Bad data, better data, and where to find them

On using (parsed) (historical) corpora to study syntactic variation and change

A problem with studying the syntax of historical stages of languages, and its diachronic development, is the need to rely on corpora, i.e. written usage data. This is problematic in more than one way:

(i) The restriction to the written medium, which is usually quite distant from how people speak (Koch/Oesterreicher 1985), means that an important part of what makes a language is inaccessible through historical corpora.

(ii) In addition, the data we have are scant, the older the more so, and what is transmitted is often due to chance, and we have at best educated estimates of what was once there (Kestemont et al. 2022).

(iii) Worse, we know from present-day languages that there is pervasive variation, and that this is in fact part of core competence (Weinreich/Labov/Herzog 1968). However, historical corpora —again, the older the more— are often mainly or exclusively produced by a non-representative segment of the population (often upper-class men), therefore barring full access to the systematic variation that is so characteristic of living languages. This is unfortunate as we know that certain sociolinguistic groups play different roles in advancing change, particularly women (Labov 1990).

(iv) A further problem is the kinds of texts preserved, including some not very 'natural' genres like poetry or law books. Studies have shown that a skew in the kinds of genres represented in a historical corpus can significantly affect what we can know about the syntax of a given period, and the course of syntactic changes in diachrony (e.g. Breitbarth 2025).

All this raises questions about how characteristic texts in historical corpora are for the language use of their period. Therefore, Labov (1982: 20), famously stated that "Historical linguistics may be characterized as the art of making the best use of bad data". Even though we can now avail ourselves of a large and growing number of parsed historical corpora, which greatly facilitate research into diachronic syntactic developments, the bad data problem as reflected in (i)–(iv) persists. Depending on the amount of surviving material, it may be difficult to assess what constraints played a role in shaping the historical development, through what stages a transition developed, and what factors ultimately caused a change or the absence of an expectable change. On the other hand, one can assume that such factors must have been present to produce the observable data, and that they don't differ too much from factors shaping data from other languages, past and present. This lies at the heart of the Uniformitarian Principle (Labov 1994, Lass 1997, Walkden 2019), viz. the assumption that "knowledge of processes that operated in the past can be inferred by observing ongoing processes in the present" (Labov 1994:21).

There are at least two ways of dealing with the bad data problem in historical corpora, and fill the gaps. One is enlarging the amount of data by comparing comparable corpora of several languages. The availability of a slowly growing number of corpora parsed according to the same protocols (e.g. corpora parsed according to the constituency-based Penn scheme,[1] or the dependency-based Syntacticus-treebanks of older Indo-European languages[2]) allows cross-corpora search queries. The other way of getting better data, and a better understanding of

[1] Besides the Penn Parsed Corpora of Historical English also Penn parsed corpora of e.g. French, Portuguese, Icelandic, Old Saxon and Middle Low German, Early New High German, and Middle Dutch

[2] https://dev.syntacticus.org/

diachronic variation and change, is looking at present-day data. In the sentence immediately preceding the quote above, Labov (1980) wrote, "There is a natural alliance between dialect geographers, who study heterogeneity in space; sociolinguists, who study heterogeneity in society; and historical linguists, whose concern is heterogeneity in time." There is still a lot of methodological headway to be made: sociolinguists, while regularly working with (recorded) interviews, often only excerpt relevant variables from them. Rarely are entire interviews transcribed, let alone POS-tagged and parsed. In dialectology, spontaneous spoken data play a minor role, so far; directly or indirectly elicited data dominate the data collection. Besides the observer's paradox (even in careful elicitation, there are priming and accommodation effects, cf. by Van Craenenbroeck et al. 2019), selection bias is a problem here: elicitation only finds what was asked for, and may therefore under-report certain phenomena. Parsed corpora of spontaneous dialect speech could help overcome these problems, but corpus-based dialectology, particularly for the study of syntactic variation and change, is at best in its infancy. Luckily, things are now beginning to change: There are already two large dialect corpora parsed using the Penn scheme available, the Portuguese CORDIAL-SIN (Martins 2000–) and the Appalachian English AAPCAppE (Tortora et al. 2017). Additionally, the first stage of the Southern Dutch GCND (Breitbarth et al. 2024) was recently released, and a parsed subset of the Spanish COSER is currently under construction (Bonilla et al. 2022).

In my presentation, I will explore the ways in which better data can be obtained with parsed historical and dialect corpora, show in what ways they can actually yield superior results to elicitation, and argue that they can profitably employed to inform theoretical insights into syntactic variation and change.

References

Bonilla, J.E., M. Bouzouita, and R.L.S. Díaz (2022). La construcción del Corpus Oral y Sonoro del Español Rural – anotado y parseado: Avances en el etiquetado de las partes del discurso. *Revista Internacional de Lingüística Iberorrománica* 40, 77–96.

Breitbarth, A. (2025). Resumption and (non-)integration after left-peripheral central adverbial clauses in Middle Low German: the role of genre. To appear in P.Larrivée and F.Pinzin (Eds.), *Syntactic change through text-types*. Berlin: De Gruyter.

Breitbarth, A., Farasyn, M., Ghyselen, A., Hellebaut, L., Lamsens, F., Depuydt, K., Does, J. de, Niestadt, J., Mertens, K. (2024). Gesproken Corpus van de zuidelijk-Nederlandse Dialecten. 1st release October 2024. https://hdl.handle.net/10032/tm-a2-z8.

Kestemont, M., F. Karsdorp, E. de Bruijn, M. Driscoll, K. A. Kapitan, P. Macháin, D. Sawyer, R. Sleiderink, and A. Chao (2022). Forgotten books: The application of unseen species models to the survival of culture. *Science* 6582, 375, 765–769.

Koch, P. and W. Oesterreicher (1985). Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36, 15–43.

Labov, W. (1982). Building on empirical foundations. In W. P. Lehman and Y. Malkiel (Eds.), *Perspectives on Historical Linguistics*, pp. 17–72. Amsterdam/Philadelphia: Benjamins.

Labov W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2(2), 205–254.

Martins, A. M. (2000–). CORDIAL-SIN: Corpus Dialectal para o Estudo da Sintaxe / Syntax-oriented Corpus of Portuguese Dialects. Universidade de Lisboa, http://www.clul.ulisboa.pt/en/10-research/314-cordial-s.

Tortora, C., B. Santorini, F. Blanchette, and C. Diertani (2017). The Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE), version 0.1. www.aapcappe.org.

Van Craenenbroeck, J., M. van Koppen, and A. van den Bosch (2019). A quantitative-theoretical analysis of syntactic microvariation: Word order in Dutch verb clusters. *Language* 95, 333–370.

Weinreich, U., W. Labov, and M. I. Herzog (1968). Empirical Foundations for a Theory of Language Change. In W. P. Lehmann (Ed.), *Directions for Historical Linguistics*, pp. 95–195. Austin: University of Texas Press.