A cross-linguistic analysis of phonological neighborhood density

Julien Eychenne & Yejin Shon

Hankuk University of Foreign Studies (South Korea),
jeychenne@hufs.ac.kr, shon.yejin@gmail.com

A number of studies have shown that lexical access, in both production and perception, is influenced by phonological neighborhood ([1][2][3] and references therein). The most common metric, neighbor-hood density (ND), is usually operationalized as the number of lexical items that differ from a target word by the addition, deletion or substitution of a single phoneme. Although neighborhood density has featured prominently in phonetic and psycholinguistic research, little is known about how it relates to other properties of the lexicon, and why some words have more neighbors than others. In a landmark study, [4] built and analyzed a database of orthographic and phonological neighborhood densities (CLEARPOND) for five Indo-European languages, namely American English, Dutch, French, German and Spanish. They showed that ND varied across these languages, and that it was correlated with word length and word frequency.

This paper builds upon the work in [4] and offers a more comprehensive analysis of neighborhood density from a cross-linguistic perspective. In order to obtain a more typologically balanced sample, we kept four languages from the CLEARPOND database, namely English, French, Dutch and Spanish, and added six other non-Indo-European languages, including three non-tonal languages (Finnish, Korean and Malay) and three tonal ones (Cantonese, Mandarin and Vietnamese). The inclusion of tonal languages seemed important since it has been shown that ND plays a role at the tonal level as well [5], and conservative estimates suggest that more than 40 % of the World's languages are tonal [6]. In addition to (log) word frequency and (log) word length, we investigated the potential effect of average phoneme probability, normalized pointwise mutual information (NPMI) [7], part of speech (POS), the number of homophones and the tonal/non-tonal nature of the language. In order to make the sample balanced, we only considered the 7,000 most frequent words in each language. (Word frequency was normalized and ND statistics were calculated on this sample.).

The data were analyzed using a negative binomial mixed effects model with the above mentioned predictors as fixed effects, and language as a random effect. In addition to confirming and nuancing the findings in [4], regarding the effect of word frequency and word length, the results of this study show that ND is positively correlated with average phoneme probability and average NPMI (words which have a more phonologically predictable shape tend to have more neighbors), as well as the number of homophones: words with higher ND tend to have more homophones, which suggests that they lie in a dense network of similar words. In addition, we find that verbs have more neighbors than the other categories. Most importantly, we find a clear difference between tonal and non-tonal languages, since we find that tonality significantly interacts with part of speech, NPMI, phoneme probability and homophone density. Our results also reveal that there is important cross-linguistic variation (see Figure 1). We discuss the significance of these results for phonetic and psycholinguistic experiments involving ND.

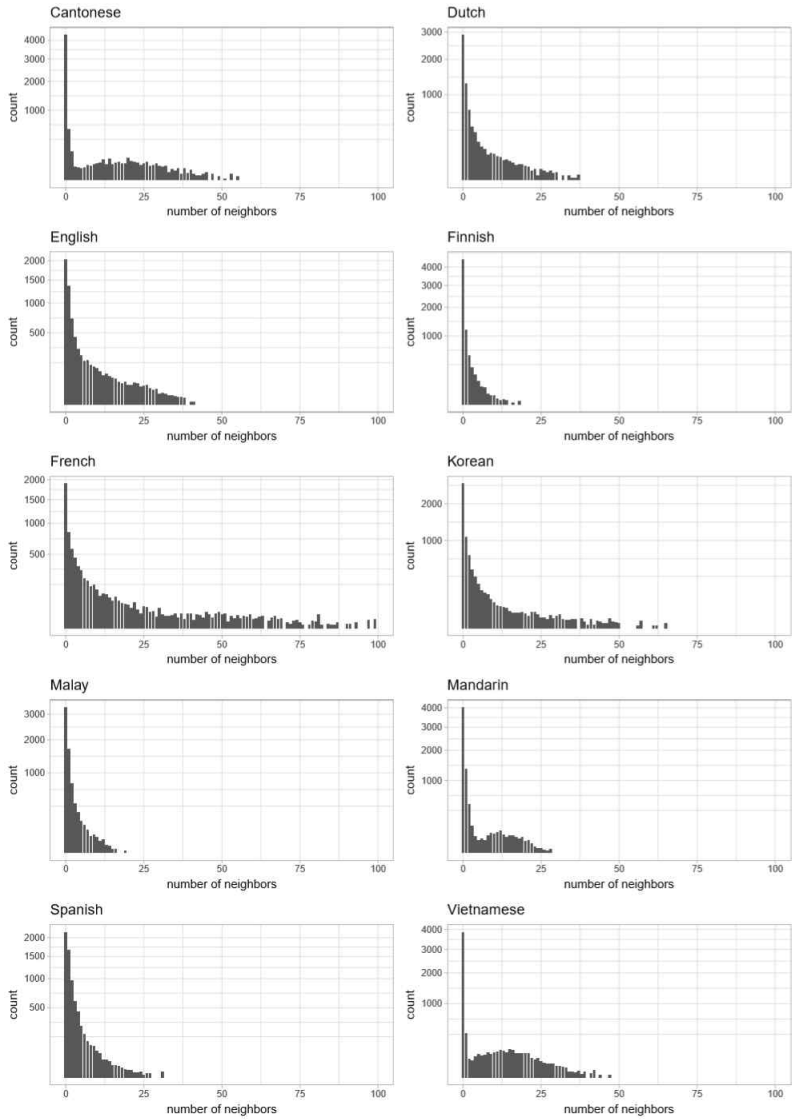

Figure 1 Number of neighbors across 10 languages in frequency spectrum

References

[1] Luce, Paul A. and David B. Pisoni. 1998. Recognizing Spoken Words: The Neighborhood Activation Model. Ear and hearing 19(1). 1-36.

[2] Munson, Benjamin and Nancy Pearl Solomon. 2004. The Effect of Phonological Neighborhood Density on Vowel Articulation. Journal of Speech Language and Hearing Research 47(5). 1048-1058.

[3] Scarborough, Rebecca. 2013. Neighborhood-conditioned patterns in phonetic detail: Re-lating coarticulation and hyperarticulation. Journal of Phonetics 41(6). 491-508.

[4] Marian, Viorica, James Bartolotti, Sarah Chabal and Anthony Shook. 2012. CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. PLOS ONE 7(8). e43230.

[5] Yao, Yao and Bhamini Sharma. 2017. What is in the neighborhood of a tonal syllable? Evidence from auditory lexical decision in Mandarin Chinese. Proceedings of the Linguistic Society of America 2. 1-14.

[6] Maddieson, Ian. 2013. Tone. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/13.

[7] Bouma, Gerlof. 2009. Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL 31-40.