On the advantages of frequency measures extracted from subtitles: The case of Portuguese

Ana Paula Soares[1], João Machado[1], Ana Costa[1], Montserrat Comesaña[1] & Manuel Perea[2]

[1]Human Cognition Lab, School of Psychology, University of Minho, Braga, Portugal & [2]Department of Methodology, University of València, València, Spain

Human Cognition Lab
Research Group in Psycholinguistics
School of Psychology, University of Minho

VNIVERSITAT DE VALÈNCIA

The predictive validity of written word-frequency measures in visual-word recognition has been recently questioned (e.g., Brysbaert & Cortese, 2011; Brysbaert & New, 2009; Brysbaert et al., 2011; Cai & Brysbaert, 2010; Dimitroupoulus et al., 2010; Keuleurs et al., 2010). In general, these studies have revealed that measures of word-frequency based on film and television series subtitle corpora explained a significantly higher percentage of variance in naming and lexical decision performance than written word-frequency measures typically used in psycholinguistic research (e.g., Kučera and Francis norms, 1967; British National Corpus, Leech et al., 2001; the Zeno corpus, Zeno et al., 1995; CELEX database, Baayen et al., 1993). In this work we present SUBTLEX-PT, a new Portuguese database which offers word frequency measures for ≅135,000 words extracted from a ≅78 million words corpus based on ≅17,500 film and television series subtitles (see Soares et al., 2012). Additionally, we validated these measures with a lexical decision study. As its international counterparts, the new SUBTLEX-PT frequency measures explained more variance in the visual word recognition times than the recently established P-PAL word frequency norms, largely based on newspaper corpora (Soares et al., in press). SUBTLEX-PT is freely available for research on http://p-pal.di.uminho.pt/about/database.

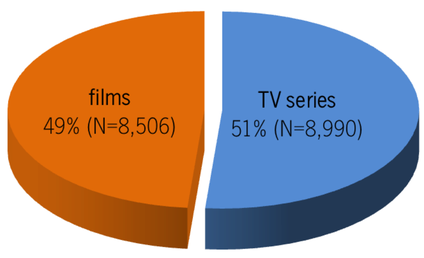


Figure 1. Distribution of the 17,496 subtitles in the SUBTLEX-PT corpus.

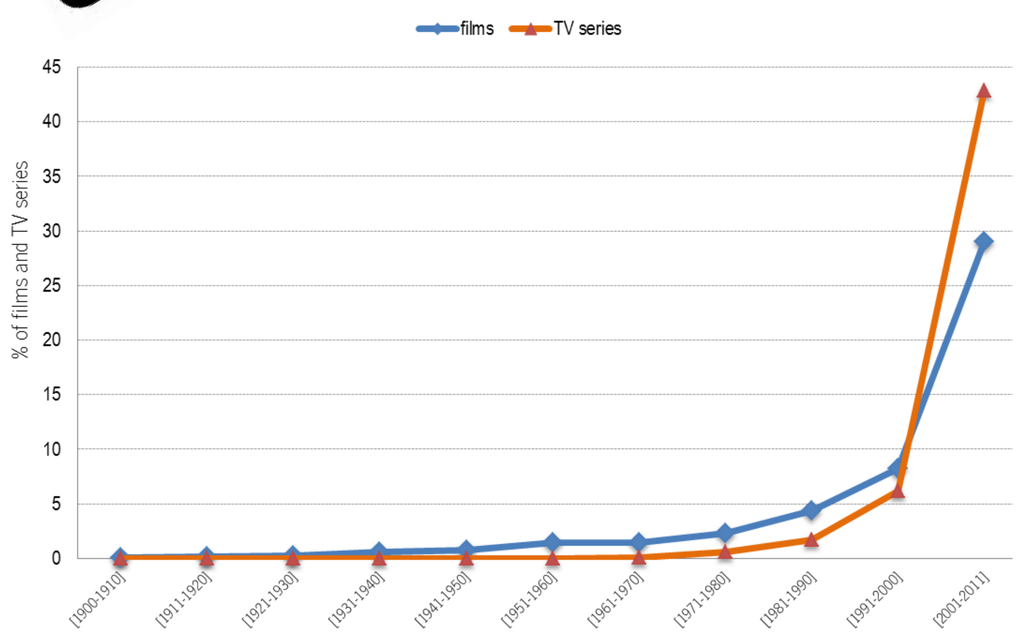


Figure 2. Film and television series subtitle distribution in the SUBTLEX-PT corpus by decade.

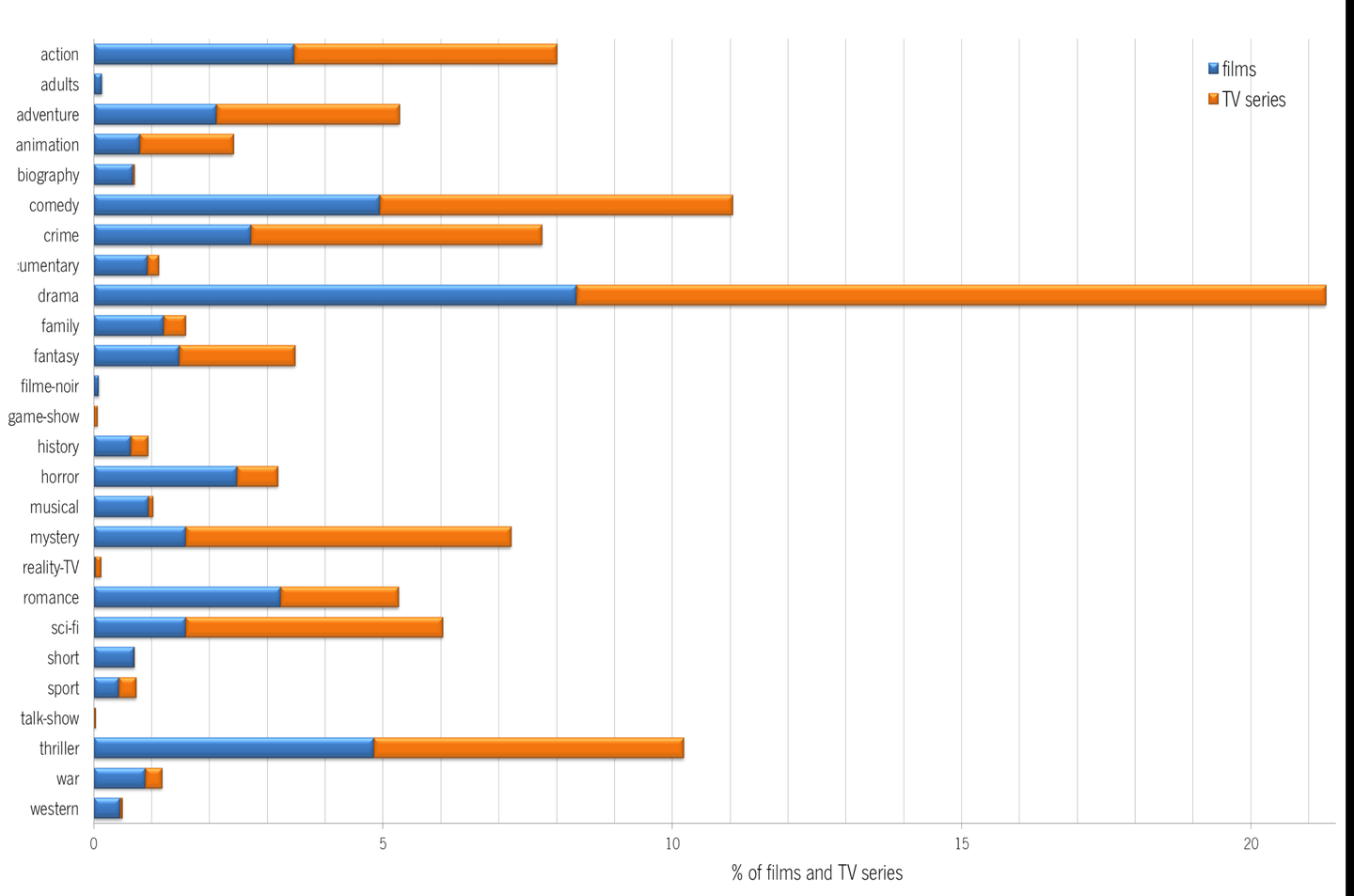


Figure 3. Film and television series subtitle distribution in the SUBTLEX-PT corpus by IMDb genre (http://www.imdb.com/).

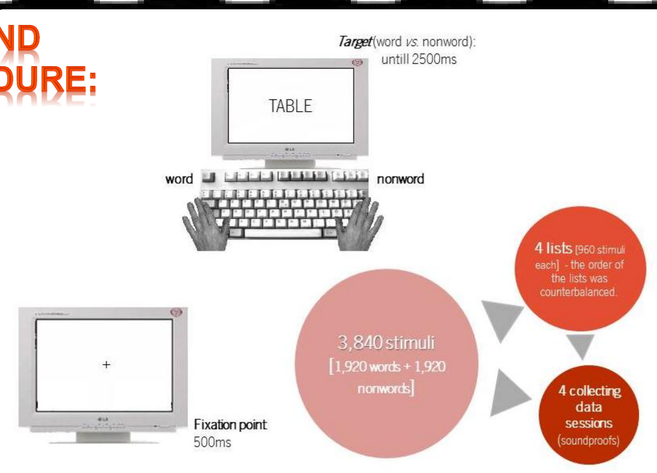
PARTICIPANTS:

- 58 psychology undergraduates from the University of Minho (Portugal) participated in the experiment in exchange for course credit (52 females; M_{age}: 21.3 years-old, *SD*=3.06).
- All participants were native speakers of European Portuguese (EP) and had normal or corrected-to-normal vision.

MATERIALS:

- 1,920 Portuguese words that vary in length (number of letters: *M*=6.89, *SD*=2.10, *range*: 2 to 15; number of syllables: *M*=2.99, *SD*=0.94, *range*: 1 to 6), and frequency per million words obtained from the P-PAL lexical database (*M*=67.33, *SD*=110.83, range: 0.44 to 1,214.45) (Soares et al., in press; available at http://p-pal.di.uminho.pt/tools) and from the SUBTLEX-PT lexical database (*M*=61.41, *SD*=142.32, *range*: 0.09 to 1,907.57) (Soares et al., 2012). Figure 4 compares the frequency values in both databases considering the total number of wordforms in each.
- 1,920 Portuguese orthographically legal nonwords for the purposes of the lexical decision task (e.g., *tordomo*, *esvoto*, *falanha*, *laide*). The manipulation of the nonword trials was the same as that for the word trials.

TASK AND PROCEDURE:



RESULTS:

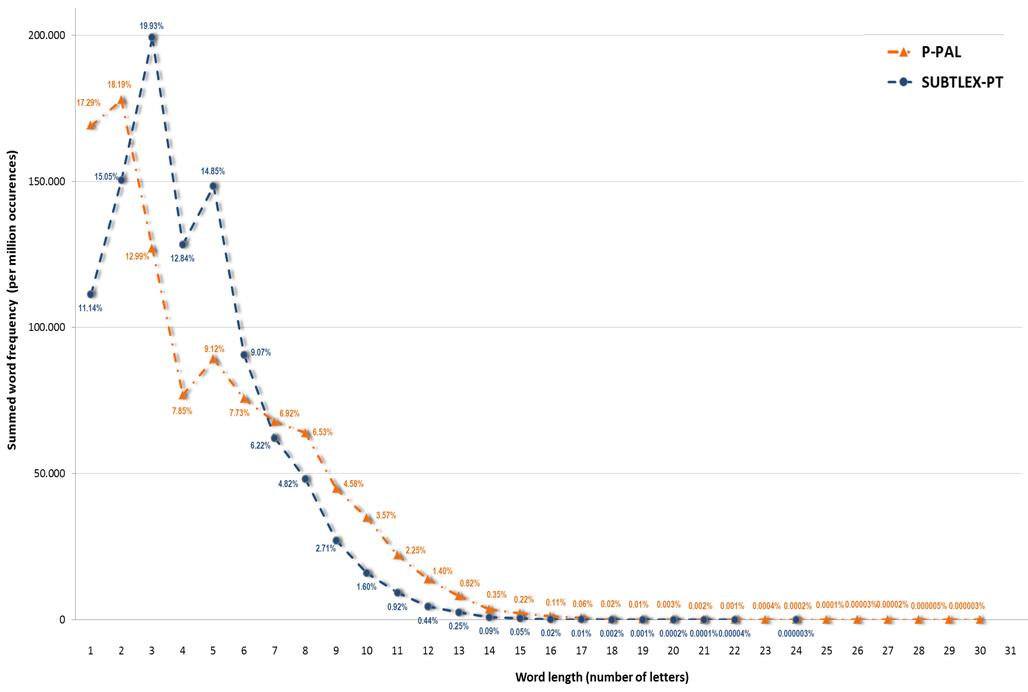


Figure 4. Summed word frequency distribution (per million occurrences) for the 133,791 and 218,518 wordforms in the SUBTLEX-PT and the P-PAL corpus respectively by word length (% of summed word frequencies by word length is also presented).

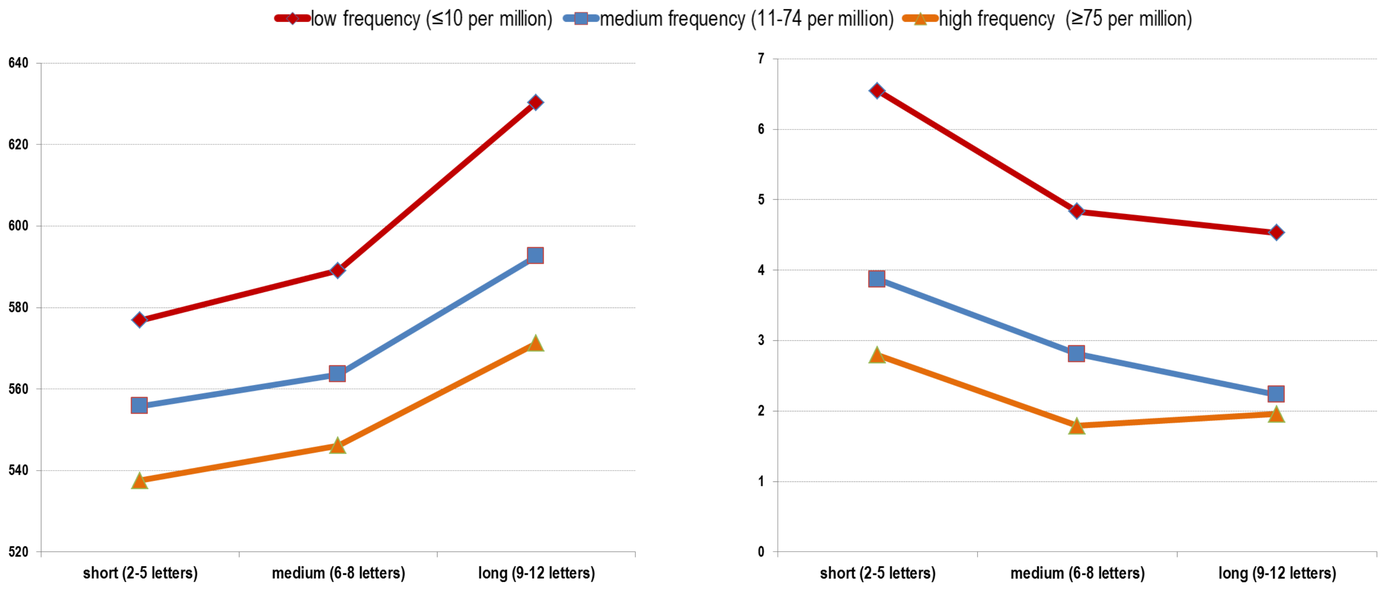


Figure 5. Reaction times (ms) of the correct word trials (*N=1,909*) by word frequency (per million occurences) and word length (number of letters).

Figure 6. Error responses (%) of the word trials (*N=1,909*) by word frequency (per million occurences) and word length (number of letters).

Table 1. Correlations between Log_{10} word frequency (Log_{10}WF) and Lexical Decision Times (TR) and Accuracy (Acc) for P-PAL levels of frequency categorization (*N*=1,909).

Corpus	Levels of word frequency	Word length (number letters)	P-PAL Log_{10}WF correlated with RT	SUBTLEX-PT Log_{10}WF correlated with RT	P-PAL Log_{10}WF correlated with Acc	SUBTLEX-PT Log_{10}WF correlated with Acc
P-PAL categorization	*All*		-.43**	-.59**	-.28**	-.29**
	Low (≤10)	all	-.32**	-.51**	-.23**	-.30**
		short	-.32**	-.46**	-.29**	-.44**
		medium	-.41**	-.55**	-.27**	-.39**
		long	-.30**	-.38**	-.17	-.19
	Medium (11-74)	all	-.19**	-.53**	-.13**	-.20**
		short	-.23**	-.48**	-.17**	-.29**
		medium	-.18**	-.49**	-.10*	-.36**
		long	-.28**	-.40**	-.13	-.11
	High (≥75)	all	-.17**	-.49**	-.01	-.05
		short	-.02	-.20*	-.05	-.06
		medium	-.11	-.39**	-.12	-.18
		long	-.34**	-.59**	-.07	-.30**

*p<.05, **p<.001

Table 2. Correlations between Log_{10} word frequency (Log_{10}WF) and Lexical Decision Times (TR) and Accuracy (Acc) for SUBTLEX-PT levels of frequency categorization (*N*=1,909).

Corpus	Levels of word frequency	Word length (number letters)	P-PAL Log_{10}WF correlated with RT	SUBTLEX-PT Log_{10}WF correlated with RT	P-PAL Log_{10}WF correlated with Acc	SUBTLEX-PT Log_{10}WF correlated with Acc
SUBTLEX-PT categorization	*All*		-.43**	-.59**	-.28**	-.29**
	Low (≤10)	all	-.26**	-.34**	-.28**	-.20**
		short	-.12	-.23*	-.17	-.36**
		medium	-.34**	-.38**	-.23**	-.34**
		long	-.37**	-.29**	-.26**	-.09
	Medium (11-74)	all	-.21**	-.32**	-.16**	-.11**
		short	-.26**	-.34**	-.15*	-.13*
		medium	-.16**	-.29**	-.14**	-.18**
		long	-.47**	-.32**	-.12	-.10
	High (≥75)	all	-.16**	-.25**	-.02	-.04
		short	-.08	-.21**	.003	.04
		medium	-.22**	-.29**	.06	-.06
		long	-.25	-.06	-.26	-.30

*p<.05, **p<.001

Table 3. Percentages of variance in Lexical Decision Times (TR) and Accuracy (Acc) explained by the P-PAL and SUBTLEX-PT frequency mesures (Log_{10} word frequency - Log_{10}WF, and Log^2_{10}WF, and Contextual Diversity- Log_{10}CD and Log^2_{10}CD).

Frequency measures		Word length	RT (*N*=1,909)	Accuracy (*N*=1,909)
$P\text{-}PAL_{WF}$	Log_{WF}	all	18.3	7.9
		short	18.3	8.9
		medium	18.9	8.9
		long	23.8	7.1
	Log_{WF} + Log^2_{WF}	all	19.3**	9.9**
		short	20.2**	12.5**
		medium	20.7**	10.2**
		long	23.8	9.0**
$SUBTLEX\text{-}PT_{WF}$	Log_{WF}	all	34.7	8.2
		short	28.7	14.0
		medium	32.8	16.4
		long	25.0	4.5
	Log_{WF} + Log^2_{WF}	all	35.9**	9.8**
		short	30.5**	20.6**
		medium	34.2**	19.6**
		long	25.0	4.5
$SUBTLEX\text{-}PT_{DC}$	Log_{DC}	all	37.4	9.8
		short	31.2	17.1
		medium	35.9	18.5
		long	27.4	5.5
	Log_{DC} + Log^2_{DC}	all	37.7**	10.9**
		short	31.2	20.9**
		medium	36.6**	21.3**
		long	27.4	5.5

R^2 changes **p<.001

asoares@psi.uminho.pt

Research project [PTDC/PSI-PCO/104679/2008] funded by FCT (Fundação para a Ciência e Tecnologia), and FEDER (Fundo Europeu de Desenvolvimento Regional) through the European programs QREN (Quadro de Referência Estratégico Nacional), and COMPETE (Programa Operacional Factores de Competitividade).