SELECTIVE FORCES IN LANGUAGE EVOLUTION

Juan A. Guerrero Montero*[1], Andres Karjus[2], Kenny Smith[1], and Richard A. Blythe[1]

*Corresponding Author: j.a.guerrero-montero@sms.ed.ac.uk
[1]Centre for Language Evolution, University of Edinburgh, Edinburgh, United Kingdom
[2]School of Humanities, Tallinn University, Tallinn, Estonia

The quantitative study of language evolution and change has seen great development, with models like Iterated Bayesian Learning (Griffiths & Kalish, 2007) and the Utterance Selection Model (Baxter et al., 2006) admitting analytical and computational explorations of fundamental questions in the field (Kirby, 2000; Thompson et al., 2016; Nowak et al., 2001). With the advent of massive digital historical corpora, data-driven methods add a new empirical perspective on the fundamental forces that shape human language over long timescales.

A key question in evolutionary linguistics is how competition between linguistic variants shapes linguistic structure. The Wright-Fisher model from population genetics provides a simple but powerful description of competition, accounting for both stochasticity in inter-generational transmission and selective forces (Crow & Kimura, 1970). This model is justified for its use in linguistic scenarios through its derivation from Iterated Bayesian Learning (Reali & Griffiths, 2010).

Nevertheless, the Wright-Fisher model is not easy to apply to corpus data since approximations are necessary for computational efficiency. These typically assume a normal distribution, which is inaccurate in certain scenarios (Karjus et al., 2020). Here, we introduce a robust algorithm based on the Beta-with-Spikes (BwS) approximation to transition probabilities in the Wright-Fisher model, whose analytical form is given by Tataru et al. (2015). In figure 1.A, we illustrate the nuance and robustness of this method by comparing to other algorithms in the analysis of a reference data set, namely competition between past forms of verbs from the Corpus of Historical American English (COHA).

We further apply the BwS method to two data sets extracted from the Google Books corpus (Michel et al., 2011). We first quantify the effect of competing motivations on language structure. We analyse the competition between inflectional simplicity (i.e. regularity) and phonological simplicity in the past forms of English verbs whose irregular forms coalesce to avoid repeating of /d/ or /t/ sounds (e.g. he knit instead of he knitted). The BwS algorithm quantifies the net force arising from these competing motivations through a selection strength parameter, revealing that selection towards phonologically simpler irregular forms is more

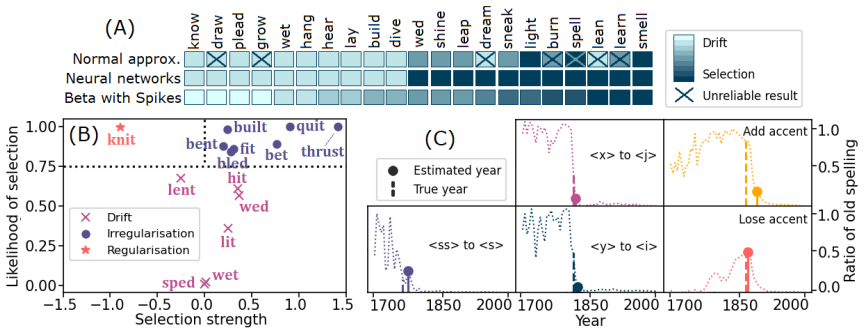

Figure 1. (A) Performance of BwS algorithm with those utilising normal approximations (Feder et al., 2014) and neural networks (Karsdorp et al., 2020), with reference to the likelihood of selection in competition between regular and irregular past forms of COHA verbs. BwS aligns with the robust neural network results, while providing a nuanced continuous classification instead of a binary one. Normal approximation is often unreliable due to non-normally distributed empirical frequency increments. (B) Application of BwS to competition between past tense verb forms that undergo coalescence in their irregular form. Positive selection strength implies favouring the irregular, phonologically simpler form, while the opposite is true for negative selection strength. Selection is concluded when likelihood is above 75%. Phonological simplicity beats inflectional simplicity in most cases, with half of all verbs in the data set showing high likelihood of selection towards irregularity, while only one shows regularisation. (C) Application of BwS to Spanish spelling reforms in the Google Books corpus. All reforms are detected with an error of 25 years or less, with three of them having an error of 10 years or less, even when trajectories are noisy (Real Academia Española, 1763, 1815, 1870).

common. If replicated in other domains, this result may shed light on how selection at different levels of linguistic structure might play out in language evolution.

Language is not only shaped by cognitive and internal factors, but also by social pressures that may change over time. This is illustrated in our second application, the dynamics of Spanish word spelling before and after reforms introduced by the Real Academia Española. Here, a new socially motivated bias appears that heavily favours a new spelling. This is reflected in the model by a change in the selection strength, thereby quantifying the level of acceptance of the reform by the literate population. Via likelihood maximisation, the BwS algorithm detects the year of introduction of the spelling reform with high precision (see figure 1.C). Thus this algorithm can detect and quantify major changes in social dynamics in the data, even when the exact date or origin of those changes is unclear.

In summary, the BwS algorithm applies to both socially and internally motivated competition, quantifies the net selective force arising from competing motivations and detects variations in this force over time. It provides a tool for the numerical analysis of diachronic linguistic data that is more insightful and reliable than previous methods. The continued development of such numerical tools opens the door to the empirical analysis of the social, cognitive and internal pressures that shape the structure of language.

References

Baxter, G. J., Blythe, R. A., Croft, W., & McKane, A. J. (2006). Utterance selection model of language change. *Phys. Rev. E*, *73*, 046118.

Crow, J., & Kimura, M. (1970). *An introduction in population genetics theory.* New York: Harper and Row.

Feder, A., Kryazhimskiy, S., & Plotkin, J. (2014). Identifying signatures of selection in genetic time series. *Genetics*, *196*(2), 509–522.

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*, 441–480.

Karjus, A., Blythe, R., Kirby, S., & Smith, K. (2020). Challenges in detecting evolutionary forces in language change using diachronic corpora. *Glossa*, *5*.

Karsdorp, F., Manjavacas, E., Fonteyn, L., & Kestemont, M. (2020). Classifying evolutionary forces in language change using neural networks. *Evolutionary Human Sciences*, *2*.

Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight (Ed.), *The evolutionary emergence of language: Social function and the origins of linguistic form* (pp. 303–323). Cambridge University Press.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Nowak, M., Komarova, N., & Niyogi, P. (2001). Evolution of universal grammar. *Science*, *291*(5501), 114–118.

Real Academia Española. (1763). *Ortografía de la lengua castellana* (3 ed.). Madrid, Spain.

Real Academia Española. (1815). *Ortografía de la lengua castellana* (8 ed.). Madrid, Spain.

Real Academia Española. (1870). *Prontuario de ortografía de la lengua castellana en preguntas y respuestas* (1 ed.). Madrid, Spain.

Reali, F., & Griffiths, T. L. (2010). Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. *Proc. R. Soc. B*, *277*, 429-–436.

Tataru, P., Bataillon, T., & Hobolth, A. (2015). Inference under a wright-fisher model using an accurate beta approximation. *Genetics*, *201*, 1133–1151.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences*, *113*, 4530–4535.