Hunting for structures in treebank forests

Considerations on the use of parsed corpora of spontaneous dialect speech

Anne Breitbarth (Ghent)

Dialectology has two main aims: (i) to catalogue the attested linguistic features per location and (ii) to find spatial patterns in the variation between these features. In pursuing the first goal, the data collection needs to strike a balance between naturalness and commensurability (cf. Nerbonne 2018: 234): while one wants to know how people speak in a normal conversation in a given place, the data from different places need to be comparable, and the effort of collecting them needs to be manageable. This has meant that until today, dialectology has mostly taken what could be called top-down approaches: elicited data dominate in data collection, whether they are gathered directly or indirectly, or corpora (as far as they exist for dialects) are searched for pre-selected features (e.g. Szmrecsanyi 2013). However, despite many advantages—comparability, replicability, and the possibility of collecting more data with fewer resources—, such top-down approaches face two problems. The more well-known one is Labov's "observer's paradox". It can be shown that even in careful elicitation, there are priming and accommodation effects (Van Craenenbroeck/Van Koppen/Van den Bosch 2019). The second problem, which has received much less attention, is that elicitation only finds what was looked for, and may therefore underreport phenomena that are in fact very characteristic of a dialect, perhaps at low frequencies, or restricted to specific discourse contexts, simply because they do not arise in translation of written questionnaires or in translation / choice tasks. For instance, the field notes of the Syntactic Atlas of the Dutch Dialects (SAND, Barbiers et al. 2006) mention for some locations that the informants rejected sentence 359, with lack of inversion of subject and finite verb after an adverbial normally requiring inversion (1), but that the field worker nevertheless reports hearing the rejected word order frequently in spontaneous speech (Haegeman/Greco 2018, Breitbarth et al. 2021).

(1) *Met zulk weer, je kunt niet veel doen*
with such weather you can not much do

'With such weather, you cannot do much' (SAND sentence 359)

A bottom-up approach, letting the relevant linguistic properties emerge from a naturalistic corpus based on recorded speech would circumvent these two problems (e.g. Wolk/Szmrecsanyi 2016). However, while using speech corpora (based on interviews) is the main source of data in variationist sociolinguistics, their use in the study of the syntax of traditional/historical dialects is only emergent, and suitable corpora are extremely rare. A principal argument against using speech corpora is that relevant syntactic features may be accidentally unattested in a particular recording, particularly if they are rare, and that therefore extremely large amounts of speech data —ideally be parts-of-speech tagged and parsed, to make it possible to compare more abstract structures between places— would be required, making such an undertaking costly and unfeasible. For this reason, such corpora are not yet available for most languages, the only currently available ones are CORDIAL-SIN (Martins 2000-; Magro 2010) and AAPCAppE (Tortora et al. 2017). Due to the enormous effort of creating such resources, they are furthermore somewhat restricted in size —CORDIAL-SIN covers 42 places in Portugal (incl. Madeira and Açores), spanning ca. 600,000 tokens, AAPCAppE has ca. 1 million tokens in total. As earlier explorative studies have suggested (Sanders 2010, Wolk 2014), this may be at the lower

end of what is necessary to extract meaningful geospatial patterns. A parsed corpus of European Spanish (COSER-UD; Bonilla et al. 2022, ca. 1.6 million tokens), and a parsed spoken corpus of Southern Dutch dialects (GCND; Breitbarth et al. 2020; Ghyselen et al. 2020; Farasyn et al. 2022, ca. 5 million tokens) are currently in preparation.

In my presentation, I will present an experiment using the Portuguese CORDIAL-SIN, sketching a programme for exploiting parsed corpora of spontaneous dialect speech for finding geospatial patterns of syntactic variation, and identifying features that may have so far escaped attention in top-down approaches, and which may help fine-tuning our knowledge of dialect areas, as syntactic isoglosses might not necessarily align with the phonological and morphological ones known from traditional dialectology. In passing, I will raise the question in how far such syntactic features can be used to formulate a "syntactic fingerprint" of a given variety thus identified and how this can be integrated into a theory of parametric variation.

References

Barbiers, S. et al. 2006. Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND). Amsterdam: Meertens Institute. http://www.meertens.knaw.nl/sand/.

Bonilla, J.E./M. Bouzouita/R.L. Segundo 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., M.Farasyn, A.-S.Ghyselen & J.Van Keymeulen. 2020. Het Gesproken Corpus van de zuidelijk-Nederlandse Dialecten. *Handelingen van de Konkinklijke Zuid-Nederlandse Maatschappij voor Taal- en Letterkunde en Geschiedenis (KZM)* LXXII: 23–38.

Farasyn, M., A.-S.Ghyselen, J.Van Keymeulen & A.Breitbarth. 2022. Challenges in tagging and parsing spoken dialects of Dutch. *Journal of Historical Syntax* 6 (4–11). https://doi.org/10.18148/hs/2022.v6i4-11.92. [Special issue 'Annotating Historical Corpora'].

Ghyselen, A.-S., A.Breitbarth, M.Farasyn, J.Van Keymeulen & A.van Hessen. 2020. Clearing the Transcription Hurdle in Dialect Corpus Building: The Corpus of Southern Dutch Dialects as Case Study. *Frontiers of Artificial Intelligence* 3: 1–17.

Haegeman, L. & C.Greco. 2018. West Flemish V3 and the interaction of syntax and discourse. *JCGL* 21: 1–56.

Martins, A. M. (Coord.) [1999-2022]. CORDIAL-SIN: Corpus Dialetal para o Estudo da Sintaxe / Syntax-oriented Corpus of Portuguese Dialects. Lisboa: Centro de Linguística da Universidade de Lisboa. https://www.clul.ulisboa.pt/projeto/cordial-sin-corpus-dialectal-para-o-estudo-da-sintaxe

Magro, C. 2010. When CORDIAL Becomes Friendly: Endowing the CORDIAL Corpus with a Syntactic Annotation Layer. LREC 2010: 3705–3711.

Nerbonne, J. & W.Wiersma. 2006. A Measure of Aggregate Syntactic Distance. In J.Nerbonne/E.Hinrichs (eds.), *Linguistic distances workshop at the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, 82–90.

Nerbonne, J. 2018. Section 2—Methods. Introduction. In C.Boberg/J.Nerbonne/D.Watt (eds.), *The Handbook of Dialectology*, 233–239. Oxford: Blackwell.

Sanders, N.C. 2010. A statistical method for syntactic dialectometry. Ph.D. diss., Indiana University.

Szmrecsanyi, B. 2013. *Grammatical Variation in British English Dialects. A Study in Corpus-Based Dialectology*. Cambridge: CUP.

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

Van Craenenbroeck, J., M.van Koppen & A.van den Bosch. 2019. A quantitative-theoretical analysis of syntactic microvariation: Word order in Dutch verb clusters. *Language* 95(2): 333–370.

Wolk, C. 2014. Integrating Aggregational and Probabilistic Approaches to Dialectology and Language Variation. Ph.D. diss., Freiburg i.Br.

Wolk, Christoph & Benedikt Szmrecsanyi. 2016. Top-down and bottom-up advances in corpus-based dialectometry. In M.-H.Côté/R.Knooihuizen/J.Nerbonne (eds.), *The future of dialects*, 225–244. Berlin: Language Science Press.