PHONOLOGICAL CUES TO SEMANTIC CLASS MEMBERSHIP ACROSS HUNDREDS OF LANGUAGES

PABLO CONTRERAS KALLENS[*1] and MORTEN H. CHRISTIANSEN[1]

[*]Corresponding Author: pc684@cornell.edu
[1] Department of Psychology, Cornell University, Ithaca, NY, USA

The cultural evolution perspective suggests that human language is primarily a product of linguistic adaptation to a variety of cognitive, communicative, and social constraints, rather than the result of biological adaptations (e.g., Beckner et al., 2009; Christiansen & Chater, 2008). But how is it possible to acquire complex language without neural mechanisms dedicated to this purpose? One suggestion is that languages, via cultural evolution, "recruit" various types of cues to facilitate learning and use. This implies that all languages should incorporate some constellation of cues (mostly probabilistic in nature) that make them easier to acquire and use (Christiansen, 2013).

Here we empirically explore this multiple-cue approach to language evolution by considering the problem of categorizing words according to their meaning, focusing on the basic distinction between words for actions (typically verbs) and words for objects (typically nouns) (Vigliocco, Vinson, Druks, Barber, & Cappa, 2011). Previous studies have shown that phonological cues, whereby words with similar meanings have some degree of sound similarity (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015), can be observed across a range of languages (Dautriche, Mahowald, Gibson, & Piantadosi, 2017) and can be used for lexical categorization (Monaghan, Christiansen, Farmer, & Fitneva, 2010).

However, most of the previous studies have focused on languages from industrialized and/or literate parts of the world, covering only a few language families. Our analysis uses the word lists available in the Intercontinental Dictionary Series (Key & Comrie, 2015) to show that phonological cues to word class are available in a variety of language families and geographical zones. This database includes phonological transcriptions for more than 200 languages from 59 families, with a median of 911 words per word list. Because the transcriptions for each language in the IDS are not readily comparable, all results are *within*

language. We used information from Concepticon (List, Greenhill, Rzymski, Schweikhard, & Forkel, 2019) to determine the broad semantic class of a word: action, object or other. We focused on this semantic distinction because it is more fundamental compared to lexical categories such as verbs and nouns. Using a normalized version of the Levenshtein edit distance (Yujian & Bo, 2007), we computed the mean distance to actions and to things for each word by averaging over all within-language pairwise distances. The difference between the former and the latter is then a measure of the a word's *phonological typicality*.

To assess the potential effect of morphological markers, we assigned languages where more than a third of the words in a class share the same final or initial phoneme (167 of the 227 languages) to a "marker" group. A two-sample Wilcoxon rank-sum test shows that the difference in typicality between actions and objects is significant for 162 of the languages with markers and 22 out of the 59 languages without markers, Bonferroni adjusted $p < .001$.

To test whether these phonological cues are useful for learning, we trained an iRNN-type (Le, Jaitly, & Hinton, 2015) recurrent neural network with a 10-unit hidden layer to learn to classify the sequences of phonemes in a word as either an action or a thing. The performance of the network was assessed through Matthews Correlation Coefficient (MCC), where chance is 0 and perfect prediction is 1. Each network's MCC was cross-validated using a 10-fold scheme. Figure 1 shows all the within-language median MCC scores as a function of marker vs no-marker group. A within language one-sided Wilcoxon ranked sum shows that 158 of the 167 languages of the marker group and 27 out of the 45 languages in the no-marker group have an MCC higher than chance.

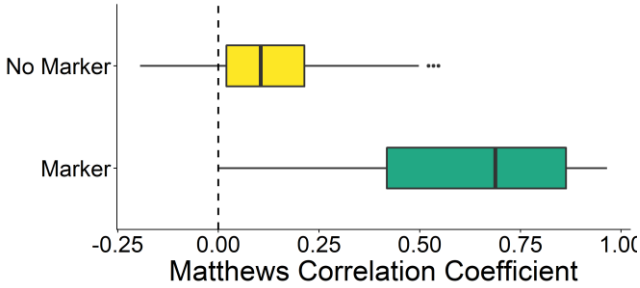

Figure 1 All within-language median MCC scores.

In conclusion, there is strong evidence that a great variety of languages encode broad semantic distinctions in subtle but useful phonological patterns, beyond the potential effects of morphological markers. This provides further evidence that languages have evolved to facilitate key parts of their acquisition.

References

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., … Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, *59*, 1–26.

Christiansen, M. H. (2013). Language has evolved to depend on multiple-cue integration. In R. Botha & M. Everaert (Eds.), *The evolutionary emergence of language: Evidence and Inference* (pp. 42–61).

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(05), 489-558.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*(8), 2149–2169.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Key, M. R., & Comrie, B. (2015). *The Intercontinental Dictionary Series* [Data set]. Retrieved from http://ids.clld.org

Le, Q. V., Jaitly, N., & Hinton, G. E. (2015). A simple way to initialize recurrent networks of rectified linear units. *ArXiv:1504.00941 [Cs]*. Retrieved from http://arxiv.org/abs/1504.00941

List, J. M., Greenhill, S., Rzymski, C., Schweikhard, N., & Forkel, R. (2019). *Concepticon 2.1.0* [Data set]. Retrieved from http://concepticon.clld.org

Monaghan, P., Christiansen, M. H., Farmer, T. A., & Fitneva, S. A. (2010). Measures of phonological typicality: Robust coherence and psychological validity. *The Mental Lexicon*, *5*(3), 281–299.

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(3), 407–426.

Yujian, L., & Bo, L. (2007). A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1091–1095.