The effect of orthographic neighborhood density on reading time in 9 languages

James Michaelov (MIT), Roger Levy (MIT)

Introduction The *orthographic neighborhood* of a word refers to how orthographically similar it is to other words in the language (i.e., more similar words are 'closer' neighbors). Orthographic neighborhood density (OND)—generally operationalized as either the number of words that can be made by switching one letter in a given word (Coltheart's *N*; Coltheart *et al.*, 1977), or, more recently, the mean distance between the 20 most orthographically similar words (Orthographic Levenshtein Distance 20 or OLD20; Yarkoni *et al.*, 2008)—has been shown to influence language processing at the word level. For example, lexical decision times are faster for words and slower for non-words with denser orthographic neighborhoods, and words with denser neighborhoods elicit larger N400 responses (see Holcomb *et al.*, 2002).

The effect of OND on naturalistic reading is less well-understood. To the best of our knowledge, three studies have been carried out directly addressing this question, with conflicting results. Pollatsek *et al.* (1999) find that words with a higher Coltheart's *N* are read more slowly in English, which they argue is due to competition with higher-frequency neighbors. Tsai *et al.* (2006), on the other hand, observe the opposite effect with Chinese (a finding replicated by Yao *et al.* (2022) for low-predictability items), which they argue may arise either due to faster verification that the word is familiar (Reichle *et al.*, 1998), or due to easier lexical access for words with higher-frequency initial characters (see Lima & Inhoff, 1985).

We revisit the question of how OND impacts reading time, going beyond previous work in two ways. First, we use OLD20 as our metric of OND, which is able to more flexibly account for orthographic neighborhood than the metrics used in the three previous studies (see Yarkoni *et al.*, 2008). Second, in addition to looking at English, we carry out the first study looking at the effect of OND on reading time in Finnish, German, Greek, Italian, Korean, Russian, Spanish, and Turkish.

Method For reading time, we use the log-transformed go-past duration of words in the Finnish, German, Greek, Italian, Korean, Russian, Spanish, and Turkish subsets of the MECO dataset (Siegelman *et al.*, 2022). We predict go-past duration using linear mixed-effects regressions, with OLD20 and language model surprisal as predictors, as well as word frequency (Speer *et al.*, 2018), word length, and the position of the word in the text. All variables were *z*-scored.

Results As can be seen in Figure 1, we see that OLD20 explains significant variance ($p<0.05$ on all likelihood ratio tests) in go-past duration above and beyond surprisal and the baseline predictors on the English, Finnish, German, Greek, Italian, Russian, and Spanish datasets. We only see this when using surprisal from 5/9 language models on the Turkish dataset, and do not see this at all for Korean words. In the latter case, we hypothesized that this may be partly due to the syllabic nature of the Korean writing system. We therefore re-ran the analysis on the same words decomposed into their component parts (*jamo*), and in this case, we do find a significant effect of OLD20. For all languages, regressions estimated a positive effect of OLD20 on go-past duration, indicating that words with denser orthographic neighborhoods are read faster.

Discussion With the exception of Turkish and Korean at the syllable level, we find consistently that words with more close orthographic neighbors are read faster. We also observe differences in the effect of OLD20 on reading time across languages. This may reflect differences between writing systems or differences in how individual languages are processed (see, e.g., Andrews, 1997), but may also be due to differences in the vocabulary items used to calculate OLD20 in each language.

Overall, our results are consistent with the idea that, across languages, a higher OND may facilitate familiarity checking, lexical access, or some other form of lexical verification or identification that occurs during reading. We hope that future work will be able to disentangle these possibilities.

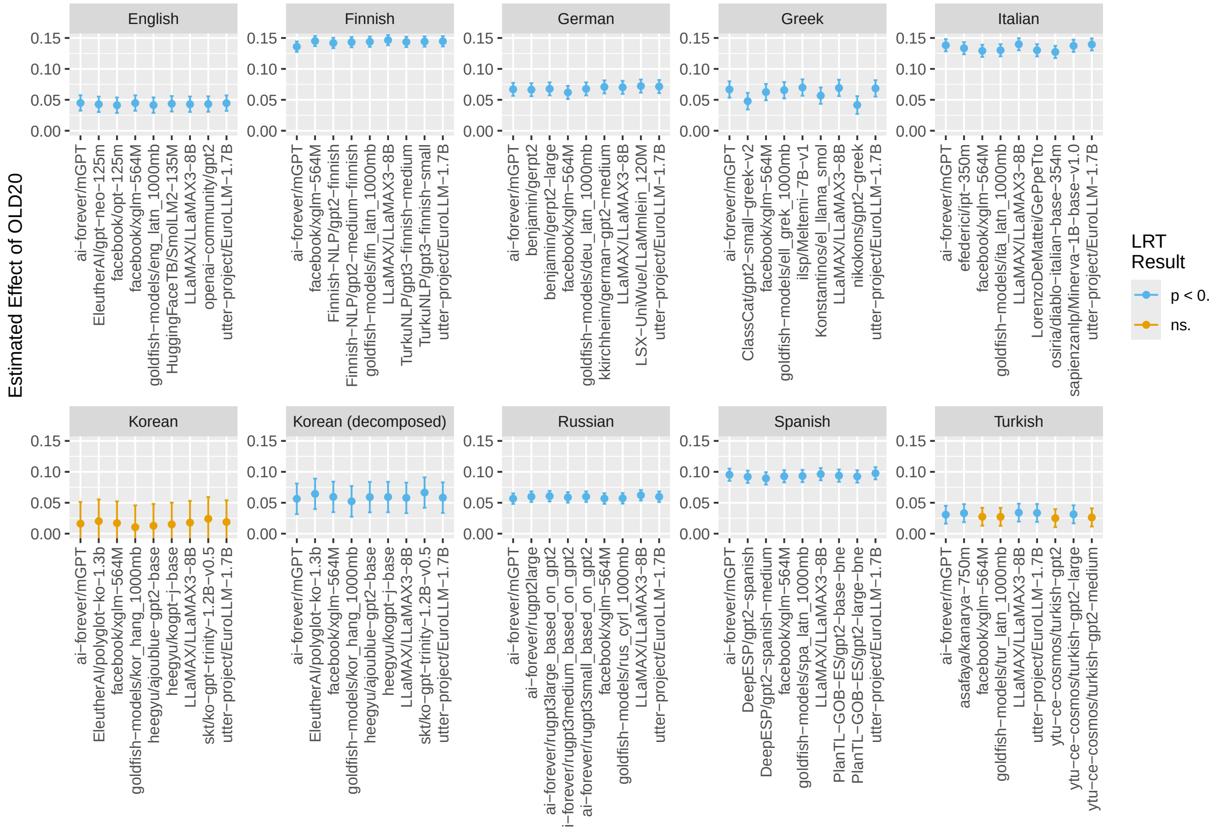

Figure 1: The estimated effects (coefficients and their standard errors in the linear mixed-effects regressions) of OLD20 on go-past duration. Note that the regression included *z*-scored OLD20 and *z*-scored log-transformed go-past durations, so these coefficients are not directly interpretable. For each language, experiments were run with the surprisal calculated from 9 language models, as labelled on the x-axis. Blue indicates that adding OLD20 to a regression already including surprisal and the baseline predictors improved fit, while orange indicates regressions for which it does not.

References

Andrews, S. (1997). *Psychonomic Bulletin & Review,* **4** (4), 439–461.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). In: *Attention and Performance VI*. Routledge.

Holcomb, P. J., Grainger, J., & O'Rourke, T. (2002). *Journal of Cognitive Neuroscience,* **14** (6), 938–950.

Lima, S. D. & Inhoff, A. W. (1985). *Journal of Experimental Psychology: Human Perception and Performance,* **11** (3), 272–285.

Pollatsek, A., Perea, M., & Binder, K. S. (1999). *Journal of Experimental Psychology: Human Perception and Performance,* **25** (4), 1142–1158.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). *Psychological Review,* **105** (1), 125–157.

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H.-D., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., Da Fonseca, S. M., Dirix, N., Duyck, W., Fella, A., Frost, R., Gattei, C. A., Kalaitzi, A., Kwon, N., Lõo, K., Marelli, M., Papadopoulos, T. C., Protopapas, A., Savo, S., Shalom, D. E., Slioussar, N., Stein, R., Sui, L., Taboh, A., Tønnesen, V., Usal, K. A., & Kuperman, V. (2022). *Behavior Research Methods,* **54** (6), 2843–2863.

Speer, R., Chin, J., Lin, A., Jewett, S., & Nathan, L. (2018).

Tsai, J., Lee, C.-Y., Lin, Y.-C., Tzeng, O., & Hung, D. (2006). *Language and Linguistics,* **7** (3), 659–675.

Yao, P., Staub, A., & Li, X. (2022). *Psychonomic Bulletin & Review,* **29** (1), 243–252.

Yarkoni, T., Balota, D., & Yap, M. (2008). *Psychonomic Bulletin & Review,* **15** (5), 971–979.