THE EVOLUTION OF REDUNDANCY IN A GLOBAL LANGUAGE

GARY LUPYAN JUSTIN SULIK
Department of Psychology. University of Wisconsin-Madison
lupyan@wisc.edu jsulik@wisc.edu

Why are there different languages? Religious mythology aside, the usual story is that languages diverge when an initial group of speakers disperses, allowing their ways of speaking to begin to drift independently, instead of together (Sapir, 1921). But consider the explanatory inadequacy of such a *neutral drift* account if applied to a biological organism. Why do birds have different beaks? Because there is random drift in beaks shapes and once a population of birds disperses, their beaks drift independently, instead of together. When explaining animal morphology, we often appeal to adaptive fit: some beaks are better suited for some environments than others. Might similar logic apply to languages as well? Might linguistic diversity reflect, in part, the adaptation of languages to different environments in which they are learned and used?

A version of this idea—the linguistic niche hypothesis—was tested by Lupyan and Dale (2010). The authors reasoned that while all natural languages must be learnable by infants, some languages (e.g., English) are further constrained to be learnable by adults. Insofar as some grammatical paradigms (e.g., complex morphology) are more difficult for adults to learn, languages with many adult learners may become adapted to be more learnable by adults via simplification of their morphological paradigms. Examining relationships between morphological complexity and the socio-demographic niches of different languages confirmed the hypothesis: languages spoken by more people (those with more nonnative speakers) have simpler morphological paradigms.

But might these difference in morphology reflect a more fundamental property? A possibility proposed by Lupyan and Dale (2010) is that languages selected to be *only* learnable by children are selected to be morphologically complex because morphology (e.g., agreement systems, obligatory markings of tense, etc.) increases redundancy, redundancy which may pose challenges for

adult learners (Dale & Lupyan, 2012) but this redundancy may facilitate learning by children by providing additional cues for cohering the linguistic signal.

The current work tests the idea that languages primarily learned by infants have greater informational *redundancy* than languages with more nonnative speakers/adult learners. First, we took advantage of a published dataset for 11 Indo-European languages (Piantadosi, Tily, & Gibson, 2011) from which we could compute the average informativeness of each word based on an N-gram model. The results showed that languages spoken by more people / those with more nonnative speakers, had much higher informativeness per word (lower redundancy), r=.84, p<.001. These differences in redundancy are in line with theoretic predictions, are confounded by numerous differences between the languages. We undertook a stronger test of our hypothesis by comparing variants of two variants of English: American (AmEng) and British English (BrEng), which differ in the environment in which they are learned and used. AmEng is learned by more adult learners than BrEng (e.g., ~95% of BrEng are native speakers, but only 80%-85% of AmEng speakers are).

We trained N-gram language models on corpora of spoken AmEng (COCA) and spoken BrEng (BNC). We then tested the models both on withheld subsets of the training corpora, and on entirely new corpora created from scripts of American and British TV shows. The results show that for a large set of starting conditions, models of AmEng texts have overwhelmingly higher perplexity (lower redundancy) than BrEng texts and that models trained on AmEng are more generalizable, hinting at the greater compositionality of AmEng.

The work provides prima-facie evidence that linguistic divergence is not random. Knowing about the (social) environment in which a language is learned and used allows us to predict its current structure and possible future trajectory. These results point to specific aspects of language that may be under active selection through cultural evolution, helping us understand the circumstances that led to languages having their current structure.

References

Dale, R. A., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems*, *15*(3), 1150017.

Lupyan, G., & Dale, R. A. (2010). Language Structure Is Partly Determined by Social Structure. *PLoS ONE*, *5*(1), e8559.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *PNAS*, *108*(9), 3526–3529.

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. Dover Publications.