Operationalising Usage-Based Construction Grammar on a Large Scale

Paul Van Eecke[1,2,3], Lara Verheyen[1] & Katrien Beuls[4]
[1]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels
[2]Itec, imec research group at KU Leuven, E. Sabbelaan 51, B-8500 Kortrijk
[3]KU Leuven, Faculty of Arts, Blijde Inkomststraat 21, B-3000 Leuven
[4]Faculté d'informatique, Université de Namur, rue Grandgagnage 21, B-5000 Namur
paul@ai.vub.ac.be
lara.verheyen@ai.vub.ac.be
katrien.beuls@unamur.be

Keywords: Fluid Construction Grammar; large-scale; usage-based

In this talk, we will (i) discuss the large-scale operationalisation of usage-based construction grammar, (ii) present a broad-coverage usage-based construction grammar operationalised in the Fluid Construction Grammar framework (Steels, 2011; van Trijp et al., 2022), and (iii) introduce a tool that makes use of this grammar to support searching through corpora from a semantic perspective.

The grammar was constructed algorithmically based on the PropBank-annotated OntoNotes (Weischedel et al., 2013) and EWT (Bies et al., 2012) corpora. In essence, it contains over a 100,000 constructions, interlinked by a categorial network consisting of over 85,000 nodes representing grammatical categories and 750,000 edges representing categorial links. The grammar can efficiently and effectively be used to map from English utterances (form) to frame-semantic-inspired representations formalised using the PropBank convention (meaning) (Palmer et al., 2005). It thereby mainly focusses on events and their participants on the semantic side and argument structure on the form side.

Apart from a general discussion about why scaling (usage-based) construction grammar is of crucial importance for the future of the field and presenting a novel methodology to automatically build large-scale usage-based construction grammars, we will also introduce a first example of a methodological tool that makes use of this methodology to support usage-based linguists in their empirical research. More specifically, the tool enables users to search for corpus observations which instantiate a particular semantic structure. For example, a user might want to search for corpus observations in which the transfer sense of the verb 'give' is expressed in combination with an agent (i.e. a giver), a theme (i.e. a thing being given) and a beneficiary (i.e. a receiver). The tool then retrieves corpus examples that instantiate this semantic structure using any morpho-syntactic realisation, e.g. 'scholars will give you a detailed analysis', 'he gives priority to diplomacy or internal affairs' or 'the Spirit gives to one person the power to do miracles'. The tool provides a user-friendly interface for defining semantic structures of interest using the PropBank rolesets. Optionally, form-related constraints can be included, in particular constraints on the order in which the semantic roles are realised, the morpho-syntactic means through which one or more of the semantic roles are expressed, or the exact strings that appear in the instantiations of the semantic roles. A beta version of the tool is available at https://ehai.ai.vub.ac.be/ccxg-explorer/.

In sum, this talk will make a case for scaling up (usage-based) construction grammar, present a novel methodology for algorithmically building large-scale usage-based construction grammar, and introduce a methodological tool that uses this methodology to support empirical linguistic research.

References

Bies, Ann, Justin Mott, Colin Warner & Seth Kulick. 2012. English web treebank ldc2012t13. Philadelphia, Linguistic Data Consortium.

Palmer, Martha, Daniel Gildea & Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics* 31(1). 71–106.

Steels, Luc (ed.). 2011. *Design patterns in Fluid Construction Grammar*. Amsterdam: John Benjamins.

van Trijp, Remi, Katrien Beuls & Paul Van Eecke. 2022. The FCG editor: An innovative environment for engineering computational construction grammars. PLOS ONE 17(6). doi:10.1371/journal.pone.0269708.

Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin & Ann Houston. 2013. Ontonotes release 5.0 ldc2013t19. Philadelphia, Linguistic Data Consortium.