estimate of the precise number of languages is called linguistics. Any estimate of the precise number of languages in them: Language - <u>Wikipédia</u>		

Figura 18. VoTec - Visualização do termo língua na área de Linguística Histórica.

#### 6. Considerações finais

Este artigo apresentou e detalhou os passos metodológicos adotados para a confecção de um dicionário bilíngue português-inglês na área de linguística histórica, direcionado para o público da área de linguística, tradução e outros profissionais que tenham interesse na área. Detalhamos como usar a metodologia da linguística de *corpus* para execução da coleta e tratamento dos *corpora* portanto, valemo-nos da ferramenta WST para leitura e análise dos dados linguísticos, confecção das listas de palavras, de palavras-chave e do concordanciador. Após tais procedimentos, explicitamos como se faz o lançamento de dados específicos na plataforma do *VoTec*, o preenchimento das fichas terminográficas *on-line* e a fase final para que a obra terminográfica seja disponibilizada na Internet. Esperamos que este trabalho seja útil para fortalecer o estudo da linguística histórica, filologia e etimologia nos países lusófonos e que sirva para a difusão e expansão da língua portuguesa, já que o acesso a obras em língua inglesa pode ser facilitado através desse trabalho.

Círculo Fluminense de Estudos Filológicos e Linguísticos

#### REFERÊNCIAS BIBLIOGRÁFICAS

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

COSERIU, E. *Sincronia, diacronia e história*. Trad.: C. A. da Fonseca e M. Ferreira. Rio de Janeiro: Presença; São Paulo: Universidade de São Paulo, 1979.

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

\_\_\_\_\_\_. Terminografia, sociolinguística, linguística de corpus, tradução, tecnologia da informação: convergências. In: MOLLICA, Maria Cecilia; GONZALES, Marcos. (Orgs.). *Linguística e ciência da informação*: diálogos possíveis. Curitiba: Appris, 2012, p. 141-158.

SCOTT, M. *WordSmith Tools*. Versão 5. Disponível em: <a href="http://lexically.net/wordsmith/version5">http://lexically.net/wordsmith/version5</a>>. Acesso em: 23-03-2013.

SILVA, Rosa Virgínia Mattos e. *Caminhos da linguística histórica – "ouvir o inaudível"*. São Paulo: Parábola, 2008.

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

VIARO, Mário Eduardo. Etimologia. São Paulo: Contexto, 2011.