Extracting neighborhood density for lexical access using phonologically-weighted Levenshtein Distance (pwLD)

Emiliano Zaccarella[1], Ute Gradmann[2], Anna Carthaus[1], & Sören E Tebay[3]

[1]Max Planck for Human Cognitive and Brain Sciences Leipzig, Department of Neuropsychology

[2]Univesität Heidelberg; [3]Universität Leipzig

zaccarella@cbs.mpg.de

INTRODUCTION. The visual recognition of a certain word can be influenced by the partial co-activation of other words that share with it varying degrees of orthographic similarity (McClelland & Rumelhart 1981). One well-known measure of orthographic similarity is Coltheart's N (ON, Coltheart et al. 1977: the neighborhood size of a word λ is equal to the amount of words of the same length of λ, which can be obtained by changing one letter in λ; MILK → MILE). A more recent measure is Orthographic Levenshtein distance 20 (OLD20), which overcomes both ON binarity (a word κ is either neighbor of λ, or not) and its inability to capture insertions, deletions or transpositions (Yarkoni et al. 2008). OLD20, which stands for the mean LD across the 20 closest Levenshtein neighbors for λ, has been shown to have higher power in predicting lexical decision latencies compared to ON (Yarkoni et al. 2008). OLD20 is now a standardly used controlling variable for psycholinguistic experiments and pseudo-word generators (Brysbaert et al. 2010). Phonological LD (PLD) measures are also available (ELP; https://elexicon.wustl.edu/), given the large literature showing how skilled readers activate phonological representations early during silent word reading (Ashby 2010). Here we reason that OLD/PLD are not sensitive enough to fine-grained similarities/differences between the phonological features forming phonological representations, following recent findings in microscopic intelligibility for Automatic Speech Recognition (ASR; Fontan et al 2016). In this respect, /beɪ/ (bay) is equally distant to /peɪ/ (pay) and /seɪ/ (say) under OLD/PLD, although /beɪ/ and /peɪ/ only differ along voicing, while /beɪ/ and /seɪ/ differ along both manner and place.

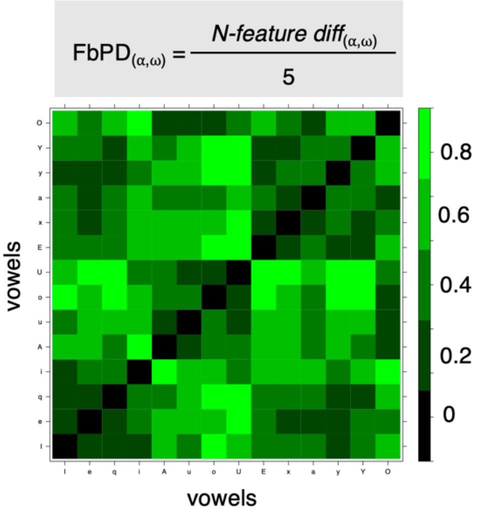
METHODS. We hypothesized that the explanatory power of a phonological feature-based measure of word similarity goes beyond the one of OLD20 and PLD20 to predict behavior during word recognition. Following Fontan et al. (2016), we first developed a Feature-based Phonological Distance (FbPD) between all phonemes of the English language. We created one FbPD matrix for all English consonants, based on four features (consonant, voice, place, manner) and one FbPD matrix for all vowels, according to five features (vowel, height, backness, roundedness, length; see Fig. 1). We then took the 2421 IPA-converted monosyllabic English words from Balota et al. (2004)'s dataset that had already been used to test the superiority of OLD20 against ON in predicting lexical decision latencies in Yarkoni et al. (2008). We used to the FbPD matrixes to calculate the phonologically-weighted LD20 for each word against a reference corpus formed by the 40481 most frequent IPA-converted words from the ELP website (pwLD20; see Fig. 2). To compare the predictive utility of OLD20, PLD20 and pwLD20, we run a series of hierarchical regression analyses (3-step-based) on the reaction times (RTs) and the accuracies of both the lexical decision and the naming task in the database (naming task analysis not reported here for space), consistently following the statistical structure of Balota et al. (2004) and Yarkoni et al. (2008).

RESULTS. Matrix correlations can be found in Fig. 3. After replicating the findings reported in the two studies above, our results show that only the Model3C, the one including pwLD20 as additional predictor variable, reveals a small yet statistically significant change (DR2 = 0.002; $p < 0.001$) compared to the previous model for both the item-based RT analysis (Table 1), the subject-based RT analysis and the accuracy data.

CONCLUSION. Overall, the results confirm human aptitude to form phonological representations of speech during reading, with high sensitivity to meaningful feature-based phonological distinctions between words. PwLD20 can integrate previous psycholinguistic measures of neighborhood density and be easily adaptable to any other testable language.

References: Ashby, *Psychon Bull Rev*, 2010; Balota et al., *J Exp Psychol Gen*, 2004; Brysbaert et al., *Behav Res Methods*, 2010; Coltheart et al., in *Attention and Performance VI,* 1977; Fontan et al., *INTERSPEECH*, 2016; McClelland & Rumelhart, *Psychol Rev*, 1981; Yarkoni et al., Psychon Bull Rev, 2008.

FIGURES AND TABLES



$$FbPD_{(\alpha,\omega)} = \frac{\textit{N-feature diff}_{(\alpha,\omega)}}{5}$$

Figure 1
Feature-based phonological distance (FbPD) between any two α and ω vowels of the English language, calculated according to five discriminant features (see abstract).

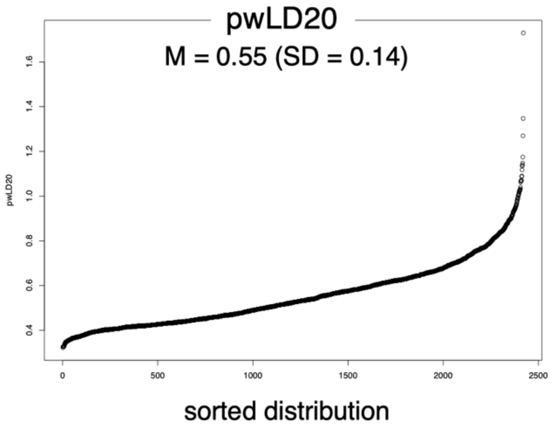


Figure 2
Phonologically-weighted LD 20 (pwLD20) for the 2421 words used in the lexical decision task from Balota et al. 2004, sorted from lowest to highest. Mean (M) and Standard Deviation (SD).

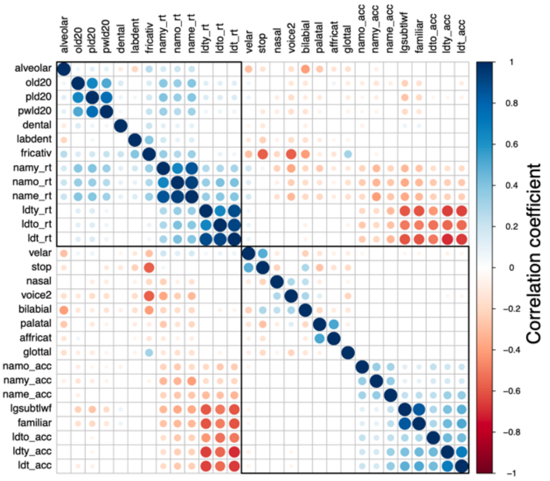


Figure 3
Correlations between predictors used in Balota et al. 2004 (plus pwLD20) with RT and ACC from Balota et al. 2004. White cells stand for correlation above threshold set at p < .01.

	LDT (all)		LDT (young)		LDT (old)	
Predictor	*β*	*p<F*	*β*	*p<F*	*β*	*p<F*
Model3A						
OLD20	-0,03		0,01		-0,06	*
Δ R^2	0,000		0,000		0,002	
R^2	0,43		0,41		0,32	
Model3B						
PLD20	-0,02		0,01		-0,04	
Δ R^2	0,000		0,000		0,001	
R^2	0,43		0,41		0,32	
Model3C						
pwLD20	-0,06	**	-0,03	.	-0,07	***
Δ R^2	0,002		0,001		0,003	
R^2	0,43		0,41		0,32	

Table 1
Results of the hierarchical regression analysis (step 3) for the lexical decision task (LDT) for both the young and old populations including OLD20, PLD20 and pwLD20 (red box) as predictors. **p* < .05, ***p* < .01, ****p* < .001