Agreement groups processing of context-free utterances: coverage, structural precision, and category information

László Drienkó

The present work tests the coverage potential of the “agreement groups” (AG) approach, a language processing model based on forming groups of similar utterances, on corpora generated by a simple context-free grammar. The idea of ‘agreement groups’ and ‘agreement groups coverage’ was presented in a series of works as a usage-based, distributional approach to modelling linguistic processing. Drienkó (2014) showed that agreement groups, i.e. groups of utterances differing from a base utterance in only one word, can account for a certain percent of novel utterances of mother-child speech, may facilitate categorisation (lexical/syntactic, semantic), and might serve as a basis for ‘real’ agreement relations. For the processing of longer utterances the idea of ‘coverage’ was introduced in Drienkó (2015, 2016). Although coverage yields a quantitative characterisation of how *much* an utterance can be covered by smaller utterance fragments (corresponding to agreement groups), it does not allow conclusions about the correctness of the coverage procedure.

To tell how *well* an utterance is covered we would need an appropriate comparison procedure. Unfortunately, such a procedure is not available in general since there can be several ‘correct’ ways to cover an utterance. By ‘correct’ we do not mean ‘grammatical in a formal sense’. For instance, the utterance *Where shall we go next* could be covered with fragments (inter alia) *where go, where next, we go, shall we, shall we go, go next, where shall we go.* Sentences of formal languages, on the other hand, have the advantage of being directly testable with respect to their constituent structure imposed on them by the rules of the language that generates them.

To capitalise on this property of formal languages, in our investigation we consider small fragments of English that can be generated by context-free (CF) grammar rules. Both the ‘training corpora’ and the test sets are generated by CF rules. As in previous work, we extract agreement groups from the training set and check to what extent the utterances of the test set can be covered by them. In addition, here we use a precision metric to characterise the correspondence between coverage structure and constituent structure, i.e. the correspondence between the covering fragments and the right-hand sides of the rules of the generating grammars.

In our experiments we set different values for various parameters: number of training corpus utterances, number of test corpus utterances, category information, and depth of rule application, i.e. how many rules generate the utterances. Category information means that in the ‘informed’ case the syntactic category of words is taken into account by the group formation algorithm, whereas in the ‘uninformed’ case no such information is available.

We find that i) the AG framework may be useful in modelling general developmental processes: larger training corpora effect higher coverage and precision, comprehension precedes production; ii) information on syntactic categories hinders processing, i.e. “less is more” (Newport, 1990); iii) agreement groups can code structural information; iv) there is a notable parallelism between coverage and structural precision for CF utterances.

References

Drienkó, L. (2014). Agreement groups analysis of mother-child discourse. In Rundblad, G., Tytus, A., Knapton, O., and Tang, C. (eds.) *Selected Papers from the 4th UK Cognitive Linguistics Conference.* London:UK Cognitive Linguistics Association. pp. 52-67.

Drienkó, L. (2015). Discontinuous coverage of English mother-child speech. Talk presented at the *Budapest Linguistics Conference,* 18-20 June 2015, Budapest, Hungary.

Drienkó, L. (2016). Agreement groups coverage of English mother-child utterances for modelling linguistic generalisations. *Journal of Child Language Acquisition and Development– JCLAD*. Vol. 4, Issue 3, pp. 113-158.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science,* 14, 11-28.