Tackling data insufficiency:
Automatically extending a (richly annotated) data set

Daphne Theijssen, Nelleke Oostdijk, Hans van Halteren and Lou Boves
Radboud University, Nijmegen

Much effort has been – and continues to be – put into developing corpora to provide linguists with suitable data in sufficient quantities to perform their research. Still, for many types of research this remains an issue: even when numerous corpora are available, most of these are too small and/or have not been annotated with the required information. This paper addresses the problem of data insufficiency and presents an approach to automatically extend a data set.

In our project, we study situations where speakers can choose between several syntactic options that are equally grammatical, but where usually one variant is most acceptable given the context. The current focus is on the dative alternation in English, where speakers and writers can choose between a double object (*She handed the student the book*) or a prepositional dative construction (*She handed the book to the student*).

In the past, researchers have tried to explain the speakers' choices in different ways, e.g. by taking a syntactic (e.g. Quirk et al 1972), semantic (e.g. Gries and Stefanowitsch 2004) or discourse approach (e.g. Collins 1995). The aim of the current project is to explore the suitability of various statistical techniques to combine these approaches in a single model (e.g. Bresnan et al. 2007). For this, we need a richly annotated data set.

To be able to find the two constructions, and to incorporate the various approaches mentioned, we need a corpus with syntactic, semantic and discourse annotations. Since such a corpus does not exist, we selected a corpus meeting at least the first requirement: the one-million-word ICE-GB Corpus, containing written and spoken language in various genres. We automatically extracted and manually filtered instances of the dative alternation. The result was a set of 915 relevant instances, which we manually enriched with the information desired.

Initial experiments with the data have indicated that the set is still too small for our purposes. We therefore developed a (semi-)automatic approach to extend the our data set, using the 100-million-word British National Corpus (BNC). Since the BNC has no syntactic annotations, we made a list of ditransitive verbs and extracted all sentences containing them. These sentences were fed to the Connexor Machinese parser, yielding syntactic dependency trees. Sentences in which the parser identified a ditransitive construction were kept.

Next, we developed algorithms for automatically enriching our data with the information desired: the animacy, concreteness, definiteness, discourse givenness, pronominality, person and number of the objects (*the book* and *him* in the example), and the semantic class of the verb. The algorithms employed the part-of-speech tags available in the BNC, the dependency parses produced by the Connexor Machinese parser, and the noun classes and synonym sets found in WordNet.

We evaluated both steps of the process by applying it to the existing data set of 915 manually annotated instances. The details of the method and the results found are presented at the conference.

References

Bresnan, J., A. Cueni, T. Nikitina and R. H. Baayen. (2007). "Predicting the Dative Alternation". In G. Bouma, I. Kraemer and J. Zwarts (eds). *Cognitive Foundations of Interpretation:* 69-94. Amsterdam: Royal Netherlands Academy of Science.

Collins, P. (1995). The indirect object construction in English: an informational approach. *Linguistics* 33, 35-49.

Gries, S. Th. and A. Stefanowitsch. (2004). "Extending Collostructional Analysis: A Corpus-based Perspective on 'Alternations'". In *International Journal of Corpus Linguistics* 9, 97-129.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik. (1972). *A Grammar of Contemporary English*. London: Longman.