Extracting construction networks from Cantonese speech corpora using clustering algorithms

Andreas Liesenfeld (Nanyang Technological University, Singapore)

Language is constructions “all the way down”, concludes Adele Goldberg (2006:18). Starting from this key insight associated with construction grammar (CxG), this study addresses an issue that is a result of CxG’s assumption that “the network of constructions captures our grammatical knowledge in toto” (ibid): the nature of the network organization of constructions.

Adopting a usage-based constructionist approach, this ongoing PhD-level study aims to model structural properties of construction networks through clustering properties of in-situ speech sequences. This data-driven grammar induction draws on recent developments in CxG that conceptualize language as complex adaptive systems (CAS) (Beckner et al. 2009). The Language-as-CAS approach holds that construction networks emerge from interrelated patterns of social interaction, experience and cognitive processes. Grounded in natural speech data, the study aims to provide empirical evidence for this emergence process, exploring ways of how construction networks can be identified, extracted and presented as complex adaptive system networks.

The goal of the project is to explore ways of how typologies of constructions can be extracted from Cantonese speech data organized in the CHAT format (minCHAT) (MacWhinney 2000). Utilizing various clustering algorithms, the subsequent network extraction is based on three key mechanisms that shape construction networks identified by Ellis (2012): frequency, recency and context. Preliminary results of this data analysis show how structural properties of speech corpus data can be extracted by applying clustering algorithms that simulate domain-general cognitive constraints in humans. The extracted construction networks can be used to complement or replace existing Cantonese grammar formalisms for various NLP tasks, such as natural language understanding, and provide a new model for data-driven grammar induction.

References

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, Jinyun, Larsen-Freeman, D., Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning, 59*(s1), 1--26.

Ellis, N. C. (2012). What can we count in language, and what counts in language acquisition, cognition, and use. *Frequency effects in language learning and processing, 1*, 7--34.

Goldberg, A. E. (2006). *Constructions at Work: the nature of generalization in language*. Oxford: Oxford University Press.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database. *Computational Linguistics, 26*(4), 657--657.