SUBTLEX_PT: Uma nova medida de frequência lexical para o português europeu baseada em legendas de filmes e séries televisivas

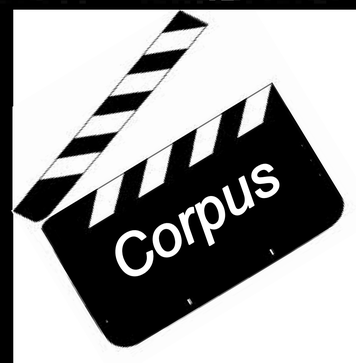
Universidade do Minho, Escola de Psicologia
Laboratório Cognição Humana

Ana Paula **Soares**[1], João **Machado**[1], Ana **Costa**[1], Alberto **Simões**[2], José João de **Almeida**[2], Álvaro **Iriarte**[2] & Montserrat **Comesaña**[1]

{[1]Escola Psicologia, [2]Departamento de Informática, [3]Instituto de Letras e Ciências Humanas}, Universidade do Minho

7º ENCONTRO NACIONAL DA ASSOCIAÇÃO PORTUGUESA DE PSICOLOGIA EXPERIMENTAL

A validade concorrente das medidas de frequência lexical utilizadas em estudos psicolinguísticos tem sido recentemente questionada. De uma forma geral esses estudos têm demonstrado que as medidas de frequência e de diversidade contextual (número de filmes em que a palavra ocorre) extraídas a partir de legendas de filmes e séries televisivas explicam uma percentagem significativamente maior de variância da precisão e dos tempos de reconhecimento e nomeação de palavras do que outras medidas de frequência classicamente utilizadas. Neste trabalho apresentamos uma nova medida de frequência lexical e de diversidade contextual para ≈136.000 palavras do português europeu extraídas a partir de um *corpus* de ≈78 milhões de palavras derivados de ≈17.500 legendas de filmes e séries televisivas obtidos a partir do OPUS (http://opus.lingfil.uu.se/).



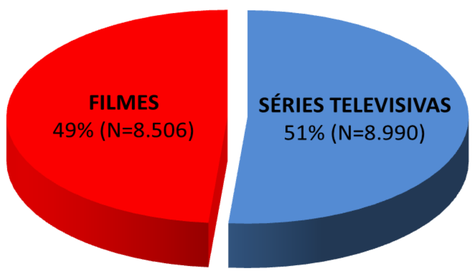


Figura 1. Distribuição das 17.496 legendas no *corpus* SUBTLEX_PT.

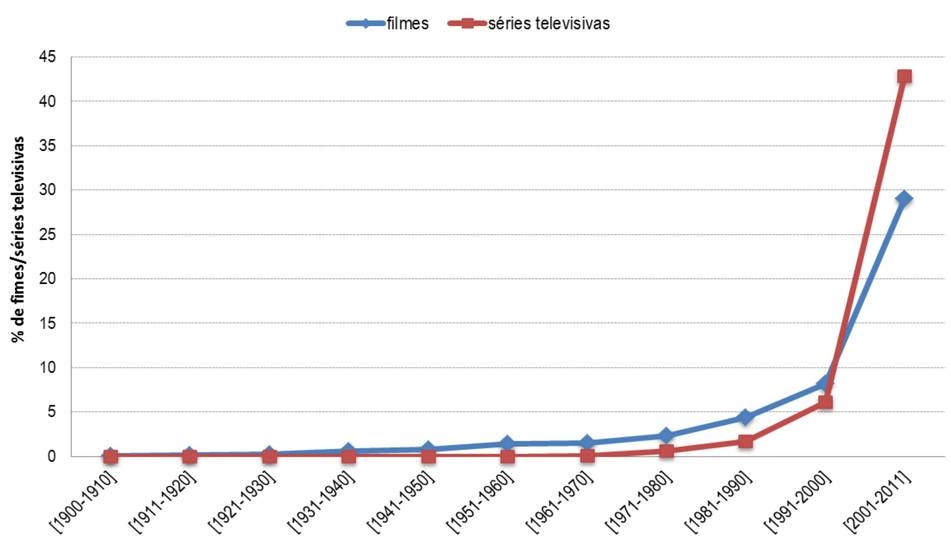


Figura 2. Distribuição dos filmes e séries televisivas do *corpus* SUBTLEX_PT por década.

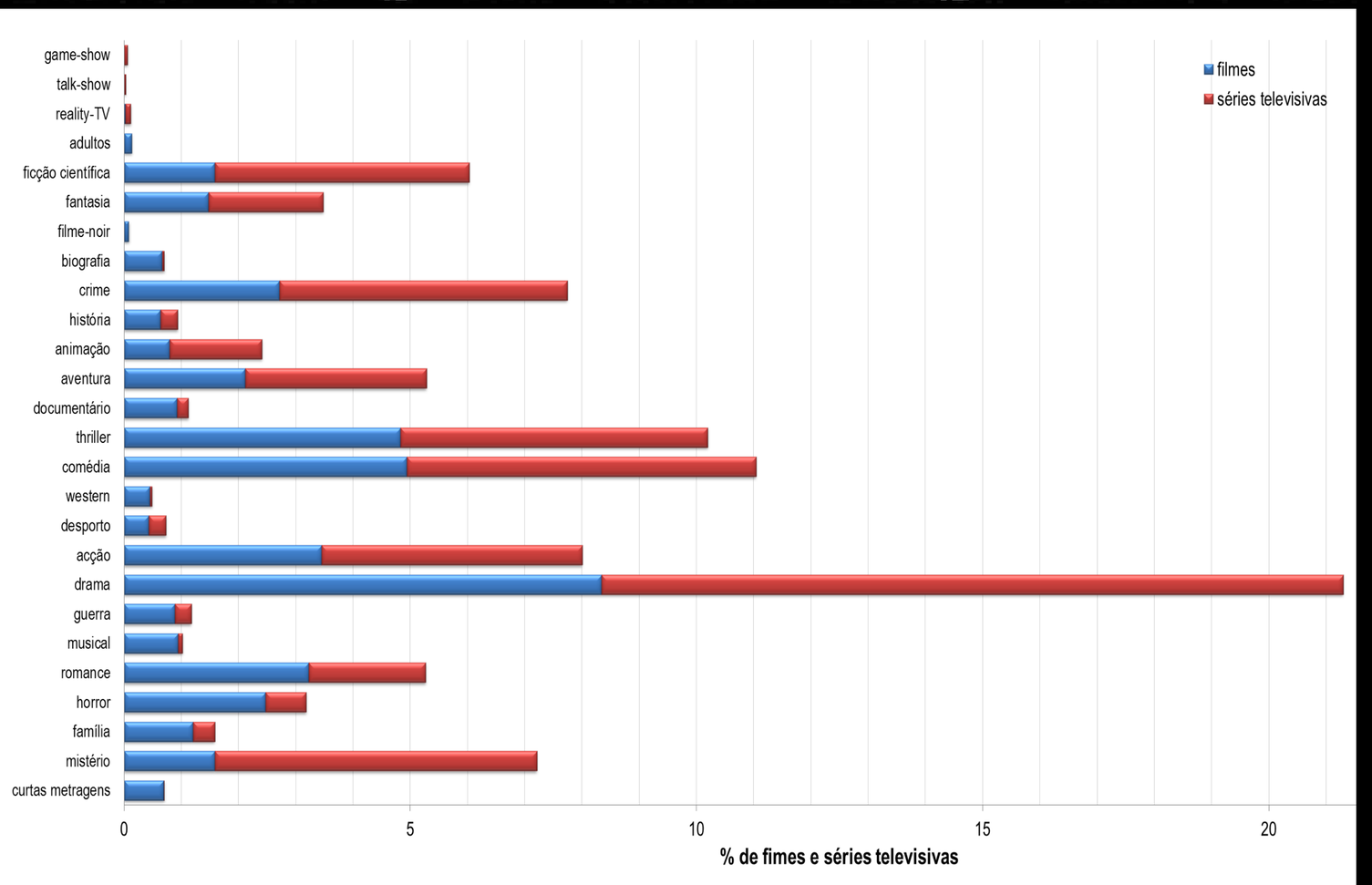


Figura 3. Distribuição dos filmes e séries televisivas do *corpus* SUBTLEX_PT por género cinematográfico (http://www.imdb.com/).

PALAVRA: contém 135.598 formas flexionadas (flexões verbais e nominais) do português europeu que ocorrem no *corpus* SUBTLEX_PT. Da base lexical fazem parte todas as formas distintas, não se diferenciando os casos das homógrafas não homófonas (ex. "forma" ['fɔrmɐ] e "forma" ['formɐ]) e as palavras homónimas (ex. "além" [nome] e "além" [advérbio]), que constituem entrada única na base.

$FREQ_{cont}$: número de vezes que a palavra ocorre no *corpus* SUBTLEX_PT (i.e., no total de 78.402.091 palavras).

DC_{cont}: número de filmes e séries televisivas nos quais a palavra ocorre (i.e., num valor máximo de 17.496 filmes e séries televisivas).

$FREQ_{milhão}$: medida standard de frequência por milhão de palavras que toma em consideração a dimensão do *corpus*. É apresentada com 2 dígitos de precisão para não perder informação das contagens absolutas.

	A	B	C	D	E	F	G
1	PALAVRA	FREQcont	DCcont	FREQmilhão	LOG10_cont	DC_%	LOG10_DC
2	a	2701655	9246	34458,97	6,4316	52,85	3,9660
3	de	2096154	9221	26735,95	6,3214	52,70	3,9648
4	o	2584427	9209	32963,75	6,4124	52,63	3,9643
5	que	2999219	9208	38254,32	6,4770	52,63	3,9642
6	no	551497	9195	7034,21	5,7415	52,55	3,9636
7	para	912398	9195	11637,42	5,9602	52,55	3,9636
8	se	838421	9194	10693,86	5,9235	52,55	3,9636
9	por	606118	9191	7730,89	5,7826	52,53	3,9634
10	como	438050	9176	5587,22	5,6415	52,45	3,9627
11	me	765240	9176	9760,45	5,8838	52,45	3,9627
12	nos	218977	9111	2793,00	5,3404	52,07	3,9596
13	ser	220595	9091	2813,64	5,3436	51,96	3,9587
14	nada	173218	9073	2209,35	5,2386	51,86	3,9578
15	está	668064	9067	8521,00	5,8248	51,82	3,9575
16	vamos	260254	9059	3319,48	5,4154	51,78	3,9571
17	ver	134103	9020	1710,45	5,1274	51,55	3,9553
18	da	425268	9010	5424,19	5,6287	51,50	3,9548
19	todos	130561	8976	1665,27	5,1158	51,30	3,9531
20	era	163585	8970	2086,49	5,2137	51,27	3,9528
21	nunca	122106	8964	1557,43	5,0867	51,23	3,9526
22	e	1299762	8961	16578,15	6,1139	51,22	3,9524
23	vez	109379	8959	1395,10	5,0389	51,21	3,9523
24	porque	188080	8916	2398,92	5,2743	50,96	3,9502
25	esta	140239	8900	1788,72	5,1469	50,87	3,9494
26	dos	128471	8888	1638,62	5,1088	50,80	3,9489
27	este	127787	8861	1629,89	5,1065	50,65	3,9475
28	ir	109193	8852	1392,73	5,0382	50,59	3,9471
29	os	476814	8769	6081,65	5,6783	50,12	3,9430
30	mas	442404	8766	5642,76	5,6458	50,10	3,9429
31	te	519723	8763	6628,94	5,7158	50,09	3,9427
32	com	600065	8723	7653,69	5,7782	49,86	3,9407

SUBTLEX_PT

$LOG10_{cont}$: valor que resulta do calculo do logarítmo de base 10 da $FREQ_{cont}$+1. Como a medida $FREQ_{cont}$ se baseia num *corpus* de 78.402.091 palavras, um valor LOG10=0,3 corresponde a palavras que ocorrem apenas uma vez no *corpus* e LOG10>5 que ocorrem mais de 100.000 vezes no *corpus*. É apresentada com 4 dígitos de precisão.

DC_%: valor que indica a percentagem de filmes e séries televisivas em que a palavra ocorre. É apresentada com 2 dígitos de precisão.

$LOG10_{DC}$: valor que resulta do calculo do logarítmo de base 10 da DC_{cont}+1. Como a medida DC_{cont} se baseia em 17.496 legendas de filmes e séries televisivas, um valor $LOG10_{DC}$<1,08 corresponde a palavras que ocorrem em menos de 10 filmes e séries televisivas e LOG10>3 que ocorrem em mais de 1.000 filmes e séries televisivas. É apresentada com 4 dígitos de precisão.

para mais informações:
asoares@psi.uminho.pt

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
UNIÃO EUROPEIA FEDER
COMPETE PROGRAMA OPERACIONAL FACTORES DE COMPETITIVIDADE
QREN QUADRO DE REFERÊNCIA ESTRATÉGICO NACIONAL

Projecto de Investigação PTDC/PSI-PCO/104679/2008 financiado pela Fundação para a Ciência e a Tecnologia (FCT) e co-finaciado pelo FEDER (Fundo Europeu de Desenvolvimento Regional) no âmbito do programas COMPETE (Programa Operacional Factores de Competitividade) e do QREN (Quadro de Referência Estratégico Nacional).