# Lingua

8

International Review of General Linguistics

Revue Internationale de Linguistique Générale

*Executive editor*

J. Rooryck

*Editors*

N.V. Smith

D. Blakemore

Volume 115, Issue 7 (2005)

ELSEVIER

Amsterdam – London – New York – Oxford – Paris – Shannon – Tokyo

*Lingua* is abstracted/indexed in *Bibliography ie Linguistique/Linguistic Bibliography, Bibliography of Linguistic Literature (BLL), Current Contents/Social and Behavioral Sciences, Linguistics Abstracts Online, Linguistics and Language Behavior Abstracts (LLBA), Modern Language, Social Sciences Citation Index, and Zeitschrift für germanistische Linguistik.*

**Notice**

No responsibility is assumed by the Publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

Although all advertising material is expected to conform to ethical (medical) standards, inclusion in this publication does not constitute a guarantee or endorsement of the quality or value of such product or of the claims made of it by its manufacturer.

CONTENTS

Vol. 115, Issue 7

*Lingua is abstracted/indexed in Bibliographie Linguistique/Linguistic Bibliography, Bibliography of Linguistic Literature (BLL), Current Contents/Social and Behavioral Sciences, Linguistics Abstracts Online, Linguistics and Language Behavior Abstracts (LLBA), Modern Language, Social Sciences Citation Index, and Zeitschrift für germanistische Linguistik, Modern Language Association (MLA) International Bibliography.*

**Articles**

**Book reviews**

0024-3841(200507)115:7;1-B

Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Lingua 115 (2005) 1619–1639
Lingua
www.elsevier.com/locate/lingua

# On the use of electronic corpora for theoretical linguistics Case studies from the syntax of German

W. Detmar Meurers

*Department of Linguistics, The Ohio State University, 222 Oxley Hall, 1712 Neil Avenue, Columbus, OH 43210-1298, USA*

Available online 11 September 2004

**Abstract**

Theoretical linguistics requires example sentences both as empirical basis for the development of theories and as counterexamples to previous generalizations. In addition to obtaining such examples by introspection, electronic corpora can be used to search for examples which are relevant for a particular theoretical issue. This second option is only rarely used in generative linguistics, possibly since it is not fully appreciated that such a use of corpora is in principle independent of the fundamental methodological issues separating empiricists and rationalists.

This paper illustrates with examples from the syntax of German how searching in corpora can help find theoretically relevant examples. Such examples are particularly interesting in that they exhibit a wide variation of potentially relevant parameters. The case studies highlight how linguistic terminology used to single out the relevant phenomenon can be reconstructed in terms of the empirical properties which are accessible directly or through annotations in a corpus.

*Keywords:* Theoretical linguistics; Syntax; Obtaining example data; Corpora; Corpus annotation

A good starting point for this paper are everyday linguistic discussions like the following:

A: Say, is it possible to extract PPs from NPs in German?

0024-3841/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.lingua.2004.07.007

B: Well, something like
*Über Chomsky habe ich eben ein Buch ausgeliehen.*
*about Chomsky have I just now a book borrowed*
sounds fine to me.

A: Hm, but why is
*Mit kurzen Haaren hat Jens eine Freundin.*
*with short hair has Jens a girlfriend*
out then?

B: That's an adjunct PP. It's well known you can't extract adjuncts from NPs.

A: Interesting you should say that since such sentences seem ok in contexts like the following:
*Letzte Woche waren in Düsseldorf wieder die neuesten Haarmoden zu sehen. Mit*
*last week were in Düsseldorf again the newest hair fashions to be seen with*
*kurzen Haaren hat man dieses Jahr nur drei Modelle gezeigt.*
*short hair has one this year only three models shown*
I guess I should have a closer look at such examples to see whether that adjunct generalization is as flaky as it seems.

The conversation introduces an issue of some theoretical relevance, the extractability of PPs from NPs in German. The issue is then explored by (a) coming up with examples for the theoretically interesting pattern and (b) evaluating the grammaticality of examples found in this way. By varying different parameters—whether the PP is a complement or an adjunct, or the effect of a particular context—certain properties which are relevant to the issue are identified and interpreted.

The current debate on linguistic methodology has primarily focused on the aspect (b) of how examples are evaluated, which potentially involves a revision of fundamental beliefs underlying generative linguistics.[1] This issue has largely overshadowed the fact that the aspect (a) of coming up with data relevant to a particular theoretical issue is in principle independent of how such data are evaluated qualitatively (e.g., by introspection or psycholinguistic experiment).[2] Sidestepping the fundamental aspects surrounding evaluation, in this paper we want to focus on the issue of coming up with theoretically relevant example data and explore the potentially useful role electronic corpora can play in this regard. This paper specifically addresses the use of corpus data for theoretical linguistics, i.e., the generative paradigm in a wide sense. It thus shares its motivation with Fillmore (1992), one of the few articles focusing on this topic.[3] For other areas of linguistic research, in particular where questions of language use, cognitive strategies, or language teaching are concerned, the use of corpora is an established methodology—a methodology which, however, differs from what we discuss in this paper since a quantitative data analysis is directly relevant to those research topics.[4]

*Obtaining relevant example data.* The traditional generative method of constructing examples by hand, as in the discussion scenario we started with, makes it possible to reduce examples to whatever is essential to the current discussion and to vary selected properties in order to explore relevant correlations. On the other hand, to obtain a complete example one has to fill the theoretically interesting pattern with lexical material and make many decisions on other syntactic, semantic, and contextual aspects which influence the issue to be tested. It is this task of filling a theoretically relevant pattern with life that searching in electronic corpora under our perspective can assist us with. As mentioned above, this makes no particular assumptions on how the data thus obtained are qualitatively evaluated. An electronic corpus in itself does not provide grammaticality judgments since finding a particular corpus instance is not a proof of the grammaticality of that utterance. This perspective on corpora as provider of examples also means that they will not help in obtaining negative results: just because a corpus does not contain an instance of a pattern, the pattern does not have to be ungrammatical.[5] Finally, the corpus in our setup does not relieve us of coming up with a theoretically interesting linguistic question—if we don't search it with a particular issue in mind, we most likely obtain uninterpreted "data cemeteries" (Marga Reis, p.c.).

Turning to the positive side of things, searching in corpora for a theoretically interesting pattern can provide realistic data with a rich variation of properties filling in the variables of the pattern to be tested. Considering such variation of properties is essential in determining which properties play a role for the pattern and how they correlate. Additionally, such examples can permit access to contextual information, which is playing an increasingly important role in theoretical linguistics. Finally, as natural examples they also include supposedly insignificant or not yet modeled properties, which in our experience makes judging the grammaticality of the relevant pattern tested with these examples significantly easier (for those who want to evaluate the data in this way). In conclusion, data obtained from corpora are a highly valuable source of empirical insights which can help verify linguistic generalizations and serve as a diverse empirical basis for the development and revision of linguistic theories.

In the main part of the paper we want to illustrate with a number of concrete examples from the syntax of German what is involved in using corpus searches to test linguistic claims and support the development of linguistic theories.

[1] See, for example, Abney (1996), Schütze (1996), McEnery and Wilson (1996, Chapter 1.3), and the papers in this issue.

[2] The independence of data gathering and data evaluation only holds when the evaluation is qualitative in nature. A quantitative analysis naturally is dependent on how the data was obtained, whether it is representative with respect to the properties to be evaluated, and related issues.

[3] A more general but related discussion of the relationship of theoretical and computational linguistics can be found in Bayer et al. (1998). The discussion between Borsley and Ingham (2002) and Stubbs (2002) is a related exchange between theoretical and corpus linguists.

[4] See, for example, Johansson and Stenström (1991) and Svartvik (1992).

[5] The absence or scarceness of a particular kind of examples can, of course, be evaluated quantitatively. As with all quantitative analysis, however, this requires additional knowledge about the corpus, its representativeness, and the recall of the search conducted.

## 1. From linguistic descriptions to examples

The setup we used for the examples in this paper is intentionally conservative, both regarding the corpus size and the degree of annotation of the data. It relies on corpora and technology which have been easily accessible since the mid-1990s. We used two German newspaper corpora, one containing 523,353 sentences (8,469,700 words) from the *Donaukurier* and another with 2,621,622 sentences (39,569,709 words) from the *Frankfurter Rundschau*.[6] The corpora were tokenized and tagged so that each corpus position is annotated with its part-of-speech (pos) category, and structural tags were inserted to delimit each unembedded sentence.[7] The part of speech annotation uses the ELWIS tagset (Feldweg, 1995), which has 46 tags and is a predecessor of the now widely used Stuttgart-Tübingen tagset (STTS) discussed in Schiller et al. (1995) and Thielen and Schiller (1996).[8] The freely available tool cqp[9] (Christ, 1994; Christ and Schulze, 1996) was used to store these corpora and provide efficient search functionality.

In order to tap into the empirical treasures hidden in a corpus, one needs to determine how one can search for the theoretically interesting patterns. This amounts to asking how one can translate the characterizations of relevant patterns as used in theoretical linguistics into language properties which can be found in a corpus. To search for examples within our corpus setup, the linguistic characterization of a phenomenon has to be translated to an expression referring to occurrences of (a) word forms and (b) part-of-speech; and those occurrences can be required to (immediately) precede each other or to occur within a certain window, e.g., within five words or within the sentence boundaries.

Turning to the linguistic specifications, for the domain of syntax we are primarily concerned with in this paper, we focus on the following properties used to characterize syntactic patterns: occurrence of a word form or part-of-speech, occurrences of multiple such elements in (pre-theoretic) serial or structural domains, topological fields, syntactic constituency, and grammatical functions. Some of the notions used in generative linguistic research are at a significantly higher level of abstraction than those mentioned here. However, at least for research interested in language outside of a conceptual utopia, one should expect that the terminology used is in principle translatable to actually observable language properties such as the ones discussed in this paper.

Before we turn to the exemplary discussion of how such a translation can be done, we should consider what properties the translation of the linguistic characterizations to the corpus query expressions needs to have in order to be useful for our purposes. There are two criteria: On the one hand, we want to know whether the translation results in the retrieval of sentences which were not characterized by the original pattern, i.e., false positives. If there are no false positives, the translation could be called *sound*; a relative measure of soundness is *precision*. On the other hand, there is the question whether the translation of the linguistic characterization into a corpus query is good enough to retrieve all instances of the linguistic pattern, in which case the translation could be called *complete*. A relative measure of how many of the intended cases we retrieve is the *recall*.

Turning to the first criterion, precision, it does not defeat the purpose of the translation if the query resulting from it retrieves some examples which turn out not to fall under the pattern we are interested in—as long as we overgenerate only to a degree which allows going through the results by hand (or whatever other means) to obtain the actual example instances. Which precision is still acceptable thus depends on the frequency of the pattern and the size of the corpus.

For corpus queries which refer not only to the words and their order in the corpus but also to annotations such as part of speech information, there is a second factor which contributes to the retrieval of false positives: incorrect annotation. Whether and how many of such errors are present in a corpus depends on a variety of factors, in particular, how rich the vocabulary used for annotation is, what empirical properties it refers to and how accessible these properties are, whether all ambiguities are intended to be resolved in the annotation, and whether the annotation is obtained manually, automatically, or by a combination of the two. In principle a text can be annotated with any linguistic notion—in the extreme, the annotation could be identical to or richer than the linguistic notions used to characterize the pattern (in which case one could query the corpus directly with the linguistic characterization). As soon as large amounts of annotated text are required, for example because the particular construction of interest is rare, performing all annotation manually is not feasible. The annotation of larger corpora must therefore be obtained automatically, generally using a tool that has been trained on a smaller, hand-annotated corpus. The accuracy of the automatic tools depends on how much training material is available and how complex it is to find and combine the empirical evidence underlying a particular classification. For the part-of-speech annotation of the corpus we are using in this study, the expected error rate of the hidden Markov model used for tagging is approximately 5%. When using the so-called gold standard corpora, for which generally human post-editing was performed, one can expect around 1.2% annotation errors (Brants, 2000). We return to annotation errors and their consequences in Section 1.3.

The second criterion, recall, is a bit trickier since not retrieving some examples which in principle match the pattern we are interested in amounts to a partial blindness for the diversity of the relevant data set—and, as we argued in the introduction, this diversity is one of the attractive properties of corpus data for theoretical linguistics. On the other hand, every datum we find in addition to the ones obtained by introspection is a gain over the previous situation, as long as we do not draw conclusions based on the apparent absence of particular data. A low recall thus can be acceptable as long as the search yields relevant examples.

Now that we have clearly identified our task, the translation step it involves, and the relevant notions of precision and recall, we proceed to the five small case studies which exemplify what is involved in searching for corpus data for theoretical linguistics.

### 1.1. Word forms and part-of-speech tags

For the first example, we zoom in on a claim made in Suchsland (1994). Suchsland argues that in German perfect tense constructions, Accusativum-cum-Infinitivum (AcI)

[6] The text of these corpora is part of the European Corpus Initiative Multilingual Text I CD-ROM. More information can be found at http://www.ldc.upenn.edu/Catalog/LDC94T5.html.

[7] The corpus preparation was done by Helmut Feldweg (SfS, Tübingen) and Oliver Christ (IMS, Stuttgart).

[8] See also http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html.

[9] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/.

verbs such as *sehen* ('see') or *hören* ('hear') are always realized in their so-called substitute infinitival form instead of as ordinary past participle. This claim is illustrated by Suchsland's example (1).[10]

(1) Er hat$_1$ ihn über die Straße gehen$_3$ **sehen$_2$/*gesehen$_2$**.
he has him over the street go see$_{inf}$/ seen$_{past\text{-}part}$

'He saw him cross the street.'

At stake here is an empirical generalization which involves the occurrence of three words which are connected through head–complement relations: (i) a perfect tense auxiliary selecting (ii) an AcI verb, and (iii) the infinitival complement of (ii).

As first step of translating the relevant pattern into a corpus search for counterexamples, we translate the reference to the class of AcI verbs by two common elements of this class, *sehen* ('see') and *hören* ('hear'). Since our task is to investigate whether counterexamples exist at all, zooming in on a subset of the general pattern is a sensible way to proceed here (it only reduces recall). According to the generalization, the form that does not occur is the past participle of these verbs when they take an infinitival verbal complement. Searching for any instance of the past participles *gesehen* ('seen') or *gehört* ('heard') is insufficient to obtain counterexamples to Suchsland's claim though, since these verbs also exist as ordinary transitive verbs, taking a nominal complement. The translation of the pattern thus needs to include the information that we are only interested in those verbs when they select an infinitival complement. Fortunately, the word order in the German verbal complex is fixed: a verbal head always immediately follows its verbal complement.[11] We therefore can avoid referring to grammatical information like head–complement, which we have no access to in our corpus, by referring to immediate precedence instead of the grammatical relation.

The resulting corpus query searches for occurrences of the AcI past participles *gesehen* ('seen') or *gehört* ('heard') immediately following an infinitive.[12] This is straightforwardly translated to the cqp query `[tpos = "VINF"] ("gesehen" | "gehört")`, which refers to VINF as the part-of-speech tag for an infinitival verb and uses "|" to express a disjunction, i.e., that either of the two AcI verbs in past participle form is to be searched for. Carrying out this search on our newspaper corpora reveals examples such as the following:

(2) Nicht wenige der Anwesenden hatten das Wesen mit der Flasche schon zu
not few of the people present had the being with the bottle already at
vergangenen Anlässen **singen gehört**, so daß sich die Frage, ob es dies nun kann
past events sing heard so that self the question whether it this now can

oder nicht, schon vorher erübrigt hatte.
or not already before been unnecessary had

'Many in the audience had already heard the being with the bottle sing at previous occasions, so that the question whether it can sing or not had already been dealt with.'

(3) so wollen Ohrenzeugen den Eintracht-Trainer schließlich in astreinem Serbo-Hessisch
so want ear-witnesses the Eintracht coach at the end in perfect Serbo-Hessian
vor sich **hinmurmeln gehört** haben
before self murmur heard have

'ear-witnesses claim to have heard the coach murmur this in perfect Serbo-Hessian'

(4) Während er sich den Vorfall nicht erklären kann, wollen Zeugen einen älteren Mann
While he self the incident not explain can want witnesses an older man
**davonfahren gesehen** haben.
drive away seen have

'While he cannot explain the incident, witnesses say an older man drove away.'

(5) Der Präsident des Nationalen Olympischen Komitees (NOK), der mit seinen 79
the president of the National Olympic Committee (NOK) who with his 79
Jahren viele Funktionäre kommen und wenige **gehen gesehen** hat, sprach von
years many officials come and few go seen has spoke of
Herrenmenschen, neuem Kolonialismus und Siegermentalität.
master race new colonialism and winner mentality

'The 79-year-old president of the NOK, who has seen many officials come and few leave, talked about master race, new colonialism and winner mentality.'

How such instances of the supposedly ungrammatical pattern are evaluated in the generative tradition is up to the linguist interpreting the data. Based on an analysis of the properties of these example one can argue that they do indeed constitute valid counterexamples to Suchsland's generalization (cf. Meurers, 2000, Chapter 3.1.1).

For the general issue of this paper the relevant point is, however, a different one; namely that with the help of linguistic background knowledge, it was possible to boil down the initial linguistic characterization of the relevant set of counterexamples—which involves three elements connected by grammatical relations—to a less complex pattern referring only to two immediately adjacent words or categories. Querying the corpus with this reduced pattern provided us with a range of potential counterexamples to the generalization we started out with.

### 1.2. From words to lemmas and pos-tags in basic domains

Our second example is concerned with a pattern that is similar but less constrained than the first in terms of its word order properties and it allows us to illustrate a downside

[10] In this and some of the later examples, subscripts are added to the verbs to clarify the embedding relationship; the most deeply embedded verb has the highest index.

[11] An exception to this rule are the so-called *Oberfeld* and *Zwischenstellung* constructions that play a role in the example of Section 1.3.

[12] Note that the corpus query does not refer to the perfect tense auxiliary as such, but only to the two past participles—despite the fact that the past participle form of a verb in German is also used in passive constructions. This is not a problem here since AcI verbs in German cannot be passivized (Höhle, 1978, p. 172).

of a direct specification of word forms. The theoretical issue concerns the interpretation of modal verbs in German (Kratzer, 1977, 1981; Öhlschläger, 1989). Since a modal verb in German can select a modal verb as verbal complement, a theoretically relevant question is whether all possible readings of modal verbs occur in such embedded contexts. We would therefore like to use a corpus query to explore the question what kind of hypotactic chains of modal verbs in what interpretations are possible in German.

The immediate problem with searching for this pattern is that information on grammatical relations is not part of our corpora so that we cannot directly search for a hypotactic chain of modals, i.e., a modal verb taking another modal verb as complement. One option at this point is to abandon the idea of using such readily available corpora and instead turn to corpora which are annotated for such grammatical relation. We turn to this very attractive possibility in Section 1.5. On the other hand, currently such richer annotations are obtained manually, so that the sizes of corpora and the variety of corpora available in that form is very limited. Since many of the phenomena of theoretical interest in linguistics are very rare, corpus size is a relevant issue for us. It therefore is relevant to explore which kind of linguistic patterns we are able to search for in corpora without more complex syntactic annotations.

For our linguistic pattern of a hypotactic chain of two modal verbs, the most basic idea is to drop the information that one of the modals selects the other modal by only searching for the occurrence of two modal verbs. Implicit in this idea is, however, that these two modal verbs should occur in a limited domain, namely within a single sentence. Basic sentence segmentation can be obtained automatically and is part of our basic corpus setup.

For the six modal verbs *dürfen* ('be allowed to'), *können* ('be able to/be possible'), *mögen* ('may'), *müssen* ('have to'), *sollen* ('shall') and *wollen* ('want to') we can come up with the following cqp expression searching for two occurrences of such verbs within a sentence:

```
[tpos="V.*"  & (word="(ge)?k[aöo]nn.*" | word="(ge)?w[oi]ll.*" |
                word="(ge)?d[aü]rf.*" | word= "(ge)?soll.*" |
                word="(ge)?m[üu][sß]s.*" | word="m[a][g].*" |
                word="(ge)?m[öo][gc].*")]
[]*
[tpos="V.*"  & (word="(ge)?k[aöo]nn.*" | word="(ge)?w[oi]ll.*" |
                word="(ge)?d[aü]rf.*" | word="(ge)?soll.*" |
                word="(ge)?m[üu][sß]s.*" | word="m[a][g].*" |
                word="(ge)?m[öo][gc].*")]
within s
```

The first property of this search expression that probably comes to mind is that it is relatively complex, primarily since it uses the so-called regular expressions to pick out all the different finite and non-finite word forms of the six modal verbs. Note that the same pattern is repeated twice to find two occurrences of such verbs and we allow any number of words ([]*) in-between the two verbs as long as they are within the same sentence (within s). The tpos="V.*" specifying that we are interested in verbs is still relatively transparent, but the regular expressions over the many different word forms which are conjoined (&) to that specification are complex regular expressions, which here approximate the different forms with the help of optionality (? and character classes in square brackets) and the expression. * standing for any sequence of letters.

The complexity arising from the use of regular expressions to characterize the different possible verb forms, and the false matches which can result due to the fact that these expressions specify some restrictions on the possible forms but do not specify them completely,[13] can be avoided if one can refer to the lemma instead of the specific instances. Lemma information can be added to a corpus automatically and is therefore something one can expect of a corpus to be used for theoretical linguistics. Using a corpus with lemma annotations, we can reduce our query to the following:

```
[tpos="V.*" & (lemma="dürfen" | lemma="können" | lemma="mögen" |
               lemma="müssen" | lemma="sollen" | lemma="wollen")]
[]*
[tpos="V.*" & (lemma="dürfen" | lemma="können" | lemma="mögen" |
               lemma="müssen" | lemma="sollen" | lemma="wollen")]
within s
```

For our modal verb example it turns out we can go one step further. The collection of lemmas in the query is not arbitrary, but refers to the modal verbs as a particular subcategory of verbs.[14] If the tagset used for annotation of the corpus is fine-grained enough, this subclass can be referred to directly. While the ELWIS tagset for German does not include a subclassification of verbs, the now widely used STTS tagset includes the relevant distinction. Using a corpus with STTS part-of-speech annotation, we can therefore search for two modal verbs within a sentence in a very straightforward way:

```
[tpos="VM.*"] []* [tpos="VM.*"] within s
```

Searching the *Donaukurier* as the smaller one of our two corpora for the initial pattern results in more than two thousand matches. Browsing through these results reveals that most of these examples are not instances of the pattern we were originally interested in. Approximating the search for a modal verb selecting another modal verb by searching for two modal verbs results in vast overgeneration. Fortunately, looking at the result also reveals the reasons for this overgeneration, namely the occurrence of the comma, *und* ('and'), and *oder* ('or') as coordinating elements between the two modal verbs in the sentence or that of interspersed direct speech. Modifying our search pattern such that it disallows these elements from occurring between the two modal verbs by restricting the []* in the search expressions above reduces the number of search results to 87 sentences,

[13] Of course, these false positives could be eliminated at the cost of making the query even longer—in the extreme case one could just list a disjunction of all possible forms.

[14] Which verbs are part of this class is a matter of definition, not deduction. One could, e.g., additionally include *brauchen* (*need to*).

of which 70 turn out to be actual examples of the linguistic pattern we wanted to find. The following examples illustrate the nature of the modal verb examples found in this way:

(6) Und irgendwann **will** ich auch ein Löschfahrzeug steuern **können**.
and at one point want I also a fire truck steer be able to

'At one point I want to be able to steer a fire truck.'

(7) Ich **möchte** dies nicht entscheiden **müssen**.
I want this not decide must

'I do not want to have to decide this.'

(8) Montags und mittwochs **sollen** sich die Mitarbeiter voll auf die Sachbearbeitung
Mondays and Wednesdays shall self the employees fully on the paperwork
konzentrieren **können**.
concentrate be able to

'On Mondays and Wednesdays, the employees are supposed to be able to concentrate entirely on their paperwork.'

With such examples at hand, the issue of the interpretation of modal verbs in embedded contexts, in particular the range of readings that occur, can be investigated in an empirically informed way. A closely related empirical topic is discussed in Ehrich (2001). The paper is a good example for the effective use of corpus data in theoretical linguistics.

The notion of a sentence as the domain in which we have been looking for two modal verbs is a rather basic, pre-theoretic one. The sentence segmentation in corpora generally is not the result of linguistic deduction but a pragmatic interpretation of the use of punctuation and similar markers. In the following section we explore the role of more linguistic topological domains and how they can be integrated into corpus queries.

### *1.3. Topological fields*

The example of this section takes a closer look at the claim by den Besten and Edmondson (1983) that speakers of Middle-Bavarian, South-Bavarian and Franconian use the otherwise non-existent verbal complex order exemplified by (9) and (10) when they "attempt to sound non-dialect like".

(9) daß er singen$_3$ hat$_1$ müssen$_2$
that he sing has must

'that he has had to sing'

(10) damit unser Lager von einer Lawine nicht getroffen$_4$ hätte$_1$ werden$_3$ können$_2$
so that our camp of an avalanche not hit had been be possible

'so that our camp had not been possible to be hit by an avalanche'

To inspect den Besten and Edmondson's claim that this particular verbal complex word order, the so-called *Zwischenstellung* (Meurers, 2000), is as exceptional as they state, we search for a verbal complex with at least three verbs in which the least embedded verbal head occurs interspersed between its verbal complement and the verbal complement of the complement—instead of following all verbs, as is normally the case, or preceding all of them in the so-called *Oberfeld* (Bech, 1955).

For our translation of the linguistic characterization into a search pattern we can rely on the fact that non-verbal elements generally cannot intervene between the verbs. As for the head–complement relations which are important to distinguish the *Zwischenstellung* from an ordinary verbal complex in the normal head-follows-complement order, if we limit our attention to verb-last sentences, which ensures that the finite verb is part of the verbal complex, we can pick out the least embedded verb in the verbal complex by looking for the finite verb. Based on this reasoning, we arrive at the following search pattern, asking for a verb followed by a finite verb which is followed by either another verb or a particle *zu* and a verb:

```
[tpos = "V.*"] [tpos = "VFIN"] ( [tpos = "V.*"] |
                                  ([tpos = "PTKZU"] [tpos = "VINF"]))
```

Running this search on the *Frankfurter Rundschau* corpus, we obtain 189 examples. Inspection of these sentences shows that 10 of these examples are instances of the pattern we were looking for, such as the ones in (11)–(14).

(11) Der Steinauer ging zuversichtlich in den dritten Quali-Lauf, in dem er gut
the Steinauer went confidently into the third qualifying run in which he well
**abschneiden$_3$ hätte$_1$ müssen$_2$**, um sich für das Finale zu qualifizieren.
finish had have to self for the finals to qualify

'The runner from Steinau confidently went into the third qualifying round, in which he would have had to run well to qualify for the finals'

(12) Nicht daß ich das ernsthaft **bezweifeln$_3$ hätte$_1$ wollen$_2$**.
not that I that seriously doubt had want

'Not that I would have seriously wanted to doubt that.'

(13) laut der der Landeszuschuß nicht bei den Betriebskosten **berücksichtigt$_4$**
according to which the subsidy not for the operating costs considered
**hätte$_1$ werden$_3$ sollen$_2$**
have be should

'according to which the subsidy should not have been considered for the operating costs'

(14) die Ortskernsanierung in Steinkirchen, die sicher 1993 **abgeschlossen$_4$ werden$_3$**
the sanitation of Steinkirchen which surely 1993 completed be
**hätte$_1$ können$_2$**
have could

'the sanitation of Steinkirchen, which surely could have been completed by 1993'

The fact that such examples of the supposedly non-existent word order occur in a national newspaper is a result which sheds doubt on the generalization of den Besten and Edmondson (1983), and one is bound to ask how such verbal complex patterns could be licensed for those speakers who find them grammatical (cf. Kathol, 1998; Meurers, 2000, 2002).

The key question in the context of this paper is a different one though: Why was the precision of the translation of the linguistic pattern into the search expression so low as to produce 189 matches of which only 10 were instances of the intended pattern? An answer to this question has to address two issues: the nature of automatic annotations, and the importance of the notion of a topological domain.

*On the nature of automatic annotations*: The search expression we used above to encode the specific verbal complex pattern relies on part-of-speech annotation to single out the verbs and on the part-of-speech tag distinction between finite and non-finite verbs as a handle on the selection relations among the verbs. However, since the finite verb in a verb-second sentence can be far away from the verbal complex, deciding whether a verb in the verbal complex is finite or non-finite cannot be done accurately by most commonly used taggers, which rely on distributional information from a relatively small window of two or three words.

Lifting this issue to a more general level, many of the phenomena of relevance for theoretical linguistics have a low frequency, so that even though current annotation tools make less than 5% errors, the qualitative nature of the errors which are made can be a significant problem for the use of these annotation for particular searches. Oliva (2001b), Oliva and Petkevic (2001), and Blaheta (2002) argue for the need of a qualitative evaluation of tagging errors and discuss rule-based means to correct some of these errors. Further approaches to error detection and correction are discussed in Dickinson and Meurers (2003). While the current research activity in this area will help reduce the number of annotation errors, one needs to keep in mind that the use of corpora for theoretical linguistics places demands on what distinctions are important which can differ significantly from more mainstream computational uses of corpora. In addition to the differences concerning the kind of distinctions which are relevant, there are also differences concerning the nature of the annotation itself. Many computational uses require full disambiguation, even when not enough information is available to make a deterministic choice. In contrast, for linguistic purposes it appears more sensible to allow for ambiguity preserving annotation (Oliva, 2001a), at least for those ambiguities which cannot be resolved with high accuracy by the efficient algorithms, possibly followed by more costly methods (automatic or manual) for ambiguity resolution. Such a methodology is, e.g., also favored by Karlsson (1992).

*The useful role of topological fields*: Turning to the second issue we wanted to investigate as a cause for the poor precision of the search, the relevant observation is that we did not specify as part of the search pattern that we are only interested in sequences of three verbs that occur as part of the verbal complex. We therefore also obtained examples in which some verbs in the three word sequence had been fronted, extraposed, were part of the so-called *Mittelfeld* (middle field), or occurred in verb-second position.

Considering what is involved when we refer to material as being part of the verbal complex, fronted, extraposed, etc., a model which views a sentence as a sequence of

topological fields is very well suited for encoding the word order contents of such characterizations. The notion of topological fields has played a prominent role in the analysis of surface word order generalizations, particularly for Germanic languages (Herling, 1821; Erdmann, 1886, Drach, 1937; Bech, 1955; Diderichsen, 1966; Engel, 1970; Reis, 1980; Höhle, 1986; Askedal, 1986; Ahrenberg, 1990; Kathol, 2000). Generally speaking, a sentence is divided into a sequence of adjacent, contiguous and non-overlapping areas, the topological fields. These fields play a role similar to that of constituents in generative linguistics, but they are not recursive and form more of a descriptive sentence skeleton, leaving many other issues involved in a constituency analysis (e.g., scope, attachment) underspecified. The basic topological model of German verb-last sentences, for example, consists of a complementizer field, followed by the *Mittelfeld* with arguments and adjuncts in relatively free order, followed by the strictly ordered verbal complex field, and finally a field with the extraposed material (*Nachfeld*).

As discussed by Reis (1980) and Höhle (1986), the different topological fields have clear empirical properties and often a direct correlate in the various theoretical architectures. That the topological field model of sentences is a good interface between word order data and their theoretical interpretation is also recognized in the more recent corpus annotation literature. Stegmann et al. (2000) specify detailed annotation guidelines for a German treebank based on topological fields, and the work reported in Braun (1999), Crysmann et al. (2002) and Müller and Ule (2002) raises the hope that automatically obtained high-quality topological field annotation will become generally available. This would significantly help in using corpora from the perspective of theoretical linguistics. This becomes particularly clear if one considers that the empirical case discussed in this section involved the verbal complex as a topological field—a field which we were able to identify (more or less) because sequences of multiple verbs outside of the verbal complex are relatively rare. Searching for material in fields with less characteristic membership, such as the fronted material in the *Vorfeld*, the freely ordered mixture of elements in the *Mittelfeld*, or extraposed material in the *Nachfeld*, is practically impossible in a corpus without topological or structural annotation.

### 1.4. Constituents

In our discussion of increasingly abstract linguistic notions that can be used to characterize example classes—from word forms via lemmas to part-of-speech tags and topological fields—we now turn to constituency as one of the fundamental notions underlying much work in syntax.

The example of this section goes back to an observation of Müller (1999, p. 376). He mentions that the sentence (15) from the text of Askedal (1984, p. 28) suggests that a past participle and an agentive *von*-PP can sometimes form a constituent (since in German only constituents are assumed to be topicalizable).[15] If this turns out to be the case, it would be a

[15] There are some cases which seem to be counterexamples to the general assumption that topicalization in German involves a (single) constituent Müller (2002b). Note that the so-called partial constituent topicalization phenomenon is not a counterexample; it only shows that constituency is more flexible than is commonly assumed (cf. De Kuthy and Meurers, 2001).

good argument for assuming that German has a passive participle that is distinct from the homonymous past participle.[16]

(15) [Von Grammatikern angeführt] werden auch Fälle mit dem Partizip intransitiver
by grammarians mentioned are also cases with the participle intransitive
Verben.
verbs

'Grammarians also mention cases with the participle of intransitive verbs'

In order to search for a fronted constituent "[*von*-PP passive-participle]" in our basic, part-of-speech annotated corpora, we need to approximate the structure of a *von*-PP and the *Vorfeld* as the topological unit preceding the finite verb in verb-second sentences. This can be done by searching for a sentence starting with *Von*, followed by anything but a finite verb, followed by a noun, a passive/past participle, and the finite (verb-second) verb:

```
<s> "Von" [tpos != "VFIN"]* [tpos = "NN"][tpos = "VPP"]
[tpos = "VFIN"] within s
```

Running this search on the *Donaukurier* corpus shows that the pattern in (15) actually occurs on a regular basis and with different types of passives, such as the *agentive passive* (*Vorgangspassiv*) in (16), the *stative passive* (*Zustandspassiv*) in (17), or a passive embedded under a raising verb in (18).

(16) [Von den Bürgern angeregt] wurde, an der Straße in Richtung Friedhof eine weitere
by the townsmen suggested was at the road in direction cemetery a further
Straßenlampe anzubringen.
street-lamp attach

'It was suggested by the townsfolk to add another street lamp at the road towards the cemetery.'

(17) [Von Baggern umklammert] ist derzeit Riedenburg.
by excavators embraced is currently Riedenburg

'Riedenburg is currently embraced by excavators.'

(18) [Von Pech verfolgt] scheint in dieser Saison Abwehrspieler Dietmar Habermeier
by bad luck followed seems in this season defense player Dietmar Habermeier
zu sein ...
to be

'This season, the defense player Dietmar Habermeier is followed by his bad luck.'

Considering why it was possible to approximate the description of a fronted constituent "[*von*-PP passive-participle]" in this way, one can point to two factors. Firstly, the pattern starts with a specific, obligatory word form, the preposition *von*. And secondly, the fronted

[16] See Müller (2002a, sec. 3.2) for a discussion of the different analyses of the German passive.

constituent we are looking for can be restricted to exclude finite verbs, so that we can approximate the right border of the fronted constituent as the first finite verb we encounter. It therefore is the specific nature of particular constituency-based characterizations which makes it possible to approximate the pattern by references to basic word forms and part-of-speech tags. In consequence, this means that many search patterns involving constituency can only be expressed if one has access to a corpus with richer annotation. Topological field information as discussed in the previous section makes it possible to approximate more constituency-based example characterizations, but other patterns will only be searchable if one has access to full syntactic tree annotations, such as in the NEGRA[17] (Skut et al., 1998), TIGER[18] (Dipper et al., 2001), or VerbMobil (Hinrichs et al., 2000) treebanks for German. High-quality syntactic annotation generally results from manual or semi-automatic[19] annotation efforts, which limits the size of such treebanks. Current work on treebanks is reported in Hinrichs and Simov (2002) and Abeillé (2003). The German treebanks mentioned above and many of those developed for other languages encode not only information about syntactic categories but also about the grammatical relations between these categories—a level of linguistic description which we turn to next.

### 1.5. Grammatical relations

For our last example, we return to the empirical issue we started the paper with, the extractability of PPs from NPs, and probe into a quote from Pafel (1995) which states that

> "arguments of the noun can be extracted, but modifiers cannot:
>
> (19) *Mit rotem Einband habe ich ein Buch gelesen.
> with red cover have I a book read
>
> 'I read a book with a read cover'
>
> [...] Unextractability of noun modifiers is attested at least for English (Huang 1982: 488; Chomsky, 1986: 80), Italian (Giorgi and Longobardi, 1991: 62), and French (Godard, 1992: 238)."[20]

In light of the fact that the basic corpora we used for the examples in this paper do not contain information on constituency or grammatical relations, we again attempt to capture the essential properties in terms of the linear order of word forms and part-of-speech tags. To narrow down the space of possible candidates for PPs, we restrict the search to one of the preposition which heads adjunct PPs, *aus* ('from'), and allow only simple NP structures consisting of a determiner, an optional (modifying) element and the noun head. Parallel to our search in Section 1.4, we look for this pattern from the beginning of the sentence to the

[17] http://www.coli.uni-sb.de/sfb378/negra-corpus/.

[18] http://www.ims.uni-stuttgart.de/projekte/TIGER/.

[19] A well-engineered tool supporting semi-automatic syntactic annotation is the freely available *annotate* http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html).

[20] We added the number, glossing and transliteration to the example.

finite (verb-second) verb. This results in the following cqp search expression, where the structural tag `<s>` fixes that the preposition "Aus" occurs at the beginning of a sentence and the question marks specify the optionality of the article and one additional word:

```
<s> "Aus" [tpos="ART"]? []? [tpos="N.*"] [tpos="VFIN"]
```

The encoding is rather poor in that it not only misses many potential examples as a result of the way we narrowed down the pattern, but it results in 1469 matches for the *Frankfurter Rundschau* corpus of which only a handful of examples turn out to be actual instances of the interesting pattern. Nevertheless, the data we find in this way are striking counterexamples to the above generalization and form the basis of alternative theories for licensing such partial NP constituents (De Kuthy and Meurers, 2001; De Kuthy, 2002):

(20) Aus dem English Theater stehen zwei Modelle in den Vitrinen.
from the English Theater stand two models in the display cases

'Two models from the English Theater are shown in the display cases.'

(21) Aus dem 17. Jahrhundert erklangen in dynamisch differenziertem Spiel und mit.
from the 17th century sounded in dynamic differentiated play and with
weich gestaltendem Ansatz Tanzsätze von Johann Christoph Pezelius und Michael
soft shaped lipping dances by Johann Christoph Pezelius and Michael
Praetorius
Praetorius

'Dances from the 17th century by J.C. Pezelius and M. Praetorius were played in a dynamically differntiated way and with a soft lipping.'

(22) Aus der A-Jugend stoßen Jens Schneider, Thomas Gölzenleuchter und Achim Nau.
from the A-youth come Jens Schneider Thomas Gölzenleuchter and Achim Nau
zu den Aktiven
to the actives

'J.S., T.G. and A.N. from the A-youth join the adult team.'

To overcome the shortcomings of the crude approximation we used in our search pattern for this example, one has to rely on more richly annotated corpora, such as the treebanks mentioned at the end of the previous section. To search in such treebanks, query languages and tools which can refer to syntactic structure or dominance relations have been developed (cf., e.g., Pito, 1994; Brew, 1999; Rohde, 2001; Mckelvie, 2001; König and Lezius, 2000; Kallmeyer, 2000; Steiner, 2001; Kepser, 2003).[21] For example, to search for example patterns such as the one in this section, Kallmeyer (2000) defines a formal

[21] A particularly well-engineered tool, including a graphical query language and import filters for many treebank formats, is the freely available TIGERSearch, cf. http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/.

language which can encode the search for "a prepositional phrase modifying the accusative object and preceding the finite verb (i.e., in the so-called *Vorfeld*), and an accusative object between finite verb and non-finite forms (i.e., in the so-called *Mittelfeld*)."[22] This general encoding of the relevant linguistic pattern also finds examples with richer internal constituent structure such as the example with coordinated NPs in (23), Kallmeyer's search result example (24), or the ones in (25) and (26) reported by Steiner (2001).

(23) In Cockpit und Kabine wurden neue Gehaltsstrukturen mit "marktkonformen"
in cockpit and cabin were new salary structures with market adequate
Anfangsgehältern vereinbart.
starting salaries agreed on

'New salary structures in cockpit and cabin with starting salaries in line with real marked conditions were agreed on.'

(24) Tja, über Flughafenverbindungen habe ich leider keine Information.
well on connections for the airport have I unfortunately no information

'Unfortunately, I have no information on connections for the airport.'

(25) Bezüglich der Unterkunft habe ich schon ein paar Informationen eingeholt.
regarding the housing have I already a few informations gathered

'Regarding the housing, I have already obtained some information.'

(26) Nach Hannover gibt es natürlich stündlich Verbindungen.
to Hannover exists it naturally hourly connections

'There are hourly connections to Hannover.'

This concludes the case studies exemplifying how one can translate theoretically relevant linguistic characterizations to queries referring to language properties found in an annotated corpus. In principle, such queries can be as complete and precise as the linguistic characterizations. In practice, one will often use partial translations which make the most of whatever annotation is available in a given corpus. Such partial translations often are sufficient since the linguistic characterizations we start out from are more precise than necessary to distinguish the set of sentences one is interested in from the others present in the corpus.

## 2. Summary

Example data highlighting theoretically interesting language properties are essential for the construction and validation of linguistic theories. How such data are obtained is in

[22] The query in terms of the German Verbmobil treebank annotation searches for a "node $n_1$ with label PX and grammatical function OA-MOD, a node $n_2$ with label VF that dominates $n_1$, a node $n_3$ with label MF and a node $n_4$ with label NX and grammatical function OA that is immediately dominated by $n_3$."

principle independent of the methodological issues surrounding the question of how natural language examples are or should be evaluated. The purpose of the paper was to illustrate that electronic corpora can be used to search for examples of linguistically relevant phenomena and to discuss what is involved in such a task.

Corpus data were characterized as particularly attractive examples for theoretical linguistics in that they exhibit a wide variation of known and unknown parameters and can include information on the context. To obtain such example data, the linguistic terminology used to single out the relevant phenomenon needs to be reconstructed in terms of the empirical notions which are accessible directly or through annotations in the corpus. This was illustrated with five case studies from the syntax of German, which involved increasingly complex linguistic patterns. Depending on the task, different levels of annotation are needed: from the basic word forms, lemmas, and part-of-speech tags via sentence segmentation and topological fields, to structural annotations and grammatical relations. The increased availability of corpora with linguistically motivated structural annotations makes it possible to search even complex syntactic patterns. In conclusion, this paper illustrates that the use of electronic corpora is a feasible and highly rewarding method for obtaining theoretically relevant example data.

## Acknowledgements

For helpful comments on a draft of this paper I would like to thank Bob Borsley, Frank H. Müller, Stefan Müller, Karel Oliva, Tylman Ule, and the anonymous reviewers. I am also grateful to the OSU College of Humanities for their Seed Grant support.

## References

Abeillé, A. (Ed.), 2003. Treebanks: building and using syntactically annoted corpora. Kluwer Academic Publishers, Dordrecht http://treebank.linguist.jussieu.fr/toc.html.

Abney, S., 1996. Statistical methods and linguistics. In: Judith, K., Philip, R. (Eds.), The Balancing Act: Combining Symbolic and Statistical Approaches to Languages. MIT Press, Cambridge, MA http://www.vinartus.com/spa/95c.pdf.

Ahrenberg, L., 1990. A grammar combining phrase structure and field structure. In: Karlgren, H. (Ed.), Proceedings of the 13th International Conference on Computational Linguistics (COLING), vol. 2, Helsinki, Finland, pp. 1–6.

Askedal, J.O., 1984. Grammatikalisierung und Auxiliarisierung im sogenannten *bekommen/kriegen/erhalten-*Passiv des Deutschen. Kopenhagener Beiträge zur germanistischen Linguistik 22, 5–47.

Askedal, J.O., 1986. Über 'Stellungsfelder' und 'Satztypen' im Deutschen. Deutsche Sprache 14, 193–223.

Bayer, S., Aberdeen, J., Burger, J., Hirschman, L., Palmer, D., Vilain, M., 1998. Theoretical and computational linguistics: toward a mutual understanding. In: John, M.L., Dry, H.A. (Eds.), Using Computers in Linguistics: A Practical Guide, Routledge, London, pp. 231–255.

Bech, G., 1955. Studien über das deutsche verbum infinitum. Historisk-filologiske Meddelelser udgivet af Det Kongelige Danske Videnskabernes Selskab. Bind 35, no. 2, 1955; Bind 36, no. 6, 1957; Kopenhagen. Reprinted 1983. Max Niemeyer Verlag, Tübingen.

den Besten, H., Edmondson, J.A., 1983. The verbal complex in continental West Germanic. In: Abraham, W. (Ed.), On the Formal Syntax of the Westgermania, Linguistik Aktuell, vol. 3. John Benjamins Publishing Co., Amsterdam, pp. 155–216.

Blaheta, D., 2002. Handling noisy training and testing data. In: Proceedings of the Seventh Conference on Empirical Methods in Natural Language Processing. pp. 111–116, http://www.cs.brown.edu/~dpb/papers/dpb-emnlp02.html.

Borsley, R.D., Ingham, R., 2002. Grow your own linguistics? On some applied linguists' views of the subject. Lingua 112, 1–6.

Brants, T., 2000. Inter-Annotator agreement for a German newspaper corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece. http://www.coli.uni-sb.de/~thorsten/publications/Brants-LREC00.ps.gz.

Braun, C., 1999. Flaches und robustes Parsen deutscher Satzgefüge. Diplomarbeit, Fachbereich Computerlinguistik, Universität des Saarlandes.

Brew, C., 1999. An Extensible Visualization Tool to Aid Treebank Exploration. In: Uszkoreit, H., Brants, T., Krenn, B. (Eds.), Proceedings of the Workshop on Linguistically Interpreted Corpora (LINC-99). Association for Computational Linguistics, Bergen, Norway, pp. 49–55, http://www.ltg.ed.ac.uk/~chrisbr/styling-trees.ps.

Christ, O., 1994. A modular and flexible architecture for an integrated corpus query system. In: Proceedings of the International Conference on Computational Lexicography (COMPLEX), Budapest, Hungary. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ:complex94.ps.gz.

Christ, O., Schulze, B.M., 1996. Ein flexibles und modulares Anfragesystem für Textcorpora. In: Feldweg, H., Hinrichs, E.W. (Eds.), Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen, Lexicographica: Series maior, vol. 73. Max Niemeyer Verlag, Tübingen. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ+schulze:tuebingen.94.ps.gz, pp. 121–134.

Crysmann, B., Frank, A., et al., 2002. An integrated architecture for shallow and deep processing. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 02). University of Pennsylvania, Philadelphia, PA, pp. 441–448, http://acl.ldc.upenn.edu/P/P02/.

De Kuthy, K., 2002. Discontinuous NPs in German—A Case Study of the Interaction of Syntax, Semantics and Pragmatics, CSLI Publications, Stanford, CA.

De Kuthy, K., Meurers, W.D., 2001. On partial constituent fronting in German. Journal of Comparative Germanic Linguistics 3 (3), 143–205 http://ling.osu.edu/~dm/papers/dekuthy-meurers-jcgl01.html.

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, pp. 107–114, http://ling.osu.edu/~dm/papers/dickinson-meurers-03.html.

Diderichsen, P., 1966. Helhed og Struktur: Udvalgte Sprogvidenskabelige Afhandlinger, G.E.C. Gads Forlag, Copenhagen, Denmark.

Dipper, S., Brants, T., Lezius, W., Plaehn, O., Smith, G., 2001. The TIGER treebank. In: Hajičová, E. (Ed.), Proceedings of the Third Workshop on Linguistically Interpreted Corpora (LINC-01), Leuven, Belgium. http://www.ims.uni-stuttgart.de/projekte/TIGER/paper/linc2001-abstract-tiger.pdf.

Drach, E., 1937. Grundgedanken der deutschen Satzlehre, 4th edition. Diesterweg, Frankfurt, Wissenschaftliche Buchgesellschaft, Darmstadt, 1963.

Ehrich, V., 2001. Was *nicht müssen* und *nicht können* (nicht) bedeuten können: Zum Skopus der Negation bei den Modalverben des Deutschen. In: Müller, R., Reis, M. (Eds.), 2001. Modalität und Modalverben im Deutschen, Linguistische Berichte, Sonderheft, vol. 9. Helmut Buske Verlag, Hamburg, pp. 140–176.

Engel, U., 1970. Regeln zur Wortstellung. Forschungsberichte des Instituts für deutsche Sprache 5, 9–148 http://www.ids-mannheim.de/pub/forber/fb05.html.

Erdmann, O., 1886. Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung, Erste Abteilung, Verlag der J. G. Cotta'schen Buchhandlung, Stuttgart.

Feldweg, H., 1995. Implementation and evaluation of a German HMM for POS disambiguation. In: From Text to Tags: Issues in Multilingual Language Analysis. Proceedings of the ACL SIGDAT Workshop, vol. 27, March 1995, Dublin, pp. 41–46.

Fillmore, C.J., 1992. "Corpus Linguistics" or "Computer-aided Armchair Linguistics". In: Svartvik, J. (Ed.), Directions in Corpus Linguistics. Trends in Linguistics: Studies and Monographs, vol. 65, Berlin and New York, NY: Mouton de Gruyter. pp. 35–60.

Herling, S.H.A., 1821. Über die Topik der deutschen Sprache. Abhandlungen des frankfurtischen Gelehrtenvereines für deutsche Sprache 394, 296–362.

Hinrichs, E., Bartels, J., Kawata, Y., Kordoni, V., Telljohann, H., 2000. The VerbMobil Treebanks. In: Schukat-Talamazzini, E.G., Zühlke, W. (Eds.), KONVENS-2000 Sprachkommunikation. Ilmenau, Germany: VDE-Verlag, pp. 107–112, http://www.coli.uni-sb.de/~kordoni/papers/treebanks.pdf.

Hinrichs, E., Simov, K. (Eds.), 2002. Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002), Sozopol, Bulgaria. http://www.bultreebank.org/Proceedings.html.

Höhle, T.N., 1978. Lexikalistische Sxntax, Die Aktiv-Passiv-Relation und andere Infinitkonstruktionen im Deutschen, vol. 67, Linguistische Arbeiten, Max Niemeyer Verlag, Tübingen.

Höhle, T.N., 1986. Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder. In: Schöne, A. (Ed.), Kontroversen alte und neue. Akten des VII. Internationalen Germanistenkongresses Göttingen 1985, Max Niemeyer Verlag, Tübingen (Bd. 3), pp. 329–340.

Johansson, S., Stenström, A.-B. (Eds.), 1991. English Computer Corpora, Selected Papers and Research Guide, Mouton de Gruyter, Berlin.

Kallmeyer, L., 2000. A query tool for syntactically annotated corpora. In: Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong, China, pp. 190–198, http://www.sfb441.uni-tuebingen.de/a1/Publikationen/emnlp2000.ps.

Karlsson, F., 1992. Comments on John M. Sinclair: "The Automatic Analysis of Corpora". In: Svartvik, J. (Ed.), Directions in Corpus Linguistics. Trends in Linguistics: Studies and Monographs, vol. 65. Berlin and New York, NY: Mouton de Gruyter. pp. 398–400.

Kathol, A., 1998. Constituency and linearization of verbal complexes. In: Hinrichs, E.W., Kathol, A., Nakazawa, T. (Eds.), 1998. Complex Predicates in Non-derivational Syntax. Syntax and Semantics, vol. 30. Academic Press, New York, pp. 221–270.

Kathol, A., 2000. Linear Syntax, Oxford University Press, Oxford.

Kepser, S., 2003. Finite structure query—a tool for querying syntactically annotated corpora. In: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03). Budapest, Hungary, pp. 179–186, http://tcl.sfs.uni-tuebingen.de/~kepser/papers/fsq.pdf.

König, E., Lezius, W., 2000. A description language for syntactically an-notated corpora. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING-00). Saarbrücken, Germany, pp. 1056–1060, http://www.ims.uni-stuttgart.de/projekte/TIGER/paper/coling2000.pdf.

Kratzer, A., 1977. What 'must' and 'can' must and can mean. Linguistics and Philosophy 1 (3), 337–355.

Kratzer, A., 1981. The notional category of modality. In: Eikmeyer, H.J., Rieser, H. (Eds.), Words, Worlds and Contexts—New Approaches in Word Semantics, Walter de Gruyter, Berlin, pp. 39–76.

McEnery, T., Wilson, A., 1996. Corpus Linguistics. Edinburgh Textbooks in Empirical Linguistics, Edinburgh University Press, Edinburgh, UK.

McKelvie, D., 2001. XMLQUERY 1.5 Manual. http://www.cogsci.ed.ac.uk/~dmck/xmlstuff/xmlquery/index.html.

Meurers, W.D., 2000. Lexical generalizations in the syntax of German non-finite constructions. No. 145 in Arbeitspapiere des SFB 340. Universität Tübingen, Tübingen. (Ph.D. thesis, Universität Tübingen, 1999). http://ling.osu.edu/~dm/papers/diss.html.

Meurers, W.D., 2002. To flip or not to flip: on the nature of irregularities in the German verbal complex. In: Van Eynde, F., Hellan, L., Beermann, D. (Eds.), Proceedings of the Eighth International Conference on Head-Driven Phrase Structure Grammar. CSLI Publications, Stanford, CA, pp. 235–246, http://csli-publications.stanford.edu/HPSG/2/meurers-pn.pdf.

Müller, F.H., Ule, T., 2002. Annotating topological fields and chunks—and revising POS tags at the same time. In: Proceedings of the COLING. http://www.sfs.uni-tuebingen.de/~fhm/Biblio/coling02-345.ps.

Müller, S., 1999. Deutsche syntax deklarativ, Head-Driven Phrase Structure Grammar für das Deutsche, vol. 394, Linguistische Arbeiten, Max Niemeyer Verlag, Tübingen.

Müller, S., 2002a. Complex predicates: verbal complexes, resultative constructions and particle verbs in German. Studies in Constraint-Based Lexicalism, vol. 13. CSLI Publications, Stanford, CA. http://www.dfki.de/~stefan/Pub/complex.html.

Müller, S., 2002b. Multiple frontings in German. In: Jäger, G., Monachesi, P., Penn, G., Wintner, S. (Eds.), Proceedings of the Formal Grammar 2002. Trento, pp. 113–124, http://www.dfki.de/~stefan/Pub/mehr-vf.html.en.

Öhlschläger, G., 1989. Zur Syntax und Semantik der Modalverben des Deutschen, vol. 144, Linguistische Arbeiten, Max Niemeyer Verlag, Tübingen.

Oliva, K., 2001a. On retaining ambiguity in disambiguated corpora. Programmatic reflections on why's and how's. Traitement Automatique des Langues (TAL) 42 (2), 487–500.

Oliva, K., 2001b. The possibilities of automatic detection/correction of errors in tagged corpora: a pilot study on a German corpus. In: Matoušek, V., Mautner, P., Mouček, R., Taušer, K. (Eds.), Proceedings of the Fourth International Conference Text, Speech and Dialogue (TSD 2001), vol. 2166, Lecture Notes in Computer Science, Zelezna Ruda, Czech Republic, September 11–13, Springer, pp. 39–46.

Oliva, K., Petkevič, V., 2001. On the need of *linguistic* linguistic interpretation of corpora. In: Hajičová, E. (Ed.), Proceedings of the Third Workshop on Linguistically Interpreted Corpora (LINC-01), Leuven, Belgium. http://wwwling.arts.kuleuven.ac.be/sle2001/abstracts/web-emp-oliva.html.

Pafel, J., 1995. Kinds of extraction from noun phrases, In: Lutz, U., Pafel, J. (Eds.), 1995. On Extraction and Extraposition in German, Linguistik aktuell, vol. 2. John Benjamins Publishing Co., Amsterdam.

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

Reis, M., 1980. On Justifying Topological Frames: 'Positional Field' and the Order of Nonverbal Constituents in German. Revue de Linguistique 22/23, DRLAV, pp. 59–85.

Rohde, D., 2001. Tgrep2. The Next-generation Search Engine for Parse Trees, Version 1.02. http://www-2.cs.cmu.edu/~dr/Tgrep2/.

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

Schütze, C.T., 1996. The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology, University of Chicago Press, Chicago, IL.

Skut, W., Brants, T., Krenn, B., Uszkoreit, H., 1998. A linguistically interpreted corpus of German newspaper text. In: Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation, Saarbrücken, Germany. http://www.coli.uni-sb.de/~thorsten/publications/Skut-ea-ESSLLI-Corpus98.ps.gz.

Stegmann, R., Telljohann, H., Hinrichs, E.W., 2000. Stylebook for the German Treebank in VERBMOBIL. Verbmobil-Report 239. Universität Tübingen, Tübingen, Germany. http://verbmobil.dfki.de/cgi-bin/verbmobil/htbin/decode.cgi/share/VM-depot/FTP-SERVER/vm-reports/report-239-00.ps.

Steiner, I., 2001. VIQTORIA (A Visual Query Tool for Syntactically Annotated Corpora). Talk at the Conference on Linguistic Data Structures, 22–24 February 2001. University of Tübingen.

Stubbs, M., 2002. On text and corpus analysis: a reply to Borsley and Ingham. Lingua 112, 7–11.

Suchsland, P., 1994. "Äußere" und "innere" Aspekte von Infiniteinbettungen im Deutschen. In: Steube, A., Zybatow, G. (Eds.), 1994. Zur Satzwertigkeit von Infinitiven und Small clauses, Linguistische Arbeiten, vol. 315. Max Niemeyer Verlag, Tübingen, pp. 19–29.

Svartvik, J. (Ed.), 1992. Directions in corpus linguistics. Trends in Linguistics: Studies and Monographs, vol. 65. Berlin and New York, NY: Mouton de Gruyter.

Thielen, C., Schiller, A., 1996. Ein Kleines und Erweitertes Tagset fürs Deutsche. In: Feldweg, H., Hinrichs, E.W. (Eds.), Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen, Lexicographica: Series maior, vol. 73. Max Niemeyer Verlag, Tübingen, pp. 215–226.