Causal inference and language evolution: A case study of Subject and Object cues

Natalia Levshina

Max Planck Institute for Psycholinguistics, Nijmegen

Understanding language evolution requires investigating how different linguistic and extralinguistic variables shape language structure. In order to test causal hypotheses of this kind, we need data about the past states of languages, which is very often missing. A popular alternative is artificial miniature language learning and communication (e.g., Kirby et al 2008; Culbertson et al. 2020; Raviv et al. 2019), which aims to reproduce language evolution in the lab. This approach has its advantages and limitations. This talk focuses on an alternative method, namely, causal inference based on synchronic data, whose potential for linguistics has been demonstrated convincingly (Blasi & Roberts 2017; Roberts et al. 2020). The recent advances in multilingual corpus creation and annotation, such as the Universal Dependencies project (Zeman et al. 2020), and typological database creation (e.g., https://glottobank.org/) make this approach increasingly attractive.

This paper will demonstrate how causal inference can be performed to model the relationships between different linguistic cues that help to identify the grammatical roles of Subject and Object, which convey "who did what to whom". These cues are case marking, rigid word order of Subject and Object, verb-medial order, and 'tight' semantics of the arguments, which is measured as association between the roles and lexemes (cf. Sapir 1921; Hawkins 1986; Sinnemäki 2010, 2014; Gibson et al. 2013). Using large web-based corpora from more than 30 languages annotated with the Universal Dependencies, and smaller Universal Dependencies corpora of over 100 languages, as well as typological data from reference grammars and the World Atlas of Language Structures (Dryer and Haspelmath 2013), I will create causal networks with the help of the Fast Causal Inference algorithm (Kalish et al. 2012; Dellert 2019). The preliminary results of the causal analyses corroborate previous findings from historical linguistics and artificial language learning (e.g., Bauer 2009; Fedzechkina et al. 2016), which means that this method can be added to the evolutionary linguist's toolkit. Most likely, the distribution of the linguistic cues is influenced by sociolinguistic factors, such as the proportion of L2 speakers and population size (McWhorter 2011; Lupyan and Dale 2010; Bentz and Winter 2013; Fenk-Oczlon and Pilz 2021).

Keywords: causal analysis, core arguments, case marking, word order, semantic tightness

References

Bauer, B.M. (2009). "Word order" in *New Perspectives on Historical Latin Syntax: Vol 1: Syntax of the Sentence*, ed. P. Baldi & Pierluigi Cuzzolin (Berlin: Mouton de Gruyter), 241-316.

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

Blasi, D.E., & Roberts, S.G. (2017). "Beyond binary dependencies in language structure" in *Dependencies in Languag*e, ed. N.J. Enfield (Berlin: Language Science Press), 117–128.

Dellert, J. (2019). *Information-Theoretic Causal Inference of Lexical Flow*. Berlin: Language Science Press.

Dryer, M.S. & Haspelmath, M. (eds.) (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2021-12-11.)

Fedzechkina, M., Newport, E.L. & Jaeger, T.F. (2016). Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive Science* 41(2), 416-446.

Fenk-Oczlon, G. & Pilz, J. (2021). Linguistic Complexity: Relationships Between Phoneme Inventory Size, Syllable Complexity, Word and Clause Length, and Population Size. *Frontiers in Communication* 6, 626032.

Gibson, E., Piantadosi, S., Brink, K., Bergen, L., Lim, E., and Saxe, R. (2013). A noisy-channel account of crosslinguistic word order variation. *Psychological Science* 24(7), 1079–88.

Goldhahn, D., Eckart, Th. & Quasthoff, U. (2012). "Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages" in *Proceedings of the 8th International Conference on Language Resources and Evaluation*, ed. N. Calzolari, Kh. Choukri, Th. Declerck et al. (Istanbul: ELRA), 759-765.

Hawkins, J.A. (1986). *A Comparative Typology of English and German. Unifying the contrasts*. London: Croom Helm.

Kalish, M, Mächler, M., Colombo, D., Maathuis, M.H. & Bühlmann, P. (2012). Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software* 47(11), 1-26.

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

McWhorter, J. (2011). *Linguistic simplicity and complexity: Why do languages undress?* Berlin: de Gruyter Mouton.

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

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt.

Sinnemäki, K. (2010). Word order in zero-marking languages. *Studies in Language* 34(4), 869-912.

Sinnemäki, K. (2014). "Complexity trade-offs: A case study" in *Measuring Grammatical Complexity*, ed. F.J. Newmeyer & L.B. Preston (Oxford: Oxford University Press), 179–201.

Zeman, D., Nivre, J., Abrams, M. et al. (2020). Universal Dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University. http://hdl.handle.net/11234/1-3226.