Testing communicative and learning biases in a causal model of language evolution: A study of cues to Subject and Object*

Natalia Levshina

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
natalia.levshina@mpi.nl

1 Introduction

If we want to understand language evolution, we should understand how different cognitive and communicative biases shape the structure of human languages. This paper investigates the relationships between different linguistic cues that help to identify the grammatical roles of Subject and Object, conveying "who did what to whom". These cues include 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 [11]. Previous research showed correlations between case marking and word order rigidity [14][21][22], between case marking and semantic tightness [11][14], and between case marking and the position of the verb in a simple clause [8][10][11][14][22]. These correlations have been explained by language users' bias towards efficient behaviour – more exactly, saving articulation costs and avoiding additional processing costs due to reanalysis.

At the same time, there are a number of causal hypotheses about the role of sociolinguistic variables, such as population size and proportion of L2 speakers, in language evolution. One of them says that a high proportion of L2 users in a community leads to morphological simplification, such as loss of case [3][15][17]. This finding has been explained by the fact that adults are poor grammar learners [24]. In contrast, a large-scale study in [13] finds no evidence of correlations between linguistic variables and the proportion of L2 users or vehicularity as an indicator for whether a language is used by L2 speakers; instead, the study reveals a relationship between the total number of speakers and information-theoretic complexity, which represents linguistic reduncancy. The question of this paper is, how all these variables interact in a causal network and whether we find evidence of the communicative and learning biases proposed in the literature.

2 Data

In order to test these hypotheses, I used diverse sources of data:

- large web-based corpora of online news [9] annotated with Universal Dependencies;
- smaller Universal Dependencies corpora of over 100 languages [25];
- the parallel corpus of Bible translations [16] and word order data inferred from this corpus [26];

*I thank Gerhard Jäger, Matías Guzmán Naranjo and Adèle H. Ribeiro for enlightening discussions of causal models.

- the World Atlas of Language Structures (WALS) [1].

From these sources I obtained the information about three variables that help to understand "who did what to whom", according to the literature:

- entropy of Subject and Object order based on the probabilities of SO and OS orders in the corpora. The lower this measure, the more a language user can rely on word order in order to identify Subject and Object;
- whether the forms of Subject and Object are the same or distinct thanks to morphological or adpositional case marking;
- the position of the lexical verb in a transitive clause. It has been claimed that the verb-middle order helps the addressee to identify Subject and Object [8].

As for the sociolinguistic variables, I took information about the population size and L2 speaker proportions from the Ethnologue database, as well as the datasets used in [23] and [3]. The genealogical and geographic information (linguistic genus and macroarea) was also included, in order to control for the genealogical and areal dependencies between the languages. I was able to obtain the linguistic and sociolinguistic data for 112 languages from 45 genera.

3 Method

In order to infer and test causal relationships, we would need information about the past states of languages. Unfortunately, it is very often missing (but see [2] on evolution of word order in Romance languages). A popular alternative is artificial miniature language learning and communication, which has shown correlations between some of the variables of interest [6][19].

This paper develops an alternative method, namely, causal inference based on synchronic data, whose potential for linguistics has been demonstrated in [4],[20] and other works. The causal graphs were created with the help of the Fast Causal Inference (FCI) algorithm implemented in the `pcalg` R package [12] (see a linguist-friendly introduction in [5]). Since the data were mixed (numeric and categorical), I wrote additional R scipts to test conditional independence between the variables. The scripts performed likelihood ratio tests on mixed-effects regression models (more exactly, Gaussian, logistic and beta regression) to determine whether or not the variables are conditionally independent. The genealogical dependencies between the languages were controlled by treating the genera as random intercepts. The geographic macroareas were treated as fixed effects, due to the fact that there were only four areas.

4 Results

The resulting causal network is displayed in Figure 1. Causal relationships were identified only between the three of the five variables: the number of users (`Total_Users`), entropy of Subject and Object order (`SO_Entropy`) and the position of the verb (`Verb`). The cirles at the end of the arrows represent uncertainty. For example, there is a possible causal effect of `Total_Users` on `SO_Entropy`, but it can also be a bidirectional relationship (which is not likely in our case). The same holds for `Verb` and `SO_Entropy`.

Contrary to expectations, there is no support for the causal links between between the proportion of L2 users (`L2Prop`) and the linguistic variables. Surprisingly, we also find no links between the same or different forms of Subject and Object (`SO_Form`), which was strong in previous studies [14]. The Area does not play a role, either.

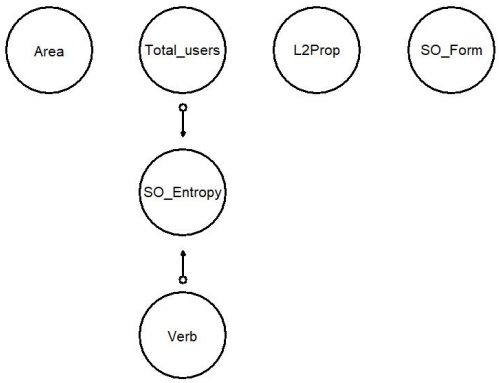

Figure 1: Causal Network

5 Discussion

The causal analysis does not support the hypothesis about the L2-related learning biases. There is no effect of proportion of L2 speakers on the main cues that help language users to distinguish between Subject and Object. Instead, the model comes to a conclusion similar to the one in [13], namely, that the total number of speakers is more likely to be relevant for language structure. In the present study, the number of speakers has an effect on how variable or fixed the order of Subject and Object is. This conclusion is intuitive. Two dominant languages, English and Chinese, have both a huge number of speakers, and very little variability in the order of Subject and Object.

How can one explain these findings? It is possible that a large number of language users increases variability of their language and cultural background and therefore increases noise in the communication channel, making the transmission of information less reliable. In this case, fixation of word order can be a response to the noise (cf. [8]). Rigid word order may be the most obvious strategy for expressing "who did what to whom" unambiguously. As for the link between the verb position and Subject - Object entropy, it is possible that rigidification helps best when the verb is in the middle.

The effect of the number of speakers on language structure has been discussed in [18] and [19]. According to the latter, larger groups of interacting participants, who learned an artificial language, developed more systematic languages over time. The increase in structure can be interpreted as a compensation for the greater linguistic variability in the larger groups. However, it is important to remember that the degree of systematicity, or language structure, was measured in that study as the correlations between string distances in the participants' linguistic output and semantic distances between the stimuli, which differed along some physical dimensions. The measure of language structure reflected morphological isomorphism and compositionality. The present study finds a causal relationship between the number of speakers and word order variability, which is a different aspect of systematicity. This is a new finding.

There are several caveats that need to be mentioned. The lack of effect of the number of speakers on case marking as one of the most salient aspects of morphological complexity may seem unexpected. But since the relevant variable in this study only describes whether the forms of Subject and Object are identical or not (including adpositional markers), perhaps this may not be the best measure for assessing the effect of L2 speakers on the overall grammatical

complexity of a language. Another problem is that the sociolinguistic variables used in this study describe the current situation, and not the previous stages, during which the grammar was shaped up. Unfortunately, historical sociolinguistic data are very difficult to find. Yet another potential factor that may affect the results is phonological complexity, which has been shown to correlate both with morphosyntactic complexity and sociolinguistic variables [7]. Nethertheless, I hope that the causal inference method presented in this study will help us to understand better how diverse cognitive and communicative biases drive language evolution.

References

[1] The world atlas of language structures online. http://wals.info, 2013.

[2] Brigitte M. Bauer. Word order. In P. Baldi and Pierluigi Cuzzolin, editors, *New Perspectives on Historical Latin Syntax. Vol 1: Syntax of the Sentence*, pages 241–316. Mouton de Gruyter, Berlin, 2009.

[3] Christian Bentz and Bodo Winter. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, 3:1–27, 2013.

[4] Damian E. Blasi and Sean G. Roberts. Beyond binary dependencies in language structure. In N.J. Enfield, editor, *Dependencies in Language*, page 117–128. Language Science Press, Berlin, 2017.

[5] Johannes Dellert. *Information-Theoretic Causal Inference of Lexical Flow*. Language Science Press, Berlin, 2019.

[6] Maria Fedzechkina, Elissa L. Newport, and T. Florian Jaeger. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive Science*, 41:416–446, 2016.

[7] Gertraud Fenk-Oczlon and Jürgen Pilz. Linguistic complexity: Relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. *Frontiers in Communication*, 6:626032, 2021.

[8] Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. A noisy-channel account of crosslinguistic word order variation. *Psychological Science*, 24:1079–1088, 2013.

[9] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 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*, pages 759–765, Istanbul, 2012.

[10] Joseph Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg, editor, *Universals of Grammar*, pages 73–113. MIT Press, Cambridge, MA, 1966.

[11] John A. Hawkins. *A Comparative Typology of English and German. Unifying the contrasts*. Croom Helm, London, 1986.

[12] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47:1–26, 2012.

[13] Alexander Koplenig. Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *Royal Society Open Science*, 6:181274, 2019.

[14] Natalia Levshina. Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. *Frontiers in Psychology*, 12:648200, 2021.

[15] Gary Lupyan and Rick Dale. Language structure is partly determined by social structure. *PLoS One*, 5:e8559, 2010.

[16] Thomas Mayer and Michael Cysouw. Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, page

3158–3163, Reykjavik, 2014.

[17] John McWhorter. *Linguistic simplicity and complexity: Why do languages undress?* de Gruyter Mouton, Berlin, 2011.

[18] Daniel Nettle. Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B*, 367:1829–1836, 2012.

[19] Limor Raviv, Antje Meyer, and Shiri Lev-Ari. Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Science*, 286:20191262, 2019.

[20] Sean G. Roberts, Anton Killin, Angarika Deb, et al. Chield: the causal hypotheses in evolutionary linguistics database. *Journal of Language Evolution*, 5:101–120, 2020.

[21] Edward Sapir. *Language: An Introduction to the Study of Speech.* Harcourt, New York, 1921.

[22] Kaius Sinnemäki. Word order in zero-marking languages. *Studies in Language*, 34:869–912, 2010.

[23] Kaius Sinnemäki and Francesca Di Garbo. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology*, 9:1141, 2018.

[24] Peter Trudgill. *Sociolinguistic typology: Social determinants of linguistic complexity.* Oxford University Press, Oxford, 2011.

[25] Daniel Zeman et al. Universal dependencies 2.10. http://hdl.handle.net/11234/1-4758, 2022.

[26] Robert Östling. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, page 205–211, Beijing, 2015.