Causal inference for historical linguistics

Seán Roberts, Cardiff University, UK

Targeted at the workshop on *Ancient languages and Algorithms*

The aim of this paper is to explain the principles behind *causal inference* (e.g. Pearl & Mackenzie, 2019) and its application in current historical linguistics.

Explaining phenomena in historical linguistics increasingly depends on synthesising hypotheses and evidence from many different fields. But before any data is engaged with, there are three main challenges: identifying causal claims from prior research about the relationship between variables; translating the causal claims into testable hypotheses; and identifying potential confounding factors. Causal inference is an approach to theory building that provides practical tools to help address these challenges.

For example, there are many theories about a potential link between the morphological complexity of a language and the number of people who speak it. Different theories suggest a different chain of causal effects that connect the two variables. These range from chains involving differences in learning abilities of children and adults (Lupyan & Dale, 2010), to those involving variation at the phonological level (Ardell, Anderson & Winter, 2016). We would like to compare these theories, and others, to test which provide the best explanation. How can we do this in a formal and systematic way?

The suggestion from causal inference is that these theories can be represented as a causal graph – a graphical method that visualises variables as nodes and causal connections as arrows between them. Expressing theories explicitly in this way allows us to visualise agreement and conflict between theories, and to identify key causal claims that should be tested.

While causal graph visualisations are helpful in themselves, they are also underpinned by mathematical principles that can provide more insight. These principles can tell us what kinds of causal claims we can make from observational data. This differs from many previous approaches which assume that little about causality can be inferred from observation alone. Causal inference, then, provides a potentially revolutionary tool for historical linguistics where direct control and experimentation is not possible.

Causal inference methods also include methods for formally identifying confounds and to find ways of controlling for them. This means that causal graphs can provide the basis for formal statistical models, helping to make clear connections between theory and inference. However, there are also limitations, and understanding these is vital for understanding the role that causal inference can play in historical linguistics. This talk will cover the assumptions and limitations of causal inference and discuss their relation to historical linguistics methodologies.

The talk will illustrate the points above by showing how 20 hypotheses about population size and morphological complexity were coded and analysed to help design a statistical study. This was done with the help of CHIELD: a suite of web tools for tracking and comparing causal graphs in evolutionary linguistics (https://correlation-machine.com/CHIELD, Roberts et al., 2020). It will be argued that causal graphs, together with digital tools, can help researchers meet the three challenges above. Causal claims can be tracked in a database of causal graphs.

References

Ardell, D., Anderson, N., & Winter, B. (2016). Noise In Phonology Affects Encoding Strategies In Morphology. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, & T. Verhoef (Eds.), The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11).

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

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Roberts, S.G., Killin, A., Deb, A., Sheard, C., Greenhill, S.J., Sinnemäki, K., Segovia-Martín, J., et al., 2020. CHIELD: the causal hypotheses in evolutionary linguistics database. *Journal of Language Evolution*, *5*(2), pp.101-120.