Abstract submitted with preference for **contributed talk**

Modelling and corpus analysis of the co-evolution of linguistic forms and functions

Juan Guerrero Montero, Dan Lassiter, Rob Truswell, Richard A. Blythe
University of Edinburgh, UK

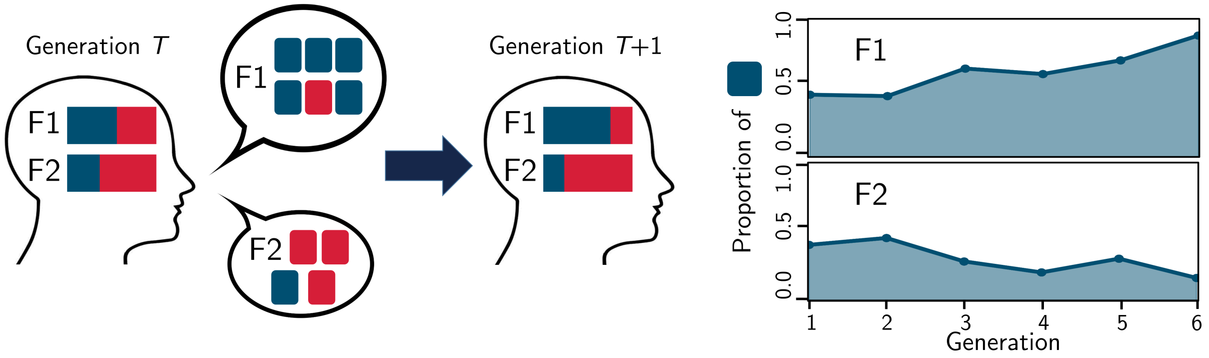
Cultural evolution and language change can be modelled as evolutionary processes, akin to those in population genetics. In these models, learning biases and social behaviours give rise to evolutionary forces such as drift, selection, and mutation. Language corpora and datasets of cultural norms and behaviours open the door to data-driven explorations of these evolutionary dynamics.

These models usually assume a simplified version of the cultural processes, in which several distinct but equivalent cultural expressions compete to represent a single, immutable cultural function. This picture is often incomplete in the field of language change, where expressions (words or structures) and functions (meanings) may be interconnected in more complex ways.

In the present work, we introduce a new model of language change that accounts for the many-to-many mappings between expressions and functions present in human languages. Based on the Iterated Bayesian Learning scheme, it is flexible enough to include effects ranging from learning biases to understanding error, selection, and mutation in production. In its data-analytic form, it takes the form of an evolutionary model that subsumes the well-established Wright-Fisher model.

We further present applications of this model to corpus data. In our analysis of the emergence of periphrastic do in Early Modern English, we are able to detect and quantify analogy as a significant evolutionary force by modelling it as mutation between functions. Our analysis of the usage of relativisers in Middle and Modern English shows that imperfect production and understanding during communication may drive forward the change of grammatical structures.

These results show that models incorporating co-evolving functions are relevant towards our empirical and quantitative understanding of language change. The model we introduce is a promising first step towards this.



Left: A cycle in our model for a system with two functions (F1 and F2) and two expressions (blue and red). Individuals in generation T use their grammar to produce utterances of each function. Individuals in generation $T+1$ use these utterances to learn the grammar. The model parametrises social and individual biases in the production and learning processes. **Right:** Corpus data of usage frequency of expressions and functions can be modelled as the result of the production phase of each of the generations in this cycle. Statistical techniques allow us to ascertain and quantify the biases in production and learning that have given rise to the observed behaviour.