Corpora and Syntax

Article 44 in Lüdeling, A. and Kytö, M. *Corpus Linguistics*. Handbooks of Linguistics and Communication Science. Berlin: Mouton de Gruyter. 2007.

W. Detmar Meurers

The Ohio State University
Department of Linguistics
222 Oxley Hall, 1712 Neil Ave
Columbus, Ohio 43210
USA

Stefan Müller

Universität Bremen
Fachbereich 10
Postfach 33 04 40
D-28334 Bremen
Germany

Contents

1 Introduction

Syntactic analysis connects empirical observations about language with theoretical generalizations and explanations. Depending on the perspective of the framework or individual researcher, syntactic research has emphasized the empirical or the theoretical aspect of the enterprise; but independent of the philosophical dispute between empiricism and rationalism about the nature of the connection between data and knowledge (cf., e.g., Markie, 2004), it is clear that neither aspect exists entirely without the other: observation of data is shaped by prior experience and current research questions, and data is needed for establishing or falsifying a theory. Leaving the philosophical dispute aside, we can thus ask how one can obtain data that is relevant for a particular theoretical issue. We address this question in this article by discussing how electronic corpora can be used in support of the creation and falsification of syntactic theories.

1.1 On the use and limits of corpora for syntactic research

Text corpora have always been used by philologists, historical linguists, and lexicographers; but over the past decades, the availability of large electronic corpora annotated with morphological and syntactic information has significantly extended the possible uses of corpora for syntactic research. (Cf. articles 2 and 4 in this volume for the historical context, and articles 17, 24–34 and 36 for a discussion of corpus annotation.) Before we turn to exploring these interesting possibilities, let us mention some relevant issues and limitations that arise when considering the use of corpora for syntactic research:

First, annotated electronic corpora exist only for very few of the world's languages; for example, the Linguistic Data Consortium (LDC, `http://www.ldc.upenn.edu`) lists corpora for 39 languages, a small fraction of the around 6000 living languages (cf. Crystal, 1997, p. 287). Traditional fieldwork with informants will thus remain the most important methodology for obtaining data from most living languages, at least as a first step (cf., e.g., the Open Language Archives Community, `http://www.language-archives.org/`).

Second, for the languages for which electronic corpora have been compiled and annotated, one needs to keep in mind that even the largest corpora can only represent a finite subset of a language's infinite potential. And given Zipf's law that the frequency of use of the n^{th} most frequently used word (or other phenomenon) in a corpus is inversely proportional to n, even the largest corpus will appear small for linguistic research. In consequence, to address questions involving parts of a language that happen not to occur in a corpus, syntactic research will also have to make use of handcrafted examples.

Finally, one needs to distinguish how data is *obtained* from how data is *evaluated*. Data exemplifying some theoretically interesting pattern can, e.g., be obtained by handcrafting examples, by searching in corpora, or by eliciting data from informants. Data can be evaluated on many dimensions in various qualitative and quantitative ways, e.g., through psycholinguistic experiments, introspection, neuroimaging, or analysis of corpus frequency. Often the evaluation method is independent of how the data to be evaluated was obtained; for example, while it is traditional in generative linguistics to handcraft examples and evaluate them introspectively, it is equally possible to search for interesting examples in corpora and evaluate those introspectively. Other evaluation methodologies, such as quantitative corpus analysis, are dependent on how the data was obtained given that such an analysis relies on representative corpora, a full understanding of the corpus query language and query tool to ensure that the relevant data set is obtained with high precision and recall, and typically a large corpus size to obtain statistically significant results. While in this article we focus on obtaining corpus data, assuming a traditional qualitative syntactic analysis, Stefanowitsch (2005) shows that the often-cited gen-

erative linguistic arguments against a quantitative corpus analysis are questionable, and article 38 in this volume provides a detailed discussion of statistical methods for corpus exploitation.

We turn to the question why it is particularly attractive to make use of corpus searches for syntactic research. To study a syntactic phenomenon, one needs to reduce examples to whatever properties are relevant for the linguistic issues being researched and to vary selected properties in order to explore the grammatical correlations. This is a complex undertaking that assumes an understanding of what properties can play a role for a given linguistic issue—which often is far from clear, as illustrated by the fact that supposedly syntactic effects in recent years have turned out to be explainable by long-overlooked contextual properties (cf., e.g., De Kuthy and Meurers, 2003).

Corpus data obtained by searching for a linguistically relevant pattern exhibits a wide variation of known and unknown parameters and can include information on the context, as needed for exploring the interaction of constraints from syntax and formal pragmatics. When searching for a particular pattern in a corpus, it thus is possible to observe the theoretically interesting pattern within sentences that exhibit a wide variation of lexical, syntactic, semantic, and contextual properties; this makes it possible to obtain a better picture of which of these properties are relevant for a given phenomenon. The fact that corpus examples generally are natural and contextualized can also be helpful whenever examples are to be evaluated through introspection.

Having situated and motivated the use of corpora for syntactic research, we are now ready to address the question how data exhibiting theoretically interesting patterns can be found, and what corpora and annotations are needed to support searching for such patterns. After discussing some basic issues in the next section, we turn to a series of small case studies involving corpora with full syntactic annotation in section 2.

1.2 Basics of syntactically motivated corpus searches

In using corpora for syntactic research, we want to find instances of some pattern of linguistic relevance in order to explore, support, or refute a linguistic claim involving that pattern. Syntactic research, at the fundamental empirical level, observes words, their form, order and cooccurrence in a sentence. The patterns of interest in syntactic research are, however, typically described in terms of generalizations and abstractions over the form and order of words (or groups thereof, for those syntactic paradigms that assume a notion of constituency). This raises the question how a syntactic pattern of interest can be characterized in terms of the properties of a particular corpus and its annotation.

Unannotated corpora, precision and recall of queries The most basic kind of corpus consists of plain text; tokenized, but without linguistic annotations or segmentation. Using such corpora for linguistic research is essentially like using a basic search engine on the web, and indeed the web has gained a significant popularity as an enormous, searchable text repository (cf., e.g., Kilgarriff and Grefenstette, 2004; Lüdeling et al., 2007; and article 55 in this volume). The use of such unannotated corpora for syntactic research requires formulating queries which explicitly list lexical possibilities and spell out entire paradigms given that no generalizations or abstractions can directly be referred to in the query. Since it is complex and often simply impossible to extensionally encode a general syntactic pattern, one has to approximate the intended pattern to be searched, which results in decreased precision and recall.

Precision here measures how many correct matches (vs. false positives) the search for a particular syntactic pattern returns, and *recall* reports how many of the relevant examples in the corpus were found by the search. From our linguistic perspective, a search with low recall for

a particular language pattern means that many instances of the pattern of interest are missed. It can still be sufficient for finding examples counter-exemplifying a particular claim, but for empirically grounding a linguistic theory, the partial empirical blindness caused by searches with low recall is a problem. Searching for a pattern with low precision, on the other hand, means that the search results will contain many false positives that one needs to weed through, generally by hand, in order to find the pattern instances one was actually interested in—which in practice might or might not be feasible.

The utility and caveats of annotation To be able to query more abstract linguistic patterns directly, one can make use of corpora that are annotated with the relevant (or related) linguistic abstractions. Meurers (2005) presents five case studies using a sentence segmented and part-of-speech (POS) annotated newspaper corpus to explore syntactic issues and address claims from the linguistic literature. Meurers discusses how increasingly complex syntactic patterns can be expressed in terms of the properties available in such a corpus. Given the increased availability of corpora with more complex syntactic annotation, the case studies in the present article will focus on the use of *treebanks* for syntactic research, which we turn to in section 2, after discussing some general issues that are relevant in the context of using annotated corpora.

Compared to working with unannotated corpora, some of the mentioned complexity resulting from approximating patterns extensionally and the resulting loss in precision/recall can be avoided by searching in corpora with relevant linguistic annotations. At the same time, the move to using annotated corpora also opens the door to a new problem that can negatively impact precision and recall of queries: errors in the annotation. Even the so-called gold-standard POS or syntactic annotation currently available contains a significant number of errors (cf., van Halteren, 2000; Květǒn and Oliva, 2002; Dickinson and Meurers, 2003, 2005; Dickinson, 2005, and references cited therein). For example, the POS assignment in the widely used Wall Street Journal corpus (WSJ, Marcus et al., 1993) has an estimated 3% error rate. Such annotation errors can result from shortcomings of the annotation scheme, its documentation, or the failure of the human annotators or correctors to apply the annotation guidelines correctly and consistently throughout the corpus. The effect of even a couple of percent of annotation errors on the use of such corpora for syntactic research should not be underestimated. Given Zipf's law, a syntactic pattern of interest can easily have only few occurrences in a corpus. In addition, an error rate such as the 3% mentioned for the WSJ above, is not evenly distributed over all annotation distinctions; instead, certain tokens are unambiguous or trivial to annotate, whereas others distinctions are very difficult to make (and to make consistently). The latter will thus exhibit an error rate many times higher than that of the corpus as a whole. In sum, the annotation errors present in current gold-standard corpora can seriously impact precision/recall of a query relying on distinctions which happen not to be made reliably in the corpus annotation.

A related point concerns the fact that large corpora, traditionally those with one million tokens or more, for practical reasons can only be annotated automatically; and even the annotation of smaller corpora typically arises from a semi-automatic annotation process, where human annotation or correction is based on the output of automatic taggers and parsers. As a result, the fact that current NLP technology cannot reliably make certain distinctions, such as the resolution of argument/adjunct or attachment ambiguities, means that these distinctions will often be incorrect in the annotated corpora or excluded from the annotation scheme to begin with (as seen by the prevalence of flat syntactic annotation in currently available treebanks).

Turning from errors in the application (or the definition or the documentation) of an annotation scheme to the foundation of the annotation itself, one needs to keep in mind that the annotation schemes used are the result of linguistic theorizing and insight. Of course, current

syntactic research frequently questions the established analyses, and a particular set of data might be interesting precisely because the delineation of a phenomenon and/or its analysis are not yet adequately understood. For example, a corpus annotated based on the traditional syntactic assumption that German only allows a single constituent to be fronted naturally would not produce many results for a query referring to this annotation when searching for examples where more than one constituent has been fronted. In a sense, writing queries referring to corpus annotation instead of the corpus data itself is much like writing a travel book based on someone else's photos instead of visiting the place oneself—with all the pros and cons that this entails.

Finally, the use of corpora with structural annotation requires the use of a more sophisticated query language in order to refer to the various linguistic properties and the dominance and precedence relations encoded by the annotation. The case studies we turn to in the next section make use of the TIGERSearch tool (Lezius, 2002), and its query language will be introduced there. The core components our discussion is based on should carry over to most other query languages designed for syntactically annotated corpora (cf., e.g., Pito, 1994; Randall, 2000; Rohde, 2001; McKelvie, 2001; Kepser, 2003; Kallmeyer and Steiner, 2003; Carletta et al., 2006). But before ending this section, let us mention an interesting, somewhat different approach to querying syntactic corpora: the Linguist's Search Engine (Resnik and Elkiss, 2005, `http://lse.umiacs.umd.edu/`), an approach that is exemplified with two linguistic case studies in Resnik et al. (2005). The basic idea of the Linguist's Search Engine is that a query is created by processing and generalizing an example. A parser processes an instance of the pattern one is interested in and the resulting parse tree can be manipulated to obtain a general pattern. That pattern is then used as a query to search in a corpus that has been processed with the same parser. Note that this setup has the interesting property that errors made by the parser do not have to be a problem given that both the initial instance of the search pattern and the corpus are processed with the same tool; the purpose of the parser is not to provide the ultimate linguistic analysis but to provide a link from the instance used to create the search pattern to other instances of that pattern in the corpus.

2 Treebank-based case studies

Following the discussion of the general issues involved, we now turn to three linguistic case studies exemplifying the use of a treebank for syntactic research. We discuss three phenomena of general interest for the architecture of grammar and show that a thorough empirical base is important both for constructing new linguistic analyses and for constructing arguments to support or refute existing theories. We focus on the question how to find the relevant data in corpora and organize the discussion based on an increasing complexity of the query that is needed to obtain the desired types of examples.

The TIGER corpus The case studies are based on the TIGER Corpus (v.1, Brants et al., 2004) and the query tool TIGERSearch (Lezius, 2002, `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/`). The TIGER corpus is a German newspaper corpus consisting of roughly 700,000 tokens (40,000 sentences), taken from the *Frankfurter Rundschau*, a national German newspaper. It was semi-automatically annotated with part-of-speech information (using the Stuttgart-Tübingen-Tagset; Schiller et al., 1995) and syntactic information. The syntactic annotation consists of tree structures with node and edge labels. The trees focus on encoding the argument structure and are relatively flat, e.g., in a prepositional phrase, the preposition, the determiner and the noun are directly dominated by a PP node. The nodes

encode the syntactic category (e.g., NP, PP), and the edge labels are used to encode grammatical functions (e.g., subject, object). There are no empty terminal nodes; instead the annotation scheme allows for discontinuous constituents. For instance, the extraposed relative clause *der lacht* in (1) is annotated as directly dominated by an NP node that also directly dominates the determiner *ein* and the noun *Mann*, but excludes the intervening verb *kommt*.

(1) Ein Mann kommt, der lacht.
a man comes who laughs

The query language of TIGERSearch The query language consists of two levels: nodes and relations. Nodes can be described by Boolean expressions over feature-value pairs. For instance, the query

```
[word="suche" & pos="VVFIN"]
```

finds all words with the orthography *suche* and the part-of-speech tag for finite verbs (`VVFIN`). As values of features, one finds the categories distinguished by the annotation scheme (in double quotes) or so-called types (without quotes), which is an abbreviation; e.g., in the TIGER corpus the type *noun* abbreviates the POS-tags for proper noun and common noun.

Relations between two or more nodes can specify constraints on immediate precedence (`.`), immediate dominance (>), immediate dominance with edge label *L* (>*L*), left corner of a phrase (`>@l`), as well as the derived node relations of general dominance (>*) or siblings (`$`). For example, the following query would find an NP dominating a sentence functioning as a relative clause, such as the one we saw in (1):

```
[cat="NP"] >RC [cat="S"]
```

In addition, a small set of special predicates can be used to describe nodes; for example, the expression `discontinuous(#n)` requires that the terminal yield of the node `#n` is not continuous.

Boolean expressions (without negation) over node relations can be used to form complex descriptions. For example, sentences that contain an S node dominating a particle (`PTKVZ`) and a finite verb satisfy the following query:

```
([cat="S"] > [pos="PTKVZ"])    &  ([cat="S"] > [pos="VVFIN"])
```

Variables are used to express coreference of nodes or feature values. For example, the query above also returns sentences that contain two separate S nodes, one dominating the finite verb and the other one the verbal particle. One can use a variable to state that the same S node is supposed to dominate both nodes:

```
(#n:[cat="S"] > [pos="PTKVZ"]) &  (#n > [pos="VVFIN"])
```

2.1 Case 1: Extraposition, Complex NPs and Subjacency

Turning to the first case study, Chomsky (1986, p. 40; among others) argues that the trace *t* in (2) 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.

(2) [$_{NP}$ 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 noun *books* in (2) in a way that reduces attachment ambiguities, one can obtain parallel German examples which are grammatical:

(3) weil viele Schallplatten mit Geschichten verkauft wurden, die ich noch lesen
because many records with stories sold were that I yet read
wollte.
wanted
'because many records with stories that I wanted to read were sold.'

This 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.

In general, there seems to be no upper limit on the number of phrase nodes that may be crossed by dislocation to the right. Example (4) shows that relative clauses can be extraposed from a deeply embedded NP, and (5) shows the same for a complement clause.

(4) Karl hat mir [eine Kopie [einer Fälschung [des Bildes [einer Frau $\__i$]]]] gegeben,
Karl has me a copy of a forgery of a picture of a woman given
[die schon lange tot ist]$_i$.
who already long dead is
'Karl gave me a copy of a forged picture of a woman who's long been dead.'

(5) Ich habe [von [dem Versuch [eines Beweises [der Vermutung $\__i$]]]] gehört, [daß es
I have of the attempt of a proof of the assumption heard that it
Zahlen gibt, die die folgenden Bedingungen erfüllen]$_i$.
numbers gives that the following conditions satisfy
'I have heard of the attempt to prove the assumption that there are numbers for which the following conditions hold.'

How can we 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 selecting 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. For our TIGER setup, we can express the query as follows:

```
#xp:[cat="NP"] >OC [] &
[cat=("NP"|"PP")] > #xp &
discontinuous(#xp)
```

The three lines of the query have the following meaning:

1. Search for a node of category `NP`; use the variable `#xp` to refer to it. The `#xp` immediately dominates a node functioning as an object clause (`OC`).
2. The `#xp` is immediately dominated by a node that is an `NP` or a `PP`. (Note that immediate dominance is sufficient here since NPs in the TIGER corpus are annotated as flat structures, i.e., the determiner and the noun are sisters in a local tree; for PPs the preposition can also be found in the same local tree.)

3. The `#xp` is discontinuous, in which case the object clause is typically extraposed (but any other discontinuous realization of the `#xp` would also be matched).

Running this query on the TIGER corpus finds examples such as the one in (6).

(6) [...] 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 Patienten
desperate attempt of the homonymous doctor been is his patients
ein für allemal von Kopfschmerzen infolge schlechter Kissen zu befreien].
once for all of headache due to bad pillow to 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.'

It is straightforward to modify this query to find extraposed relative clauses: the labeled dominance constraint >OC in the first line of the query has to be replaced by >RC. To find sentences with extraposed relative clauses that cross one more maximal projection, we can use the following query:

```
#xp:[cat="NP"] >RC [] &
discontinuous(#xp)  &
#yp:[cat=("NP"|"PP")] > #xp &
[cat=("NP"|"PP")] > #yp
```

Here, the additional maximal projection between the topmost `NP` or `PP` and the `#xp` is the node called `#yp`, which is required to be an `NP` or `PP` node itself. This query finds sentences such as the one in (7).

(7) Der 43jährige will nach eigener Darstellung damit [$_{NP}$ den Weg [$_{PP}$ für [$_{NP}$ eine
the 43 year old will after own account thereby the way to a
Diskussion [$_{PP}$ über [$_{NP}$ den künftigen Kurs [$_{NP}$ der stärksten
discussion of the future direction of the strongest opposition
Oppositionsgruppierung]]]]]] freimachen, [die aber mit 10,4 Prozent der
party clear which however with 10,4 percent of the
Stimmen bei der Wahl im Oktober weit hinter den Erwartungen zurückgeblieben
votes at the elections in October far behind the expectations remained
war].
were
'By his own account, the 43 year old thereby wants to clear the way to a discussion of the future direction of the strongest opposition party, which had, however, fallen far behind the expectations by receiving only 10,4 percent of the votes at the elections in October.'

The specification with regard to `#yp` ensures that the extraposition crosses more than one `NP` or `PP` node.

Based on corpus examples such as these, which we take to be well-formed ordinary 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.

2.2 Case 2: The Structure of the German Clause 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 (8) to be grammatical.

(8) * [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.

Based on a syntactically annotated corpus, such as the TIGER corpus used in this study, we can formulate the following query:

```
[pos="PTKVZ"] . [pos=finite]
```

The query looks for a word with part-of-speech `PTKVZ` (separated verbal particle) followed by a `finite` verb (the type `finite` is an abbreviation for the finite auxiliary, modal, and main verbs tags). This query yields 36 sentences for the TIGER corpus, including sentences of the kind we are looking for (9), but also verb-final sentences like (10), which are irrelevant for our issue.

(9) a. 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.'

b. Verloren ging dabei endgültig das Selbstverständnis der Einheimischen.
lost went there.at finally the self-understanding the natives
'Due to this the way the natives saw themselves got finally lost.'

(10) dem Anfang der neunziger Jahre Hohn und Spott zuteil wurde
who beginning of the nineties year derision and sneer part of become
'who was derided at the beginning of the nineties'

To exclude such verb-final sentences, we can extend the query in the following way:

```
#s:[cat="S"]  > #part:[pos="PTKVZ"] &
#part . [pos=finite] &
#s >@l #leftcorner &
#leftcorner:[pos= ! (prorel | prointer | conjunction)]
```

This query searches for a sentence that dominates a verbal particle which is adjacent to a finite verb. The additional constraints rule out certain clause types (relative clauses, embedded interrogative clauses, and subordinated clauses) that are verb-final and thus are not interesting

in the present context. The operator >@l is used to find the leftmost terminal symbol in a tree. The last three conjuncts of the query above state that the leftmost terminal must not be a relative pronoun, an interrogative pronoun, or a conjunction. This query results in a set of examples, all of which are relevant for the question under discussion (i.e., the query has a 100% precision).

In sum, searching for fronted particles in a syntactically annotated corpus provides a range of examples showcasing this supposedly impossible pattern.

2.3 Case 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 (Erdmann, 1886, ch. 2.4; Paul, 1919, pp. 69 and 77). 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 (cf., e.g., Bußmann, 1983, p. 446).

However, as discussed in Müller (2003), there are well-formed example sentences such as those in (11), which according to other constituent tests include more than one constituent in front of the finite verb.

(11) a. [Gar nichts mehr] [mit dem Tabakkonzern] hat Jan Philipp Reemtsma zu tun,[1]
nothing at all more with the tobacco company has Jan Philipp Reemtsma to do
'Jan Philipp Reemtsma has nothing at all to do with the tobacco combine any more.'

b. [Mit ihm] [auf der Anklagebank] sitzen zwei 18-Jährige,[2]
with him on the dock sit two 18 year olds
'Two 18 year olds are in the dock with him . . . '

Müller (2005) proposes that such examples can be analyzed by assuming an empty verbal element as the head of the fronted projection, which therefore can only include dependents of that verb. To explore and test this proposal, we want to search for the pattern in the TIGER corpus and write the following pattern:

```
#s:[cat="S"] >HD #fin:[pos=finite] &
#s >@l #sleftcorner &
#s > #vf1 &
#vf1 >@l #sleftcorner &
#vf1 >@r #vf1rightcorner &

#s > #vf2 &
#vf2 >@l #vf2leftcorner &
#vf2 >@r #vf2rightcorner &
#vf1rightcorner . #vf2leftcorner &

#vf2rightcorner .* #fin
```

This query searches for a node #s with the category S that dominates a finite verb #fin. The node #s has the left periphery #sleftcorner and immediately dominates a node #vf1

[1] From the national German newspaper *taz*, 16.01.2003, p. 6
[2] From the national German newspaper *taz*, 03.04.2003, p. 9

which also has the left periphery `#sleftcorner`. This ensures that the node `#vf1` starts at the same position as `#s`. The right corner of `#vf1` is `#vf1rightcorner`. The query asks for a second node that is also dominated by `#s`, namely `#vf2`. The node `#vf2` has to be adjacent to `#vf1`, which is ensured by the constraint that the node at the right corner of `#vf1` (i.e., `#vf1rightcorner`) immedeatly preceedes the node at the left corner of `#vf2` (i.e., `#vf2leftcorner`). Note that this precedence constraint cannot be encoded directly by a statement like `#vf1 . #vf2`, since the precedence operator `.` compares the left corners of two nodes, which would restrict the `#vf1` node to nodes with exactly one word. Since there are sentences with more than two constituents in front of the finite verb, we do not require that the right edge of `#vf2` is immediately adjacent to `#fin`, but in the last line instead require that the right edge of `#vf2` (i.e., `#vf2rightcorner`) is placed somewhere to the left of the finite verb (`#fin`).

Unfortunately, this query returns several classes of false positives: It admits verb-final sentences containing relative or interrogative pronouns and some other constituent before the verb. And the search results include sentences with complex coordinations of relative or interrogative sentences in which the relative phrase part is not part of the conjunction. Finally, the query also returns examples with adverbials such as the one in (12).

(12) Hier wiederum mangelt es an Opferbereitschaft.
here again lacks it of readiness to make sacrifice
'There is an insufficient readiness to make sacrifices here.'

Such examples have been analyzed differently in the literature and do not constitute evidence for multiple frontings.

Extending the query to eliminate these three classes of false positives results in a rather complex query, which returns six results, two of which are given in (13):

(13) a. [Am schwersten] [mit der Selbstkritik] tat sich Jürgen Kocka.
at the heaviest with the self-criticism did self Jürgen Kocka
'Jürgen Kocka had the most difficulties with self-criticism.'

b. [Negativ] [auf den Gewinn] wirkten sich vor allem
negative on the profit have an effect self before all
Wechselkursschwankungen aus.
exchange rate variations PART
'In particular exchange rate variations had a negative effect on the profit.'

While such examples are illustrative of the phenomenon, a set of six corpus examples is not sufficient to study and reach an understanding of the restrictions and properties of the phenomenon.

We conclude that, as a consequence of Zipf's law, many infrequent but theoretically relevant phenomena can only be found in very large corpora, which given their size cannot be manually annotated or corrected. While searching for the constituency issue discussed in this section requires full syntactic annotation with reliable attachment disambiguation, for other rare phenomena large automatically annotated corpora, such as the 200 million token "Tübingen Partially Parsed Corpus of Written German" (TüPP-D/Z; F. H. Müller, 2004, Ule, 2004) can be an interesting option.

3 Summary

Following an introduction characterizing the context of using corpora in syntactic research, we investigated how unannotated and annotated corpora can be searched to find data exemplifying

patterns of interest to theoretical syntax. Based on three case studies making use of a syntactically annotated newspaper corpus, we illustrated that searching for relevant corpus examples can serve as an important component of empirically grounded syntactic research.

Literature

Brants, S./ Dipper, S./ Eisenberg, P./ Hansen-Schirra, S./ König, E./ Lezius, W./ Rohrer, C./ Smith, G./ Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation* 2(4), 597–620.

Bußmann, H. (1983). *Lexikon der Sprachwissenschaft*. Stuttgart: Alfred Kröner Verlag.

Carletta, J./ Evert, S./ Heid, U./ Kilgour, J. (2006). The NITE XML Toolkit: Data Model and Query Language. *Language Resources and Evaluation* 39(4), 313–334. http://www.ltg.ed.ac.uk/NITE/papers/NXT-LREJ.web-version.ps.

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

Crystal, D. (1997). *The Cambridge encyclopedia of language*. Cambridge, UK: Cambridge University Press, 2. ed.

De Kuthy, K./ Meurers, W. D. (2003). The secret life of focus exponents, and what it tells us about fronted verbal projections. In S. Müller (ed.), *Proceedings of the Tenth Int. Conference on HPSG*. Stanford, CA: CSLI Publications, 97–110. http://ling.osu.edu/~dm/papers/dekuthy-meurers-hpsg03.html.

Dickinson, M. (2005). Error detection and correction in annotated corpora. Ph.D. thesis, Department of Linguistics. The Ohio State University.

Dickinson, M./ Meurers, W. D. (2003). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest, Hungary, 107–114. http://www.aclweb.org/anthology/E/E03/E03-1068.

Dickinson, M./ Meurers, W. D. (2005). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. 322–329. http://www.aclweb.org/anthology/P/P05/P05-1040.

Erdmann, O. (1886). *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung*, vol. 1. Stuttgart: Verlag der J. G. Cotta'schen Buchhandlung. Reprint: Hildesheim: Georg Olms Verlag, 1985.

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

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

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

Kallmeyer, L./ Steiner, I. (2003). Querying treebanks of spontaneous speech with VIQTORYA. *Traitement Automatique des Langues* 43(2).

Kepser, S. (2003). Finite structure query: a tool for querying syntactically annotated corpora. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 179–186.

Kilgarriff, A./ Grefenstette, G. (2004). Introduction to the special issue on the Web as corpus. *Computational Linguistics* 29(3), 333–348.

Květǒn, P./ Oliva, K. (2002). Achieving an Almost Correct PoS-Tagged Corpus. In P. Sojka/ I. Kopeček/ K. Pala (eds.), *TSD 2002*. Heidelberg: Springer, 19–26.

Lezius, W. (2002). Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. Ph.D. thesis, IMS, Universität Stuttgart. Appeared as Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 8, number 4. http://www.ims.uni-stuttgart.de/projekte/corplex/paper/lezius/diss/disslezius.pdf.

Lüdeling, A./ Evert, S./ Baroni, M. (2007). Using Web Data for Linguistic Purposes. In M. Hundt/ N. Nesselhauf/ C. Biewer (eds.), *Corpus Linguistics and the Web*, Amsterdam/New York, NY: Rodopi, vol. 59 of *Language and Computers – Studies in Practical Linguistics*. http://purl.org/stefan.evert/PUB/LuedelingEvertBaroni2005.pdf.

Marcus, M./ Santorini, B./ Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330. http://www.aclweb.org/anthology/J/J93/J93-2004.

Markie, P. (2004). Rationalism vs. Empiricism. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Stanford University. http://plato.stanford.edu/archives/fall2004/entries/rationalism-empiricism/.

McKelvie, D. (2001). XMLQUERY 1.5 manual. Technical report/Web page. University of Edinburgh. http://www.cogsci.ed.ac.uk/~dmck/xmlstuff/xmlquery/index.html.

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. (2004). *Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universität 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*. Tübingen: Max Niemeyer Verlag. 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/ S. Urban (eds.), *Verb-Particle Explorations*, Berlin, New York: Mouton de Gruyter. 119–139. http://www.cl.uni-bremen.de/~stefan/Pub/syn-morph-part.html.

Müller, S. (2003). Mehrfache Vorfeldbesetzung. *Deutsche Sprache* 31(1), 29–62. http://www.cl.uni-bremen.de/~stefan/Pub/mehr-vf-ds.html.

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.

Paul, H. (1919). *Deutsche Grammatik. Teil IV: Syntax*, vol. 3. Halle an der Saale: Max Niemeyer Verlag. 2nd unchanged edition 1968, Tübingen: Max Niemeyer Verlag.

Pito, R. (1994). TGREPDOC. Manual page for tgrep. http://mccawley.cogsci.uiuc.edu/corpora/tgrep.pdf.

Randall, B. (2000). CorpusSearch user's manual. University of Pennsylvania. Technical report/Web page. http://www.ling.upenn.edu/mideng/ppcme2dir/.

Resnik, P./ Elkiss, A. (2005). The Linguist's Search Engine: An Overview. In *Proceedings of the ACL-05 Interactive Poster and Demonstration Sessions*. 33–36. http://www.aclweb.org/anthology/P/P05/P05-3009.

Resnik, P./ Elkiss, A./ Lau, E./ Taylor, H. (2005). The Web in Theoretical Linguistics Research: Two Case Studies Using the Linguist's Search Engine. In *Proceedings of the 31st Meeting of the Berkeley Linguistics Society (BLS-31)*. Berkeley, USA: Berkeley Linguistics Society.

Rohde, D. (2001). Tgrep2. The next-generation search engine for parse trees. Version 1.02. Technical report/Web page. Carnegie Mellon University. http://www-2.cs.cmu.edu/~dr/Tgrep2/.

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

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.

Schiller, A./ Teufel, S./ Thielen, C. (1995). *Guidlines für das Taggen deutscher Textcorpora mit STTS*. Tech. rep., IMS-CL, Univ. Stuttgart and SfS, Univ. Tübingen. `http://www.cogsci.ed.ac.uk/~simone/stts_guide.ps.gz`.

Stefanowitsch, A. (2005). New York, Dayton (Ohio), and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 1(2), 295–301.

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

van Halteren, H. (2000). The Detection of Inconsistency in Manually Tagged Text. In A. Abeillé/ T. Brants/ H. Uszkoreit (eds.), *Proceedings of LINC-00*. Luxembourg.

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*, Tübingen: Stauffenburg Verlag. 211–234.