This is the accepted version of a paper presented at Protolang 8, Rome, September 27-28, 2023.

Citation for the original published paper:

Johansson, S. (2023)
Simulating the spread and development of protolanguages
In: Protolang 8: Book of Abstracts (pp. 41-42).
Ways to (proto)language conference series

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:du-47919
Simulating the spread and development of protolanguages

Languages change over time, due to various processes that likely have been operative since the dawn of language. But our understanding of the relative importance of different processes in the distant past remains limited. Methods for reconstructing language change are hampered by shortage of training data.

Simulating language change in software can help, testing processes and producing simulated language data as input for reconstruction tests. In simulation, the processes are known and controllable, and the true diversification path is known. Tuning process strength in simulation until the results resemble real language diversity may inform theories of language dynamics.

But simulated data will only be helpful if the simulation reproduces relevant aspects of reality closely enough. Several items in List (2019) *Open problems in computational linguistics* concern simulation issues. Extant simulations are mainly of two types:

- Detailed short-term simulations of within-language dynamics, often agent-based (e.g. Nolfi & Mirolli, 2010).
- Macro-scale long-term simulations, but with linguistic and/or geographical details abstracted away (e.g. Wichmann, 2017; Kapur & Rogers, 2020).

Neither type covers the middle ground where within-language and between-language dynamics meet. This work aims to fill that gap, with a simulation that has sufficient linguistic, geographic and anthropological detail to produce realistic data, and sufficient scope to cover macro-scale dynamics over millennia.

The basic simulation unit is a speech community with typically 100-1000 speakers, speaking a common language. Their language has an explicit vocabulary with word-forms and meanings. Real languages from CLICS3 (Rzymski et al., 2019) are used as seed languages, which then evolve through regular sound change, word gain and loss, semantic shift, language contact, and areal effects. All processes are adjustable and can be disabled.

The geography of the real world is used, with topography from De Ferranti (2015), rivers from Kelso (2016) and climate/ecology from NASA (2016). Each speech community lives in a 50x50 km grid square, which may be shared with other communities up to a carrying capacity. Population may increase or decrease depending on food availability, and surplus population may migrate to greener pastures, forming a new community whose language then evolves independently. Travel depends on real terrain and available technology (innovations occur occasionally, starting from paleolithic level).

Simulation results are available as Swadesh matrices, or in formats suitable for automated reconstruction such as CLDF or NEXUS. True trees and true cognate sets are saved separately.

Software and sample output available at https://github.com/[ANONYMIZED]/LangChangeSimulator/tree/master

http://www.viewfinderpanoramas.org/dem3.html


