Corpus Evidence for Syntactic Structures and Requirements for Annotations of Tree Banks

Stefan Müller and Detmar Meurers
University of Bremen and The Ohio State University
Stefan.Mueller@cl.uni-bremen.de, dm@ling.osu.edu

In our talk we want to illustrate the usefulness of corpora to validate/falsify claims made in the linguistic literature. We will do so using three case studies, the first one dealing with extraposition and subjacency, the second one with the German clause structure and particle verbs, and the third with multiple frontings.

1 Case Study 1: Extraposition and Subjacency

Turning to the first case study, Chomsky (1986, p. 40; among others) argues that the trace *t* in (1) cannot be the source of the extraposition and explains this by the principle of subjacency, which says that only one Barrier may be crossed by such movement. See also Baltin 1981 on extraposition and subjacency.

(1) [$_{NP}$ Many books [$_{PP}$ with [stories t]] t'] were sold [that I wanted to read].

Grewendorf (1988, p. 281), Haider (1996, p. 261), and Rohrer (1996, p. 103) assume that subjacency also plays a role for extraposition in German. But if one substitutes the head noun in (1) in a way that reduces attachment ambiguities, one can obtain parallel German examples which are grammatical:

(2) weil viele Schallplatten mit Geschichten verkauft wurden, die ich noch
because many records with stories sold were that I yet
lesen wollte.
read wanted
'because many records with stories that I wanted to read were sold.' (The sentence describes a situation where the speaker goes to a record shop and for certain audio book records there he realizes he wants to read those stories.)

The example in (2) seems to falsify the subjacency claim frequently found in the literature—which raises the question whether one can find more examples to empirically explore this issue. Even with an unannotated corpus, examples with such extraposed complement clauses can be found by looking for sentences that contain a

complementizer and a noun that typically selects a clausal complement. The precision of such searches is quite low, though, since in many of the matches the complement clause is not extraposed.

Using a syntactically annotated corpus one can formulate a more precise query that includes the requirement that the complement clause be extraposed. We used the TIGER treebank (Brants et al., 2002), a syntactically annotated German newspaper corpus consisting of roughly 700,000 tokens (40,000 sentences), taken from the *Frankfurter Rundschau*, a national German newspaper. In our talk we will discuss a query which for this corpus returns sentences such as the following:

(3) [. . .] die Erfindung der Guillotine könnte [$_{NP}$ die Folge [$_{NP}$ eines
the invention of the Guillotine can the consequence of a
verzweifelten Versuches des gleichnamigen Doktors] gewesen sein, [seine
desperate attempt of the homonymous doctor been is his
Patienten ein für allemal von Kopfschmerzen infolge schlechter Kissen zu
patients once for all of headache due to bad pillow to
befreien].
free
'The invention of the Guillotine may have been the consequence of a desperate attempt of a doctor by the same name to, once and for all, free his patients of headaches caused by bad pillows.'

Based on corpus examples such as these, which we take to be ordinary, well-formed sentences of German, one can conclude that subjacency or related constraints such as the Complex NP Constraint of Ross (1967) do not universally hold for movement to the right.

2 Case Study 2: German Clause Structure and Particle Verbs

The second case study addresses the frequently made claim that particles of particle verbs cannot be fronted in German (cf. Müller, 2002, for an overview). The empirical issue has been used to define the class of particle verbs (Zifonun, 1999, p. 212), and it has played an important role in a number of syntactic arguments. For instance, Haider (1990) claimed that verb traces cannot be a part of the fronted projection, since if they were, one would expect sentence like (4) to be grammatical.

(4) * [Ein Buch auf t_i] schlug$_i$ Hans.
a book open (PARTICLE) beat Hans
'Hans opened a book.'

Turning to corpus searches intended to explore the empirical side of this issue, if one wants to use an unannotated corpus, one can try to look for fronted particles by searching for a particle that is separated by a space from its corresponding verb. According to orthographic conventions this would be the way to write particle and verb if the particle is fronted and the finite verb is in second position. But this requires spelling out all possible particle verbs and it clearly is questionable to rely on orthographic conventions for finding cases that supposedly do not exist at all.

Using a syntactically annotated corpus, it is easy to search for adjacent particles and finite verbs. For the TIGER corpus, one obtains sentences such as those shown in (5).

(5) Fest steht, daß dort 580 der insgesamt 4650 Arbeitsplätze wegfallen.
solid stands that there 580 of the in total 4650 jobs are cut
'It is certain, that 580 of the 4650 jobs are cut.'

Searching for fronted particles in a syntactically annotated corpus thus provides a range of examples showcasing this supposedly impossible pattern.

3 Case Study 3: Fronting as a Constituent Test

The third case study will lead us to the most complex query—and to the limits of what can be found in currently available corpora. German is a so-called verb-second language and a generally accepted empirical generalization is that only one constituent can appear in front of the finite verb in declarative main clauses. The strongest claim found in the literature is that the ability of material to appear in front of the finite verb is both sufficient and necessary for constituenthood.

However, as discussed in Müller (2003), there are well-formed example sentences such as those in (6), which according to other constituent tests include more than one constituent in front of the finite verb.

(6) a. [Gar nichts mehr] [mit dem Tabakkonzern] hat Jan Philipp Reemtsma
nothing.at.all more with the tobacco company has Jan Philipp Reemtsma
zu tun,
to do
'Jan Philipp Reemtsma has nothing to do with the tobacco combine.'

b. [Mit ihm] [auf der Anklagebank] sitzen zwei 18-Jährige,
with him on the dock sit two 18 year olds
'Two 18 year old people are in the dock with him . . . '

In order to collect more data, we again searched the TIGER treebank. The rather complex query needed for this search will be discussed in the talk. However, the

query did not return any matches for this corpus. The conclusion to be drawn from this is that, as a consequence of Zipf's law, many infrequent but theoretically relevant phenomena can only be found in exceedingly large corpora. In consequence, we tried to find the pattern in a larger corpus, the 200 million token "Tübingen Partially Parsed Corpus of Written German" (TPP-D/Z; F. H. Müller, 2004a; Ule, 2004). The corpus nicely supports syntactic queries thanks to an automatically obtained shallow syntactic annotation. For the phenomenon at issue here, however, the annotation was not rich enough to formulate queries that do not return overwhelmingly many false positives.

4 Conclusion

We used three case studies to showcase how corpora with syntactic annotation, the so-called treebanks, can be used to find linguistically interesting data. The case studies presented here thus complement those using more basic corpora with part-of-speech information discussed in Meurers (2005). However, treebanks result from a semi-automatic mark-up process, so that the size of treebanks is significantly smaller than that of part-of-speech annotated corpora. As illustrated by our third case study, currently existing treebanks are not large enough to include instances of many relevant linguistic patterns. In the future, a convergence of high-quality manual annotation efforts and automatically obtained shallow syntactic annotation will hopefully make larger corpora directly accessible for linguistic research.

References

Baltin, M. (1981). Strict bounding. In C. L. Baker and J. J. McCarthy, eds., *The Logical Problem of Language Acquisition*. The MIT Press, Cambridge, USA.

Brants, S., S. Dipper, S. Hansen, W. Lezius, and G. Smith (2002). The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria. `www.bultreebank.org/proceedings/paper03.pdf`.

Chomsky, N. (1986). *Barriers*. The MIT Press, Cambridge, USA; London, UK.

Grewendorf, G. (1988). *Aspekte der deutschen Syntax. Eine Rektions-Bindungs-Analyse*. Gunter Narr Verlag, Tübingen.

Haider, H. (1990). Topicalization and other puzzles of German syntax. In G. Grewendorf and W. Sternefeld, eds., *Scrambling and Barriers*, pp. 93–112. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Haider, H. (1996). Downright down to the right. In U. Lutz and J. Pafel, eds., *On Extraction and Extraposition in German*, pp. 245–271. Benjamins, Amsterdam.

Meurers, W. D. (2005). On the use of electronic corpora for theoretical linguistics. case studies from the syntax of German. *Lingua*, **115**(11):1619–1639. http://ling.osu.edu/~dm/papers/meurers-03.html.

Müller, F. H. (2004a). *Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universitt Tübingen. http://www.sfb441.uni-tuebingen.de/a1/Publikationen/stylebook-04.pdf.

Müller, S. (1999). *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche*. Max Niemeyer Verlag, Tübingen. http://www.cl.uni-bremen.de/~stefan/Pub/hpsg.html.

Müller, S. (2002). Syntax or morphology: German particle verbs revisited. In N. Dehé, R. S. Jackendoff, A. McIntyre, and S. Urban, eds., *Verb-Particle Explorations*, pp. 119–139. Mouton de Gruyter, Berlin, New York. http://www.cl.uni-bremen.de/~stefan/Pub/syn-morph-part.html.

Müller, S. (2003). Mehrfache Vorfeldbesetzung. *Deutsche Sprache*, **31**(1):29–62.

Müller, S. (2004b). Complex NPs, subjacency, and extraposition. *Snippets*, **8**:10–11. http://www.cl.uni-bremen.de/~stefan/Pub/subjazenz.html. 20.12.2005.

Müller, S. (2005). Zur Analyse der scheinbar mehrfachen Vorfeldbesetzung. *Linguistische Berichte*, **203**:297–330. http://www.cl.uni-bremen.de/~stefan/Pub/mehr-vf-lb.html. 20.12.2005.

Rohrer, C. (1996). Fakultativ kohrente Infinitkonstruktionen im Deutschen und deren Behandlung in der Lexikalisch Funktionalen Grammatik. In G. Harras and M. Bierwisch, eds., *Wenn die Semantik arbeitet. Klaus Baumg"artner zum 65. Geburtstag*, pp. 89–108. Max Niemeyer Verlag, Tübingen.

Ross, J. R. (1967). *Constraints on Variables in Syntax*. Ph.D. thesis, MIT, Cambridge, USA. Appeared as Ross (1986): *Infinite Syntax*. Norwood, USA: Ablex Publishing Corporation.

Ule, T. (2004). *Markup Manual for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universitt Tübingen. http://www.sfs.uni-tuebingen.de/tupp/dz/markupmanual.pdf.

Zifonun, G. (1999). Wenn *mit* alleine im Mittelfeld erscheint: Verbpartikeln und ihre Doppelgänger im Deutschen und Englischen. In H. Wegener, ed., *Deutsch kontrastiv. Typologisch-vergleichende Untersuchungen zur deutschen Grammatik*, pp. 211–234. Stauffenburg Verlag, Tübingen.