Erratum for "Using LSTMs to Assess the Obligatoriness of Phonological Distinctive Features for Phonotactic Learning"

Nicole Mirea[*] and **Klinton Bicknell**[‡,*]

[*]Northwestern University [‡]Duolingo
nimirea@u.northwestern.edu
klinton@duolingo.com

The authors note that the subset of the CELEX2 database that used for the experiment reported in the paper contained repeated wordforms. The training, test, and validation sets, therefore, were not non-overlapping. The authors have repeated their analyses using non-overlapping data sets; although the argumentation of the paper remains identical, the authors note the following quantitative differences:

- In Experiment 1, all 38,882 unique wordforms from the CELEX2 lemma database were used to train and test the model. 23,330 (60%) of these lemmas were used to train the model, and the remaining 40% were randomly divided into validation and test sets of 7776 lemmas each.
- The effect size for Experiment 1 is $W = 1.90 \times 10^{10}; p < .05$. On average, the feature-aware models assigned a log likelihood of -21.27 to the words in the test set, and the feature-naive models assigned an average log likelihood of -21.21.
- The correlations with human-derived data in Experiment 2 ranged from 0.46 to 0.81, and the effect size of the difference according to the Wilcoxon rank sum test was $W = 365; p = 0.32$.

All corrected figures are included below.

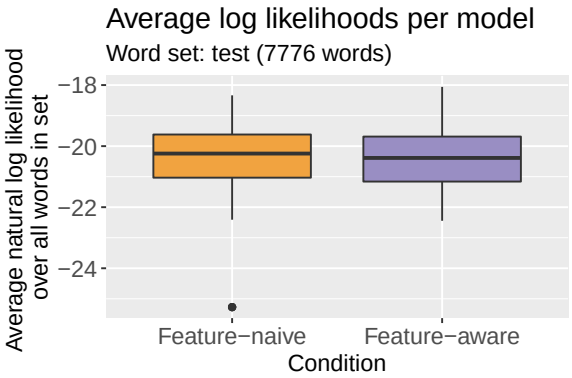

Figure 1: Box-plot of log likelihoods per model in each experimental condition. Each observation used to generate this plot ($N = 50$) is the average log likelihood assigned to each word in the test set, for a single model.

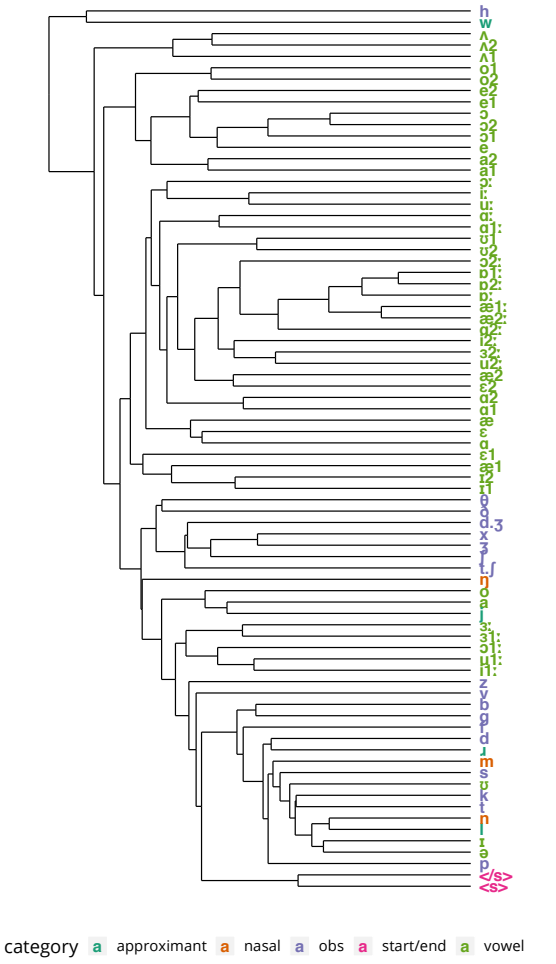

Figure 2: Dendrogram created using agglomerative clustering on trained embeddings from the feature-aware model that achieved the highest log likelihood on the test corpus. <s> and </s> signify start- and end-of-word symbols, respectively, and numbers after vowels indicate primary (1) and secondary (2) stress.

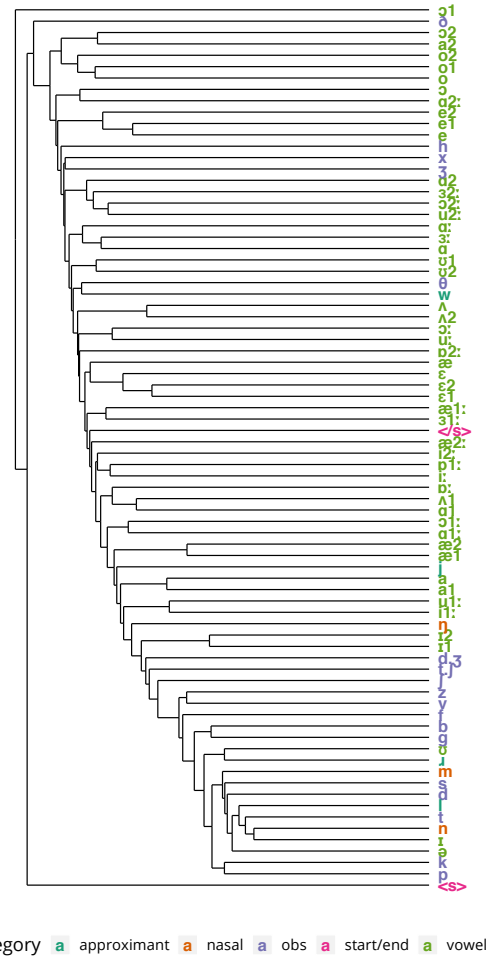

Figure 3: Dendrogram created using agglomerative clustering on trained embeddings from the feature-naive model that achieved the highest log likelihood on the test corpus.

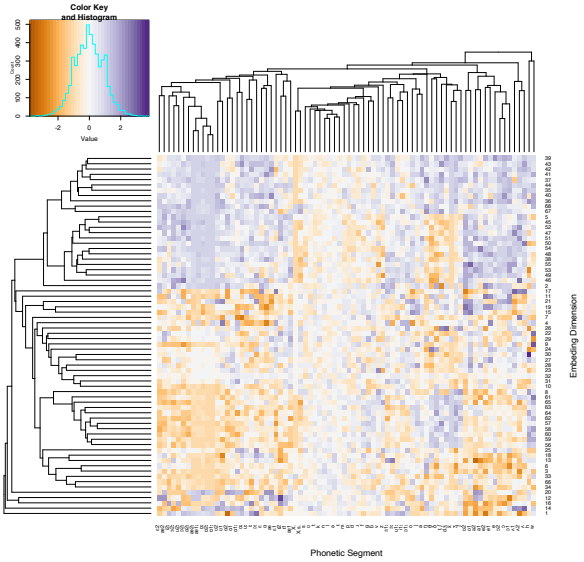

Figure 4: Heatmap of trained embeddings created from the feature-aware model that achieved the best performance on the test set. Clusterings along top axis are based on trained embeddings of each segment.

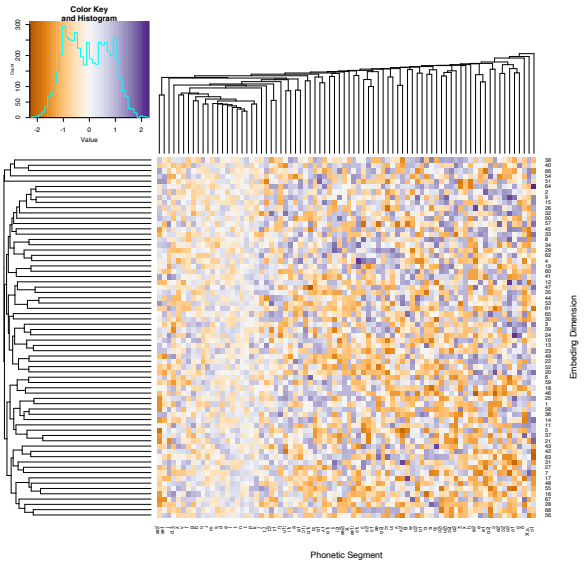

Figure 5: Heatmap of trained embeddings created from the feature-naive model that achieved the best performance on the test set.