Simulating the spread and development of protolanguages

Sverker Johansson[*1]

[*]Corresponding Author: sja@du.se
[1]Dalarna University, Falun, Sweden

Simulation of language change can (1) provide training and validation data for language reconstruction, (2) explore protolanguage hypotheses where real data are unavailable, and (3) test the importance of different processes in language evolution. But previous simulation work is mainly either micro-scale or too abstract. Neither pole captures the full range of language dynamics. The aim here is to fill that gap between micro and macro. The simulation combines explicit models for language, demography, and geography in sufficient detail to produce output that in relevant respects mimics real language data, with adequate scale and scope to model language history across millennia.

1. Why simulation?

Languages change over time, but our understanding of the relative importance of different processes in the distant past remains limited. The development of methods for reconstructing language change is also hampered by a shortage of suitable data.

Simulating language change in software can help alleviate these problems (cf. List, 2019). Virtually unlimited amounts of simulated language data can be produced where the processes are known and controllable, and the true diversification path is known.

Furthermore, tuning process strength in simulation until the results resemble real language diversity may inform theories of language dynamics, within the limits set by the problem of equifinality (Kandler & Powell, 2018). Early forms of protolanguages may be simulated by adding suitable constraints to the simulation (cf. Gong et al., 2022).

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 dynamics within a single language, often agent-based (e.g. Cangelosi & Parisi, 2002; Nolfi & Mirolli, 2010, as well as many Evolang entries over the years, e.g. Wang & Steels, 2008).
- Macro-scale long-term simulations that cover the dynamics between languages, but with linguistic and/or geographical details abstracted away (e.g. Hochmuth et al., 2008; Wichmann, 2017, 2021; Kapur & Rogers, 2020; Ciobanu & Dinu, 2018; Gergel et al., 2021).

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

This simulation is not tailored to test specific hypotheses, but to produce simulated language data that is useful for further processing and testing.

2. Simulation framework

The simulation contains the following core models:

- Explicit *geography* model.
    - Topography, climate, vegetation.
    - Climate change.
- Explicit *population* model.
    - Population growth and decline.
    - Migration & population split.
    - Population interactions.
- Explicit *language* model.
    - Within-language processes.
    - Between-language dynamics.
- Technological development.

The basic simulation unit here is a speech community with typically 100-1000 speakers, speaking a common language. To initialize the simulation, real languages 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. Lexical data for seed languages are taken from CLICS3 (Rzymski et al., 2019), and grammatical data from Grambank (Skirgård et al., 2023).

The geography of the real world is used, with topography from De Ferranti (2015), rivers from Kelso (2016), climate/ecology from NASA (2016) and climate change from Snyder (2016). Each speech community lives in a grid square (default size 50x50 km, which is also the resolution of the geographic model) which may be shared with other communities up to a carrying capacity. The carrying capacity depends on climate, vegetation, and access to water, and may fluctuate from year to year (modelling drought etc.).

The population of each community may increase or decrease over time, depending on food availability, and surplus population may migrate to greener pastures, forming a new community. The new community speaks a clone of the parent community language, but their languages then evolve independently. The distance travelled in migration depends on real terrain and available technology.

Technological innovations occur occasionally, starting from a paleolithic level. An innovation may increase food production in some or all environments, open up new environments to exploitation, or enhance mobility. Some innovations are prerequisites for others. One community may learn a technology from a neighboring community. The technological model reaches Bronze Age technology, and is inspired by the computer game Civilization (Meier, 2021). This allows communities to evolve from hunter-gatherers to horticulturalists, pastoralists, and farmers.

3. Language model

Each language in the simulation has an explicit vocabulary. The word forms are strings of phonemes, each paired with one or more meanings. Each meaning may be covered by zero, one, or more words. Grammar is modelled using typological parameters from GramBank (Skirgård et al., 2023).

New languages are born when a community splits in two. Languages can die in two different ways: the whole community may starve, or the community may be assimilated into a more powerful neighboring community.

The processes affecting a language can be divided into *endogenous* processes internal to it, and *exogenous* processes due to contact with other languages.

3.1 Endogenous processes

Regular sound change is modelled as one random phoneme in the language being replaced by another phoneme with similar features, at random points in time at some rate. Sound change may be either unconditional or conditional. Words may

undergo metathesis, swapping two nearby phonemes in the word. Other sound change processes are not implemented yet.

Semantic shifts are implemented using colexification data from CLICS3 (Rzymski et al., 2019). The meaning of a word may be broadened to include another meaning, with a rate proportional to the colexification rate between the original and the new meaning. A word may also lose semantic scope, especially if it has synonyms in one of its senses.

Grammatical change is modelled as random changes in typological parameters, with care taken to keep the resulting grammar consistent.

3.2 Exogenous processes

Languages that are in contact regularly borrow words from each other. Terrain and travel technology affect which languages are regarded as in contact. Long-range borrowing beyond the regular travel range happens occasionally. Borrowing rate is enhanced in the following cases:

- Two languages occupy the same grid square.
- Two languages are closely related (either short time since split or short lexical distance between vocabularies).
- One language lacks a word for a concept.

When technology is transferred between communities, the vocabulary for the new technology is also borrowed.

If multiple languages are present in the same grid square, minorities are heavily affected by the most powerful group. For each generation, some fraction of the minority will shift to the majority language.

3.3 Areal effects

Areal effects are exogenous processes, where language features spread broadly in a region so that unrelated neighboring languages come to resemble each other. This is modelled for sounds, words, and grammatical features in roughly the same way: if a large fraction of the languages in a region share the same feature, the languages that don't have it are likely to adopt it.

4. Flexibility

All processes in the model can be switched on and off under user control, either through runtime switches or through parameter files. All processes can also have their rates adjusted through parameter files. The parameter space is by necessity unexplorably large in any detailed simulation (30-odd primary parameters in this

case, plus huge transition matrices), but having the parameters visible instead of fixed in code makes this issue explicit. Both time step and total running time for the simulation can be chosen at runtime.

A smaller geographical region than the whole world can be chosen at runtime, in order to run contact-rich scenarios. It is also possible to create and load alternative geography data, in order to test hypotheses about the importance of geographical structure.

Seed languages can either be selected from a list, or a random selection of a given number of languages can be provided.

5. Output

Simulation results are available in several different formats:

- Language metadata: seed, birth year, birth place, death year
- Swadesh matrices (tab-separated text; can be read by e.g. MS Excel).
- Word lists in CLDF format.
- Grammar data as list of typological features.
- True phylogenetic tree in NEXUS format
- True cognate lists

Simulation results and the underlying true data are consistently saved in separate files, that can be cross-referenced with unique identifiers.

6. Validation

The aim of the simulation is to produce data that mimic patterns in real language data in relevant respects. On visual inspection, the output generally looks plausible, though with some unusual phoneme sequences; phonotactic constraints are not implemented. But visual impression is of course not a scientific validation.

Comparing statistical patterns in the output with real language data is a more reliable validation method than visual impressions. In Figure 1 below is one example. The similarity between word forms is quantified using weighted Levenshtein (1966) distances. Between unrelated words, the distances should be randomly distributed, whereas cognates can be expected to have smaller distances. In each part of Figure 1, the dashed curve is for words from different language families, and the solid curve is for words with the same meaning within the same language family. The left diagram shows real data from CLICS3, the right is simulated data. The two diagrams are qualitatively similar but not identical; part of the difference may be due to the simulated families all having the same age. If the simulation is left running long enough (15,000+ years) the two curves will eventually converge, as cognates are no longer discernible.

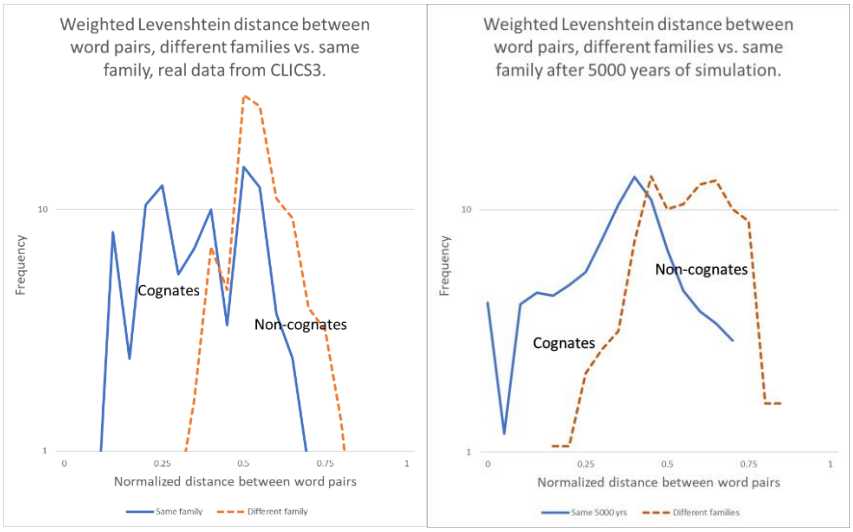

Figure 1. Levenshtein distance between word pairs, comparison between real and simulated data.

Another validation option is to use the output from the simulation as input to the same phylogenetic reconstruction methods that are used in computational historical linguistics with real data (cf. Jäger, 2019). This works fine, the standard program PAUP (Swofford, 1996) for phylogenetic reconstruction recovers the true tree with a plausible level of accuracy. Automated cognate detection likewise recovers the true cognate sets from simulated data with reasonable accuracy.

7. Summary

The simulation basically works as intended, producing reasonable-looking data in large quantities. Some tuning work is still needed, but the model passes basic validation tests. The output works as input to phylogenetic reconstruction.

The software runs on a regular PC, generating thousands of languages over thousands of years in a matter of hours. But over very long time scales with very large numbers of languages, it will bog down computationally.

8. Supplementary Materials

Software and sample output available at
https://github.com/Lsjbot/LangChangeSimulator/tree/master

References

Cangelosi, A. & Parisi, D. (2002) *Simulating the evolution of language.* London: Springer.

Ciobanu, A. M. & Dinu, L. P. (2018) Simulating language evolution: A tool for historical linguistics. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 68–72 Santa Fe, New Mexico, USA, August 20-26, 2018.

De Ferranti, J. (2015) *Viewfinder Panoramas Digital Elevation Model.* http://www.viewfinderpanoramas.org/dem3.html

Dellert, Johannes. (2019) *Information-theoretic causal inference of lexical flow (Language Variation 4)*. Berlin: Language Science Press.

Gergel, R., Kopf-Giammanco, M., & Puhl, M. (2021). Simulating semantic change: A methodological note. *Experiments in Linguistic Meaning*, *1*, 184-196.

Gong, T., Shuai, L. & Yang, X. (2022). A simulation on coevolution between language and multiple cognitive abilities. *J Lang Evo* 2022:120-145

Hochmuth, M., Lüdeling, A., & Leser, U. (2008). *Simulating and reconstructing language change*. Unpublished manuscript, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Informatik. https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/research/publications/2008/tr_language_change.pdf

Jäger, G. (2019) Computational historical linguistics. *Theoretical Linguistics*, 45(3-4):151-182. https://doi.org/10.1515/tl-2019-0011

Kandler, A. & Powell, A. (2018) Generative inference for cultural evolution. *Phil. Trans. R. Soc. B*373:201700562.0170056 http://doi.org/10.1098/rstb.2017.0056

Kapur, R & Rogers, P (2020) *Modeling language evolution and feature dynamics in a realistic geographic environment.* In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona.

Kelso, N V (2016) *Natural Earth Data.* https://www.naturalearthdata.com/downloads/

Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. **10** (8): 707–710.

List, Johann-Mattis (2019): *Open problems in computational historical linguistics.* Invited talk presented at the 24th International Conference of Historical Linguistics (2019-07-01/05, Canberra, Australian National University).

Meier, Sid (2021) *Sid Meier's Memoir.* W W Norton.

NASA (2016) *NASA Earth Observations.* https://neo.gsfc.nasa.gov/

Nolfi, S & Mirolli, M (2010) *Evolution of Communication and Language in Embodied Agents.* Springer.

Rzymski, Christoph and Tresoldi, Tiago et al. (2019). *The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic poly-semies*. DOI: 10.1038/s41597-019-0341-x

Skirgård et al. (2023). Grambank v1.0 (v1.0.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7844558

Snyder, C. (2016) Evolution of global temperature over the past two million years. *Nature* **538**, 226–228. https://doi.org/10.1038/nature19798

Swofford, D. (1996). *PAUP*: Phylogenetic analysis using parsimony (and other methods), version 4.0*. Sunderland, MA: Sinauer Assoc.

Wang, E. & Steels, L. (2008) Self-interested agents can bootstrap symbolic communication if they punish cheaters, In Proceedings of Evolang 7.

Wichmann, S. (2017) Modeling language family expansions. *Diachronica 34:1, 79-101.*

Wichmann, S. (2021) A world language family simulation. *Physica A: Statistical Mechanics and its Applications*, 572: 125913. https://doi.org/10.1016/j.physa.2021.125913 .