Reference corpora that can make a difference

Stella E. O. Tagnin – USP

CILC 2013

Alicante, Spain

Outline

- What is a reference corpus?
- Different types of reference corpora
- RC in Terminology
- RC in Literature
- RC in Translation
- Final comments

What is a reference corpus?

Content-based concept:

- usu. of general language :
  - BNC, Russian National Corpus, the Reference Corpus for Contemporary Portuguese, Corpus del Español, or the COCA
  - mostly "national corpora".

- Also corpus of specific genres / specialized language:
  - Reference Corpus for Web Genres
  - DeRiK: A German Reference Corpus of Computer-Mediated Communication.

Function-based concept:

- used for "comparative purposes" (Scott 2010)

→ to elicit the specific vocabulary of the corpus being investigated.

Corpus for comparison

To elicit specificities of study corpus = **keywords**

study corpus

x

reference corpus

Comparison cancels out similar vocabulary

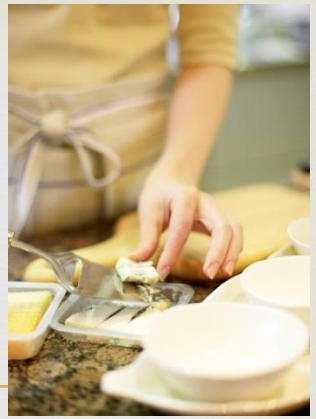
statistically

Types of reference corpora

- General language corpora
- Specialized language corpora
- Customized reference corpora
- RC usually 3 to 5 times larger than study corpus

... but not always...

Same size corpora



Cooking recipes

Portugal  vs Brazil 

UK  vs USA 

- 10 cooking recipes in 11 categories = circa 24.000 words each corpus

Categories covered

- Appetizers
- Soups
- Entrées: meat
- Entrées: poultry
- Entrées: fish
- Entrées: pasta
- Side dishes
- Salads
- Desserts
- Cakes and Pies
- Breads

Processing the corpus

	Tokens	Types	T/T ratio
BP	24.816	2.066	8,33
EP	**17.661**	**1.942**	**11**
AE	26.990	2.287	8,47
BE	24.587	2.150	8,74

Keywords

- UK  vs USA 
- Positive keywords = UK
- Negative keywords = USA

BritishEnglish

Positive

N	WORD	FREQ.	TOTIOB.LST %	FREQ.	TOTIOA LST %	KEYNESS	P
1	THE	1.689	6,86	1.164	4,31	160,3	0,000000
2	OZ	178	0,72	35	0,13	119,0	0,000000
3	TBSP	132	0,54	20	0,07	103,2	0,000000
4	METHOD	61	0,25	3	0,01	70,0	0,000000
5	GAS	47	0,19	3	0,01	50,8	0,000000
6	PINT	42	0,17	2		48,5	0,000000
7	LEAVE	47	0,19	4	0,01	46,8	0,000000
8	MINS	26	0,11	0		38,5	0,000000
9	SERVES	66	0,27	18	0,07	33,8	0,000000
10	CORNFLOUR	21	0,09	0		31,1	0,000000
11	LITTLE	60	0,24	17	0,06	29,6	0,000000
12	MARK	26	0,11	2		26,7	0,000000
13	PRAWNS	18	0,07	0		26,7	0,000000
14	TEASP	17	0,07	0		25,2	0,000001
15	COURGETTES	17	0,07	0		25,2	0,000001
16	FRYING	30	0,12	4	0,01	25,0	0,000001
17	FRY	46	0,19	12	0,04	24,5	0,000001
18	TIN	34	0,14	6	0,02	24,3	0,000001

American English

Negative

N	WORD	FREQ.	TOTIOB.LST %	FREQ.	TOTIOA LST %	KEYNESS	P
19	SHRIMP	*1*		25	0,09	25,4	0,000000
20	HAM	*0*		21	0,08	27,2	0,000000
21	SKILLET	*4*	*0,02*	38	0,14	28,8	0,000000
22	MEDIUM	*29*	*0,12*	93	0,34	29,8	0,000000
23	YIELD	*0*		24	0,09	31,1	0,000000
24	TABLESPOON	*18*	*0,07*	76	0,28	33,4	0,000000
25	POUND	*2*		37	0,14	35,2	0,000000
26	BROTH	*3*	*0,01*	41	0,15	35,7	0,000000
27	TB	*0*		30	0,11	38,9	0,000000
28	OUNCES	*0*		32	0,12	41,5	0,000000
29	DEGREES	*5*	*0,02*	56	0,21	45,5	0,000000
30	DIRECTIONS	*1*		42	0,16	46,5	0,000000
31	LET	*5*	*0,02*	58	0,21	47,7	0,000000
32	T	*0*		39	0,14	50,6	0,000000
33	TABLESPOONS	*20*	*0,08*	102	0,38	53,1	0,000000
34	SERVINGS	*3*	*0,01*	57	0,21	54,6	0,000000
35	TEASPOON	*19*	*0,08*	109	0,40	62,1	0,000000
36	CUPS	*5*	*0,02*	118	0,44	118,8	0,000000
37	CUP	*28*	*0,11*	284	1,05	222,4	0,000000

Lexical oppositions BE/AE

Broth	3	41
Stock	55	19
Skillet	4	38
Frying pan	21	0
Can	6	11
Tin	7	0
Tin (baking)	36	4
Cornstarch	0	13
Cornflour	21	0
Gas (mark)	47	1

Unexpected findings

- Geographical differences
  - **Europe**: precise measurements – dl, kg, oz, pint
  - **America**: bulk measurements – tsp, Tsp, cup

Terminology - Brazilian cuisine

- Brazilian recipes vs General language RC

→ Brazilian cuisine + general cooking terms

- Brazilian recipes vs General **Cooking** corpus

→ **Brazilian cuisine**

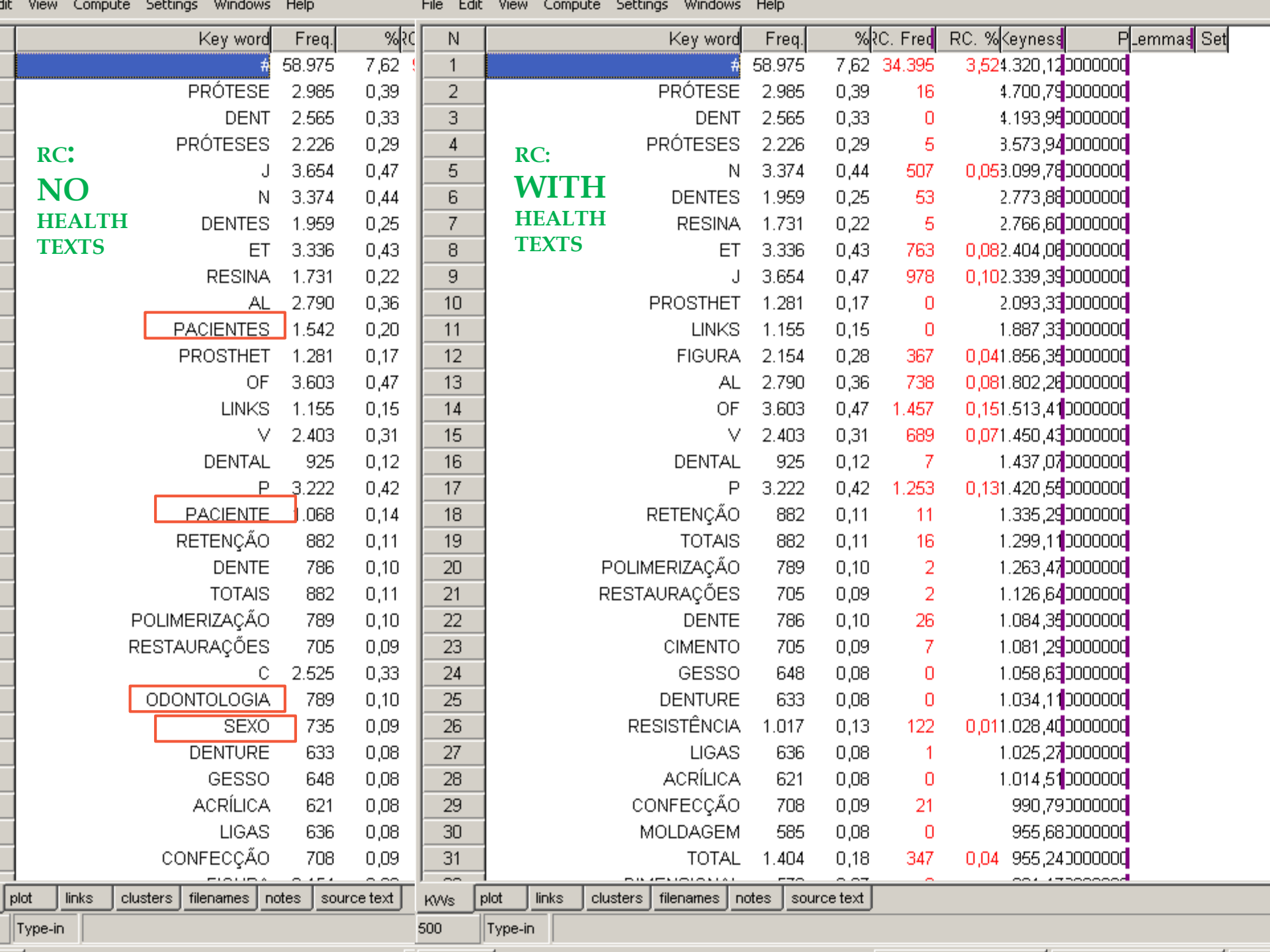
Terminology – Prostodontics

- Prostodontics vs General Language Corpus

→ Prostodontics terms + general medical terms

- Prostodontics vs **Health Sciences** corpus →

→ **Only Prostodontics terminology**



Key word	Freq.	%
#	58.975	7,62
PRÓTESE	2.985	0,39
DENT	2.565	0,33
PRÓTESES	2.226	0,29
J	3.654	0,47
N	3.374	0,44
DENTES	1.959	0,25
ET	3.336	0,43
RESINA	1.731	0,22
AL	2.790	0,36
PACIENTES	1.542	0,20
PROSTHET	1.281	0,17
OF	3.603	0,47
LINKS	1.155	0,15
V	2.403	0,31
DENTAL	925	0,12
P	3.222	0,42
PACIENTE	1.068	0,14
RETENÇÃO	882	0,11
DENTE	786	0,10
TOTAIS	882	0,11
POLIMERIZAÇÃO	789	0,10
RESTAURAÇÕES	705	0,09
C	2.525	0,33
ODONTOLOGIA	789	0,10
SEXO	735	0,09
DENTURE	633	0,08
GESSO	648	0,08
ACRÍLICA	621	0,08
LIGAS	636	0,08
CONFECÇÃO	708	0,09

N	Key word	Freq.	%	RC. Freq	RC. %	Keyness	P	Lemmas	Set
1	#	58.975	7,62	34.395	3,52	4.320,12	0,000000		
2	PRÓTESE	2.985	0,39	16		4.700,79	0,000000		
3	DENT	2.565	0,33	0		4.193,95	0,000000		
4	PRÓTESES	2.226	0,29	5		3.573,94	0,000000		
5	N	3.374	0,44	507	0,05	3.099,78	0,000000		
6	DENTES	1.959	0,25	53		2.773,88	0,000000		
7	RESINA	1.731	0,22	5		2.766,60	0,000000		
8	ET	3.336	0,43	763	0,08	2.404,08	0,000000		
9	J	3.654	0,47	978	0,10	2.339,39	0,000000		
10	PROSTHET	1.281	0,17	0		2.093,33	0,000000		
11	LINKS	1.155	0,15	0		1.887,33	0,000000		
12	FIGURA	2.154	0,28	367	0,04	1.856,38	0,000000		
13	AL	2.790	0,36	738	0,08	1.802,28	0,000000		
14	OF	3.603	0,47	1.457	0,15	1.513,41	0,000000		
15	V	2.403	0,31	689	0,07	1.450,43	0,000000		
16	DENTAL	925	0,12	7		1.437,07	0,000000		
17	P	3.222	0,42	1.253	0,13	1.420,55	0,000000		
18	RETENÇÃO	882	0,11	11		1.335,29	0,000000		
19	TOTAIS	882	0,11	16		1.299,11	0,000000		
20	POLIMERIZAÇÃO	789	0,10	2		1.263,47	0,000000		
21	RESTAURAÇÕES	705	0,09	2		1.126,64	0,000000		
22	DENTE	786	0,10	26		1.084,35	0,000000		
23	CIMENTO	705	0,09	7		1.081,29	0,000000		
24	GESSO	648	0,08	0		1.058,63	0,000000		
25	DENTURE	633	0,08	0		1.034,11	0,000000		
26	RESISTÊNCIA	1.017	0,13	122	0,01	1.028,40	0,000000		
27	LIGAS	636	0,08	1		1.025,27	0,000000		
28	ACRÍLICA	621	0,08	0		1.014,51	0,000000		
29	CONFECÇÃO	708	0,09	21		990,79	0,000000		
30	MOLDAGEM	585	0,08	0		955,68	0,000000		
31	TOTAL	1.404	0,18	347	0,04	955,24	0,000000		

plot | links | clusters | filenames | notes | source text

KWs | plot | links | clusters | filenames | notes | source text

Type-in

500 | Type-in

RC in Literature

(Gonçalves 2006)

James Joyce

(1882–1941)
67,940 words

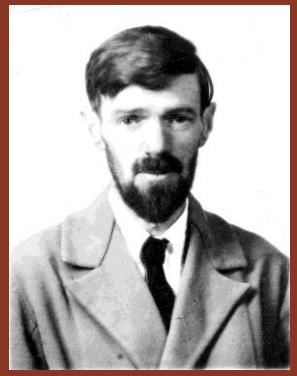
IRISH

Katherine Mansfield
(1888–1923)
79,845 words

212.591 WORDS
BRITISH

Virginia Woolf
(1882–1941)
53,659 words

D. H. Lawrence
(1885–1930)
79,087 words



Joyce revealed

- **City**: urban novel, citizen – *street, bottles, stout, whisky, drank, bar*
- **Positive keywords**: *Mr* (922,3), *he* (117,6), ***his*** (193,8), *friends* (67,5), **semantic field: music**

  tenor (45,4); *concert* (43,8); *artistes* (34,0); *concerts* (24,9); *baritone* (22,7); *clapping* (22,7); *song* (20,5); *opera* (17,0); *piano* (15,7); *music* (14,6); *sing* (14,4); *musical* (14,4); *artiste* (14,2); *accompanist* (14,2); *waltz* (13,8); *melody* (11,8); *singers* (11,3).

- **Negative keywords: nature, feelings, colors**
- **Music**: background **BUT ALSO** to distinguish characters (concordances).

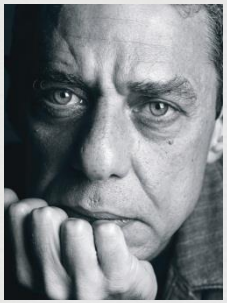
Joyce revealed

- SHE: keyness 376,6

- Criticics say woman is subservient in Joyce's stories
- 684 occurrences as subject, 207 distinct verbs

Concordances show:

- Volitional verbs – deliberate action: 99,4% woman is NOT in submissive or oppressive situations
- Intellectual verbs (mental processes): none indicate subservience
- Affective verbs (emotion) 99,03% do NOT point to oppression



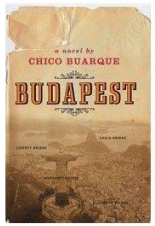
RC in Literature

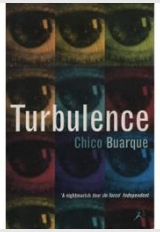
(Aguiar 2010)

Chico Buarque's novels

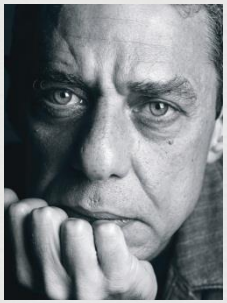
vs

General Language RC





N	Key word	Freq.	%	RC.	Freq.	Keyness
1	**EYES**	170	0.14	5	78.60
2	WHEN	456	0.38	62	77.26
3	**HAND**	180	0.15	9	68.39
4	**HAIR**	116	0.10	1	66.93
5	**MOUTH**	97	0.08	0	63.46
6	**FACE**	188	0.16	12	62.87
7	CASTANA	95	0.08	0	62.15
8	**HEAD**	160	0.13	8	60.79
9	MYSELF	87	0.07	0	56.92
10	GOT	124	0.10	4	55.74



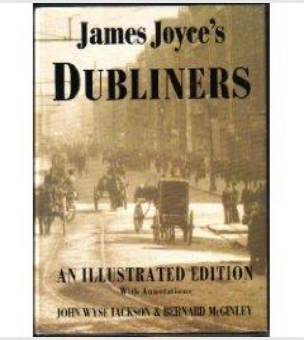
RC in Literature
(Aguiar 2010)

Typical of one novel or of the author?
Each novel vs other two → **body parts disappear**
Therefore: peculiar to all of Chico's 3 novels

Budapest vs **Turbulence** = peculiarities of Budapest– NO BODY PARTS

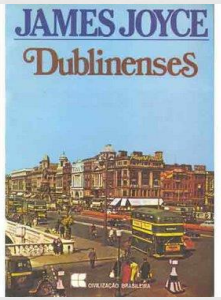
N	Key word	Freq.	%	RC. Freq.	Keyness
1	WAS	836	1.87	157	370.67
2	HAD	435	0.97	40	300.89
3	**VANDA**	138	0.31	0	161.58
4	WOULD	321	0.72	56	150.47
5	**KRISKA**	110	0.25	0	128.76
6	I	1,748	3.90	910	114.80
7	**BOOK**	109	0.24	4	99.51
8	**HUNGARIAN**	83	0.19		97.14
9	**ÁLVARO** 51	0.11	0	59.67	
10	NOT	259	0.58	85	56.96

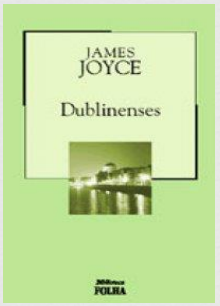
RC in Translation
(Gonçalves 2006)



15 short stories

2 Brazilian translations:
Hamilton Trevisan &
José Roberto O'Shea





The Sisters

Nº	JOYCE	oc.	Trevisan	oc.	O'Shea	oc.
1	aunt	19	tia	13	tia	20
2	head	10	cabeça	7	padre	10
3	house	7	titia	7	cabeça	8
4	night	7	casa	6	noite	7
5	room	7	noite	6	olhos	7
6	mind	6	olhos	6	casa	6
7	snuff	6	padre	6	dia	6
8	chair	5	tempo	6	rapé	6
9	children	5	cálice	5	sinal	6
10	father	5	Deus	5	Deus	5
11	priest	5	lareira	5	caixão	4
12	chapel	4	caixão	4	cálice	4
13	friends	4	coisa	4	capela	4
14	God	4	cortinas	4	gente	4
15	uncle	4	guarda	4	guarda	4

Tabela 6.2: Os 15 substantivos mais ocorrentes (texto original e traduções)

As Irmãs

Final remarks

- Reference corpora are not ONLY general language corpora
- Customized comparison corpora can generate better results
- Recommendation is that reference corpora should be 3 to 5 times bigger than study corpus
- BUT size is NOT ALWAYS relevant.
- Comparison results – KEYWORDS – are first step towards in-depth analyses.

Bibliography

- Aguiar, Sergio Marra. 2010. *As vozes de Chico Buarque em inglês: tradução e linguística de corpus*. PhD dissertation. University of São Paulo, Brazil.
- Gonçalves, Lourdes Bernardes. 2006. *"Dubliners" sob a lupa da Linguística de Corpus. Uma contribuição para a análise e a avaliação da tradução literária*. PhD dissertation. University of São Paulo, Brazil.
- Scott, Michael. 2010. *Introduction to WordSmith Tools*. Available at http://www.lexically.net/downloads/version5/HTML/index.html?referencecorpus.htm
- Tagnin, S.E.O & Teixeira. E.D. 2004. British vs. American English, Brazilian vs. European Portuguese: How close or how far apart? - A corpus-driven study. In *Practical Applications in Language and Computers - PALC 2003*, Frankfurt am Main: Peter Lang, 193-208.

Thank you

Stella

seotagni@usp.br

Baker Street – Jô Soares

- Mixture of 19thC and 20thC language
- Corpus of 19th century texts (authors mentioned in JS's novel)
- Corpus graph
- → 20thC language