Combining corpus research with agent-based modelling in the study of morphological and syntactic variation

Agent-based modelling has been used extensively in evolutionary linguistics (see references in Steels 2011) and is beginning to find its way to historical linguistics (see references in Pijpops, Beuls and Van de Velde 2015). Meanwhile, its use in variational linguistics has been more limited (for exceptions, see Fagyal et al., 2010; Karjus & Ehala, 2018). Moreover, the few available variational studies that employ agent-based modelling do not combine the technique with corpus research. This is regrettable, as both methodologies are highly complementary, as I aim to show in this talk.

To do so, agent-based modelling and corpus research will be combined in the study of lectal contamination. This is an effect whereby the lexical items that are more often used in one lect, e.g. a socio-, regio-, ethno- or dialect, exhibit a preference for a morphosyntactic variant typical of that same lect, even among language users of a different lect. This effect is theorized to emerge under four preconditions: (i) a different distribution of the morphosyntactic variants in both lects, (ii) probabilistic differences in lexical usage between both lects, (iii) language contact between both lects, and (iv) the cognitive storage of ready-made forms by language users. These four preconditions are implemented in an agent-based model, and the effect is shown to emerge consistently.

Next, lectal contamination is tested using observational data extracted from the Corpus of Spoken Dutch and the Sonar corpus of written Dutch (Oostdijk et al. 2002, 2013). Two case studies are put under scrutiny: the morphological alternation between the variant with versus without *-s* ending of the partitive genitive construction (Pijpops and Van de Velde 2018), and the syntactic choice between the determiners *zulk* 'such' versus *zo'n* 'so a' in front of plural nouns and singular mass nouns (Ghesquière and Van de Velde 2011; Van Olmen 2019). Both cases concern variation from the nominal domain, with the respective first variant being more widespread in the Netherlandic regiolect of Dutch, while the respective second variant is more popular in the Belgian regiolect. It is then predicted that lexical items that are more often used in the Netherlandic regiolect will more often exhibit the 'Netherlandic' variant, and vice versa, both in the language use of Belgians and in the language use of people from the Netherlands. To test this prediction, mixed model regression analysis is employed.

In sum, agent-based modelling can show us which theoretical preconditions are required for an effect to emerge, and how an emergent effect is predicted to behave in various ecological settings. Meanwhile, corpus research can provide an empirical confirmation of these models. Combined, these techniques can feed into one another, forming a continued circle of refinement of both theory and analysis.

References

Fagyal, Zsuzsanna, Samarth Swarup, Anna María Escobar, Les Gasser and Kiran Lakkaraju. 2010. Centers and peripheries: Network roles in language change. *Lingua* 120(8). 2061–2079.

Ghesquière, Lobke and Freek Van de Velde. 2011. A corpus-based account of the development of English such and Dutch zulk: Identification, intensification and (inter)subjectification. *Cognitive Linguistics* 22(4). 765–797.

Karjus, Andres and Martin Ehala. 2018. Testing an agent-based model of language choice on sociolinguistic survey data. *Language Dynamics and Change* 8(2). Leiden | Boston: Brill. 219–252.

Oostdijk, Nelleke, Wim Goedertier, Frank Van Eynde, Louis Boves, Jean-Pierre Martens, Michael Moortgat and Harald Baayen. 2002. Experiences from the Spoken Dutch corpus project. *Proceedings of the third international conference on language resources and evaluation (LREC)*, 340–347.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste and Ineke Schuurman. 2013. The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In Peter Spyns & Jan Odijk (eds.), *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, 219–247. Heidelberg: Springer.

Pijpops, Dirk, Katrien Beuls and Freek Van de Velde. 2015. The rise of the verbal weak inflection in Germanic. An agent-based model. *Computational linguistics in the Netherlands Journal* 5. 81–102.

Pijpops, Dirk and Freek Van de Velde. 2018. A multivariate analysis of the partitive genitive in Dutch. Bringing quantitative data into a theoretical discussion. *Corpus Linguistics and Linguistic Theory* 14(1). 99–131.

Steels, Luc. 2011. Modeling the cultural evolution of language. *Physics of Life Reviews* 8(4). 339–356.

Van Olmen, Daniël. 2019. A diachronic corpus study of prenominal zo'n "so a" in Dutch: Pathways and (inter)subjectification. *Functions of Language* 26(2). 216–247.