Finding Long-distance Dependencies in Treebanks

Gosse Bouma (University of Groningen)

A central topic in theoretical syntax is the proper analysis of non-local dependencies of the kind found in *wh*-questions and relative clauses. Rather different solutions have been proposed in various theoretical frameworks such as Transformational Grammar (Chomsky, 1977), Categorial Grammar (Steedman, 2000), GPSG (Gazdar et al., 1985), HPSG (Bouma et al., 2001), and LFG (Kaplan and Zaenen, 1989). One of the surprising facts is that there is still considerable disagreement about what the relevant data are and whether these are to be accounted for in syntax or by an appeal to general cognitive constraints (Hofmeister and Sag, 2010). Another observation that is somewhat at odds with the claims of most studies in theoretical syntax is that in actual usage, sentences involving a true long-distance dependency (LDD) are rare, and often involve the same matrix verb and subject, suggesting that these are all variants of a small set of constructions (Verhagen, 2006). Corpus-based research on these issues has been hindered by the fact that LDDs are difficult to find using search patterns consisting of combinations of lexical items and part-of-speech tags only. Manually annotated syntactic treebanks offer the right kind of annotation, but are too small to be of use in this respect. Accurate computational grammars allow large corpora to be annotated with syntactic relations automatically, and can produce large corpora with the right level of annotation.

In this paper, we present the results of searching for four kinds of LDDs in an automatically annotated treebank for Dutch. We concentrate on phenomena that have recently been subject to debate, and where conflicting claims have been made regarding the question whether these constructions actually occur with some frequency in corpora, i.e.:

1. To what extent do we find collocational effects in LDDs (cf. Verhagen (2006))?
2. Do we find LDDs involving infinitival clauses introduced by the optional complementizer *om*?
3. What is the relationship between resumptive prolepsis (Hoeksema and Schippers, 2012) and the (absence) of LDDs? and
4. Do parasitic gap constructions involving R-pronouns (Everaert et al., 2015) occur in actual usage?

It has been observed that even the best statistical parsers are not very good at handling non-local dependencies (Rimell et al., 2009). As we are using a corpus that was automatically annotated using the Alpino parser (van Noord, 2006), this study also gives some insights in the accuracy of Alpino.

LDDs involving a 'gap' in a (tensed or infinitival) subordinate clause are all covered by the Alpino parser, and thus can be searched for directly. The results show that dependencies of this type are quite infrequent in the corpus and do provide support for claims that there are collocational effects. Manual inspection of the results showed that the precision of the parser on these constructions is not very high (0.27-0.35, but comparable with state-of-the-art). Manual inspection of sentences containing a relevant matrix verb and subordinate clause suggests that recall is acceptable. LDDs involving a *to*-infinitive are somewhat more frequent and frequently occur with the matrix verb *achten* (*consider*). LDDs involving an *om te*-infinitive are considerably less frequent, but do occur. Resumptive prolepsis and R-pronominal parasitic gaps are outside the scope of the grammar. For the resumptive pronoun construction, an approximate query turned out to be quite accurate, and gave rise to a high number of results. The distribution of matrix verbs in this construction supports the findings of Hoeksema and Schippers (2012). For R-pronominal parasitic gaps, it is much harder to come up with a good approximate query. However, after manual filtering we did nd a number of positive examples. In this case, the main advantage of using a syntactically annotated corpus is that it avoids the inherently limited recall of search patterns based on strings only.

References

Gosse Bouma, Rob Malouf, and Ivan Sag. Satisfying constraints on adjunction and extraction. *Natural Language and Linguistic Theory,* 19:1-65, 2001.

Noam Chomsky. On wh-movement. In Akmajian Adrian Culicover Peter, Wasow Thomas, editor, *Formal Syntax*. Academic Press, New York, 1977.

Martin Everaert, Riny Huybrechts, Noam Chomsky, Robert Berwick, and Johan Bolhuis. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences,* 19:729-743, 2015.

Gerald Gazdar, Ewan Klein, Geoffrey Pullum, and Ivan Sag. *Generalized Phrase Structure Grammar*. Blackwell, 1985.

Jack Hoeksema and Ankelien Schippers. Diachronic changes in long-distance dependencies. *Historical Linguistics 2009: Selected Papers from the 19th International Conference on Historical Linguistics*, Nijmegen, 10-14 August 2009, 320:155, 2012.

Philip Hofmeister and Ivan A Sag. Cognitive constraints and island effects. *Language*, 86(2): 366-415, 2010.

Ronald M. Kaplan and Annie Zaenen. Long-distance dependencies, constituent structure and functional uncertainty. In Mark R. Baltin and Anthony S. Kroch, editors, *Alternative Conceptions of Phrase Structure.* University of Chicago Press, 1989.

Laura Rimell, Stephen Clark, and Mark Steedman. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing:* Volume 2-Volume 2, pages 813-821. Association for Computational Linguistics, 2009.

Mark Steedman. Information structure and the syntax-phonology interface. *Linguistic inquiry,* 31(4):649-689, 2000.

Gertjan van Noord. At last parsing is now operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, TALN06. *Verbum Ex Machina*. *Actes de la 13e conference sur le traitement automatique des langues naturelles,* pages 20-42. 2006.

Arie Verhagen. On subjectivity and 'long distance Wh-movement'. In Angeliki Athanasiadou, Costas Canakis, and Bert Cornillie, editors, *Subjectification: Various Paths to Subjectivity,* pages 323-346. Mouton de Gruyter, Berlin, 2006.