CHIELD: CAUSAL HYPOTHESES IN EVOLUTIONARY LINGUISTICS DATABASE

Seán G. Roberts

sean.roberts@bristol.ac.uk
excd.lab, University of Bristol, Bristol, UK

Evolutionary linguistics is now a well established field with several conferences and its own journal. The ultimate goal of the field is to explain how complex communication systems emerge and change. A coherent, comprehensive explanation would involve a long chain of causal claims, stretching from genetics to cognition and from prehistorical adaptations to modern language change, supported by a range of methods from experiments to computer simulations. Because of the range of disciplines feeding into language evolution theories, producing such an explanation is a daunting task. In order to help this process, this paper presents a schema and implementation for a database of causal hypotheses about language evolution. Researchers can edit and contribute through a custom web application or through a GitHub repository.

1. Introduction

New databases and web technology are being used in many fields to synthesise knowledge. For example, the *D-PLACE* database (Kirby, Gray, Greenhill, Jordan, & al., 2016) integrates cultural, linguistic and phylogenetic data. Databases are also being created to collect hypotheses, too, such as the *Explaining Human Culture* database, a collection of over 3,000 hypotheses in cultural anthropology (Ember, 2016). Hypotheses are drawn from publications, and the database includes which variables were used, the statistical method and the main statistical results. The database is searchable by hypothesis or by variable, making it easy to find studies linking any two variables. Collections of studies like this can be used to guide research. For example, *Metalab* (Lewis et al., 2015) includes experimental results from 282 publications to support meta-analyses and power analyses in language acquisition paradigms. Collaboration tools are also helping to refine definitions and converge on hypotheses. For example, Glottolog (Hammarström, Forkel, & Haspelmath, 2017), a database of languages and language families, hosts its data on GitHub. Anyone can suggest edits and discuss issues in a simple web interface, allowing the research community to collaborate on maintaining and refining knowledge about linguistic history.

A similar resource for language evolution would be invaluable. The paper presents a schema and initial implementation for a database of causal hypotheses in evolutionary linguistics.

DOI:10.12775/3991-1.099

2. Motivation

The motivations for creating a database of causal hypotheses include:

Surveying the field. Language evolution is a very broad field, both in terms of scope and methods (Christiansen & Kirby, 2003), and surveying it is no easy task. Computational methods can help here (Bergmann & Dale, 2016), but the fundamental problem is simply the very large number of studies. Causal processes can be represented conveniently as graphical networks (Pearl, 2009), helping to visualise the field.

Converging on definitions. Part of the work of coding the database is to translate a hypothesis into an explicit series of causal links between variables. This forces transparent interpretations of theories and the use of common variable names. There will, of course, be disagreement on the interpretation of studies and on the terminology used for variables. However, if the debates can be centralised and directed towards concrete issues then this is a healthy process for a field.

Finding competing and supporting hypotheses. The database can identify competing explanations (alternative paths between variables or conflicting causal links). These are candidates for critical comparison studies. Similarly, the causal network could also identify evidence that supports a hypothesis, such as replications or tests using alternative methods. This aids a robustness approach to theory building (Irvine, Roberts, & Kirby, 2013).

Linking hypotheses together. The database could reveal some surprising links between theories, or identify missing or weakly supported links. It could also provide researchers with evidence for the preconditions for the topics they study, suggest wider downstream implications of their hypotheses or provide more detailed mechanisms that link higher-level concepts. Network analyses could identify 'broker' theories that bridge two areas. This would help extend theories and guide future research and collaboration.

Articulating causal processes. Even though causal arguments should be at the heart of any hypothesis investigation, coding articles for causal claims was often surprisingly difficult. Creating a visible framework for thinking about hypotheses as a network of causal processes will encourage more rigorous and transparent definitions of hypotheses. Using the schema below, it would be possible to publish a formal definition of the causal network alongside publications.

Research and teaching resource. The database will aid systematic literature review and provide an accessible entry point for students or researchers from outside of the field.

Given these motivations, there are several desiderata for a database of causal claims: it is openly accessible; the research community can contribute, edit and discuss issues; it should recognise contributors; causal claims can be represented visually and interactively; and the type of support for the claim should be coded; entries should be sourced widely and in an unbiased way.

3. Methodology

3.1. *Framework*

Causal claims can be represented as a directed graph (Pearl, 2009). Nodes represent variables and edges represent causal processes. The definition of variables is, at this point, left vague. This is because they might include a number of different kinds of concepts, depending on the research topic. For example, some variables might be concrete and measurable such as presence of a genetic allele, but others might represent higher-level concepts like a selection pressure for efficient communication. Also, variables might measure concepts on different scales, such as the age of an individual or the size of a population. While this is perhaps conceptually weak, in practice the interpretations are reasonably clear. Directed causal graphs can be easily visualised and analysed with a range of tools to discover weak, conflicting or supporting links (e.g. DAGitty, Textor, Hardt, & Knüppel, 2011).

3.2. *Sources*

The condition for entry into the database is that the hypothesis makes causal claims that relate to some part of the evolution of communication and that it is published in a peer-reviewed publication. Entry into the database does not mean that the hypothesis is correct nor widely accepted nor even empirically supported. The aim is not that the database be a single coherent, consistent theory of the evolution of communication, but a reflection of the field.

Existing digital databases will serve as initial sources of publications, such as the *Language Evolution and Computation Bibliography* (`http://groups.lis.illinois.edu/amag/langev/`), *the Universals Archive* (`https://typo.uni-konstanz.de/archive`), the EvoLang conferences (`http://evolang.org/neworleans/`) and relevant journals such as the *Journal of Language Evolution* and *Interaction Studies*. The research community can also contribute entries through a custom web application or directly through GitHub.

3.3. *Coding scheme*

An entry in the database encodes a single causal link between two variables. A minimal entry contains: bibtex reference for the source; label for variable 1; label for variable 2; type of causal relation; and the direction of the effect (positive or negative). A publication may be coded with multiple entries. The type of relation is drawn from table 1 (borrowing from the *lavaan* package in R, Rosseel, 2011).

The direction of the effect is necessary not only for interpreting the claim, but also so that causal claims from multiple studies can be integrated under the same variables (e.g. a process that increases morphological simplicity can be coded under a negative effect on morphological complexity).

Table 1. Causal relation syntax.

Syntax	Meaning	Syntax	Meaning
X > Y	A change in X causes a change in Y	X / > Y	X does not causally influence Y
X <=> Y	X and Y co-evolve	X >> Y	X is a necessary precondition for Y
X ∼∼ Y	X and Y are correlated	X = ∼ Y	X is an indicator of (measured by) Y

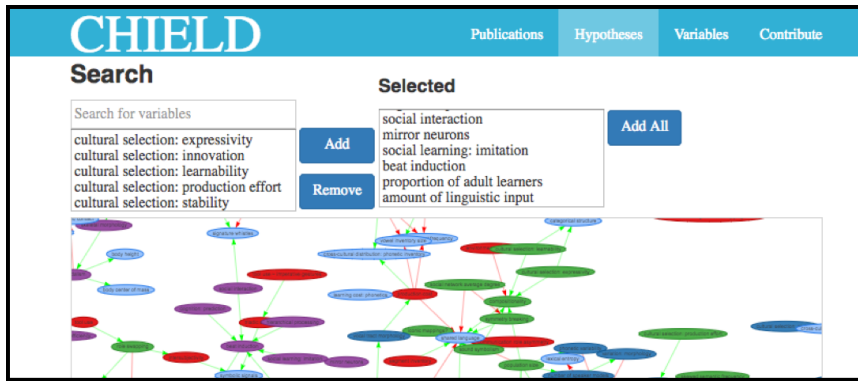

Figure 1. The current web interface for searching the database

Two studies might make claims about the same underlying concept, but measure it in different ways. To unify the theories, the causal link is represented as a latent variable: e.g. population size > morphological complexity (main link)
morphological complexity = ∼ presence of nominal case (indicator link)
morphological complexity = ∼ WALS feature score (indicator link)

Entries can be extended to include: Process: popular label for the process (e.g. "iterated learning"). Topic: e.g. phonetics, syntax etc. Stage: preadaptation, coevolution, cultural evolution, language change (Scott-Phillips & Kirby, 2010). Type: type of evidence (hypothesis (logical argument), review (other work), experiment, model, simulation). Subtype: subtype of evidence: e.g. iterated learning experiment, communication game etc. Confirmed: Whether the hypothesis was supported or not. Quote: A quote from the paper which states or clarifies the causal claim. Coder: Identity of the coder. These fields are important for the searchability of the database. For example, identifying the evolutionary stage at which the causal process applies helps to locate the link, but also to visualise the network of causal links.

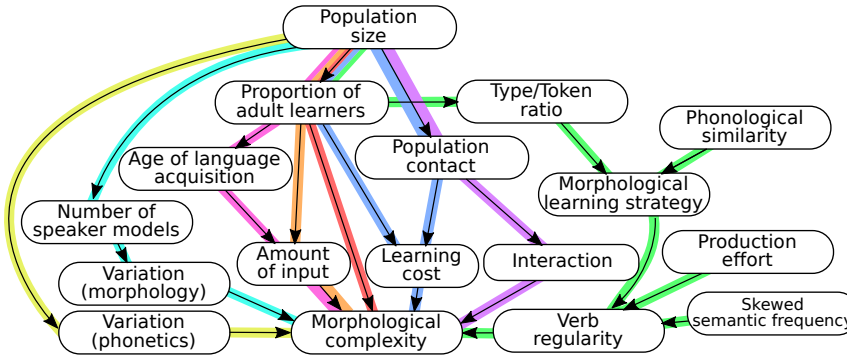

Figure 2. A sub-graph from CHIELD linking population size and morphological complexity. Yellow: Ardell, Anderson & Winter (2016); Cyan, Purple: Atkinson, Smith & Kirby (2016); Pink: Bentz & Berdicevskis (2016); Red: Bentz & Winter (2013); Blue: Lupyan & Dale (2010); Green: Cuskley & Loreto (2016).

4. Current implementation

The database currently contains 222 causal links from 50 publications. The web interface can be accessed at `http://chield.excd.org/`, and the GitHub repository is live at `https://GitHub.com/CHIELDOnline/CHIELD`. The current interface (figure 1) allows users to interactively visualise different parts of the causal network, and submit their own links through a graphical interface. Coding of new links is guided by the interface's suggestions of variable labels already present in the database, helping to unify hypotheses. The data is hosted openly on GitHub, which also provides tools for curation, editing and debate.

Figure 2 shows part of the CHIELD database linking population size and morphological complexity (with an example of part of the coded data in table 2). While Bentz and Winter (2013) make a direct link between the proportion of adult learners and morphological complexity, two other studies discuss the intermediate step of the amount of the amount of linguistic input.

Table 2. An example of some entries in the database, summarising Lupyan & Dale (2010).

Variable 1	Relation	Variable 2	Cor	Type	...
population size	>	proportion of adult learners	pos	statistical	...
proportion of adult learners	>	learning cost: morphology	pos	review	...
learning cost: morphology	>	morphological complexity	neg	statistical	...

5. Conclusion

This paper presented a schema and initial implementation of a database of causal hypotheses in evolutionary linguistics. Its aim is to provide an extendable resource for researchers. The major challenge is in the coding, both in terms of amount of time and coming to an agreement on interpretations and labels. The web interface tools and integration with GitHub are designed to address these challenges. However, there are also conceptual issues specific to language evolution (capturing arguments about the timing of the emergence of traits or properties such as population size having different connotations during preadaptation and language change). It is also unclear how observational work (e.g. animal communication) or arguments using evolutionary analogy fit in. Usefully visualising the network will also be a challenge, though there are many existing tools to help. Despite these difficulties, this paper argues that it is a worthwhile project which will has the potential for high impact in the field.

Acknowledgements

Supported by Leverhulme fellowship ECF-2016-435.

References

Ardell, D., Anderson, N., & Winter, B. (2016). Noise in phonology affects encoding strategies in morphology. In S. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 11th international conference (evolangx11).*

Atkinson, M., Smith, K., & Kirby, S. (2016). Adult language learning and the evolution of linguistic complexity. In S. Roberts & al. (Eds.), *The Evolution of Language: Proceedings of the 11th Conference (EVOLANGX11).*

Bentz, C., & Berdicevskis, A. (2016). Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence. In *Coling 2016* (pp. 222–232).

Bentz, C., & Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, *3*(1), 1–27.

Bergmann, T., & Dale, R. (2016). A scientometric analysis of evolang: Intersections and authorships. In S. Roberts & al. (Eds.), *The evolution of language: Proceedings of the 11th Conference (EVOLANGX11).*

Christiansen, M., & Kirby, S. (2003). Language evolution: The hardest problem in science? *Studies in the evolution of language*, *3*, 1–15.

Cuskley, C., & Loreto, V. (2016). The emergence of rules and exceptions in a population of interacting agents. In S. Roberts & al. (Eds.), *The Evolution of Language: Proceedings of the 11th Conference (EVOLANGX11).*

Ember, C. R. (2016). *Explaining human culture.* New Haven, Ct.: Human Relations Area Files. `http://hraf.yale.edu/ehc`.

Hammarström, H., Forkel, R., & Haspelmath, M. (2017). Glottolog 3.0. *Jena: Max Planck Institute for the Science of Human History. http://glottolog.org*.

Irvine, L., Roberts, S. G., & Kirby, S. (2013). A robustness approach to theory building. In *Proceedings of the 35th CogSci* (pp. 2614–2619).

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., & al. (2016). D-place: A global database of cultural, linguistic and environmental diversity. *PloS one*, *11*(7), e0158391.

Lewis, M., Braginsky, M., Bergmann, C., Tsuji, S., Cristia, A., & Frank, M. (2015). Metalab: A tool for power analysis and experimental planning in developmental research. In *Proceedings of 40th BUCLD.*

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.

Pearl, J. (2009). *Causality.* Cambridge university press.

Rosseel, Y. (2011). *lavaan: an r package for structural equation modeling and more version 0.4-9 (beta).* Ghent University.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in cognitive sciences*, *14*(9), 411–417.

Textor, J., Hardt, J., & Knüppel, S. (2011). Dagitty: a graphical tool for analyzing causal diagrams. *Epidemiology*, *22*(5), 745.