Filling in sparse data: Estimating missing response times for morphologically complex words

Response times (RTs) have been instrumental in advancing psycholinguistic research, and today researchers have access to millions of RTs. While this sounds like a substantial amount of data, the actual word coverage is small. As an illustration, consider the English Lexicon project (ELP; [1]), which contains RTs from more than 800 participants to over 40,000 words in a lexical decision task, and contrast it with the approximately 160,000 entries in the SUBTLEXUK [7]. This equates to missing RT values for 75% of the entries in SUBTLEXUK. The missing data problem is particularly pressing for morphologically complex words [4,6]. For example, for *pack* the ELP provides RTs for only 17 word-forms. However, it lacks RTs for the other 42 word-forms associated with *pack* in the Merriam-Webster dictionary. Since obtaining RTs for missing words is not economically viable, we tested two approaches to predict RTs for morphologically complex words: a generalised linear model (GLM) and a transformer-based [5] neural network (NN).

To maximize word-form coverage, RTs were obtained from the ELP and the BLP (British Lexicon Project; [3]) and word frequency information was operationalized as contextual diversity [2]. In the GLM, the RT for a multimorphemic word is given by:

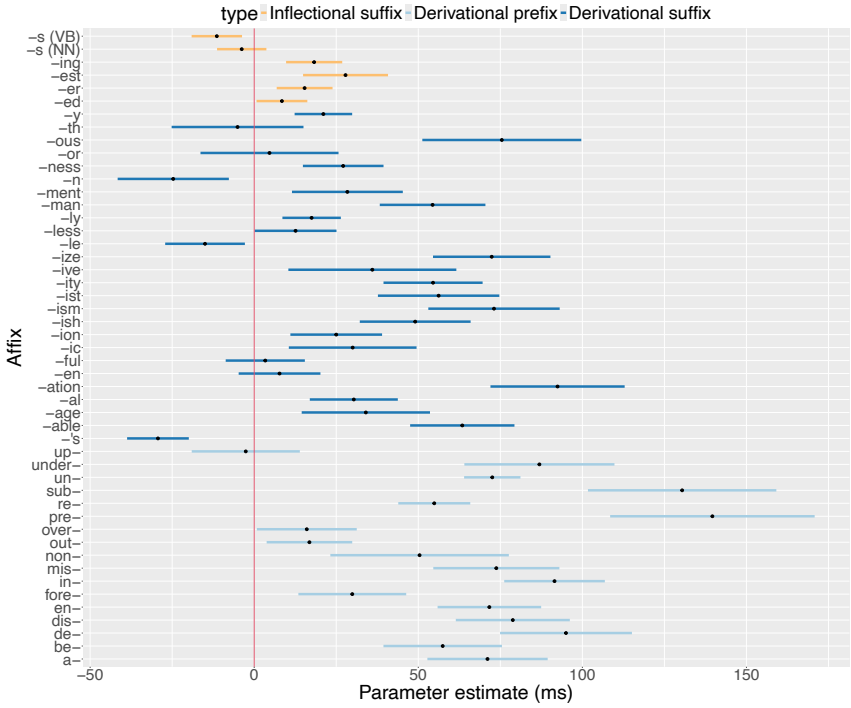
$$RT_{\text{morph}} = 122.414 + 0.805\, RT_{\text{stem}} - 35.55\, \Delta WF + 5.115\, \Delta L + 14.923 \times 1_{\text{ELP}} + 0.018 RT_{\text{stem}} \times 1_{\text{ELP}} + RT_{\text{pre}} + RT_{\text{suf}}.$$

We see that increasing the stem RT (RT_{stem}) and word length difference (ΔL) between the morphologically complex word and its stem increases RT_{morph}. In contrast, increasing the difference in word frequency (ΔWF) between the morphologically complex word and its stem decreases RT_{morph}, and words from the ELP are responded to more slowly than those from the BLP. Figure 1 shows that suffixes (RT_{suf}) generally increased RTs less than prefixes (RT_{pre}), and that the difference between derivational and inflectional suffixes seems to be small. A crucial result is that the probability distribution that best fits the data comes from the Box-Cox family. For the NN, explicitly providing frequency information to a text-embedding only implementation improved model performance only marginally, strongly suggesting that the NN already extracts relevant frequency information from pure text-embeddings. A direct comparison of the GLM with the NN using cross validation revealed a similar mean absolute error (MAE) of 63.59 $\pm$ 0.52ms and 60.66 $\pm$ 0.43ms, respectively. To put this into perspective, the MAE for a mean-only baseline model is 94.90 $\pm$ 0.62ms.

Taken together, our work provides methods for estimating RTs for morphologically complex words, thus sidestepping the challenging task of collecting RTs across morphological families. The GLM, which contains only a small number of predictors and uses a fraction of the computational resources of the NN, exhibits a similar predictive performance to the NN while yielding interpretability of the factors that shape RTs. On the other hand, the NN can predict RTs when the GLM lacks some of the covariates of morphologically complex words. The similar performance of GLM and NN suggests that the NN may inherit some of the explainable nature of the GLM while maintaining its flexibility.

Figure 1

Parameter estimates and 95% confidence intervals for all affix specific contributions.



References

[1] Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. Behavior Research Methods, 39(3), 445-459. https://doi.org/10.3758/BF03193014

[2] Johns, B. T. (2021). Disentangling contextual diversity: Communicative need as a lexical organizer. Psychological Review, 128(3), 525-557. https://doi.org/10.1037/rev0000265

[3] Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. Behavior Research Methods, 44(1), 287-304. https://doi.org/10.3758/s13428-011-0118-4

[4] Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. Behavior Research Methods, 50, 1568-1580. https://doi.org/10.3758/s13428-017-0981-8

[5] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *arXiv [cs.CL]*. http://arxiv.org/abs/1910.01108

[6] Stevens, P., & Plaut, D. C. (2022). From decomposition to distributed theories of morphological processing in reading. Psychonomic Bulletin & Review, 29(5), 1673-1702. https://doi.org/10.3758/s13423-022-02086-0

[7] van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A New and Improved Word Frequency Database for British English. Quarterly Journal of Experimental Psychology, 67(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521