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.

© 2005 Elsevier B.V. All rights reserved

This journal and the individual contributions contained in it are protected under copyright by Elsevier B.V., and the following terms and conditions apply to their use:

Photocopying
Single photocopies of single articles may be made for personal use as allowed by national copyright laws. Permission of the Publisher and payment of a fee is required for all other photocopying, including multiple or systematic copying, copying for advertising or promotional purposes, resale, and all forms of document delivery. Special rates are available for educational institutions that wish to make photocopies for non-profit educational classroom use.

Permissions may be sought directly from Elsevier's Rights Department in Oxford, UK; phone (+44) 1865 843830; fax (+44) 1865 853333; e-mail: permissions@elsevier.com. Requests also may be completed on-line via the Elsevier homepage (http://www.elsevier.com/locate/permissions).

In the USA, users may clear permissions and make payment through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA; phone (+1) (978) 7504744, and in the UK through the Copyright Licensing Agency Rapid Clearance Service (CLARCS), 90 Tottenham Court Road, London W1P 0LP, UK; phone: (+44) 20 7631 5555; fax: (+44) 20 7631 5500. Other countries may have a local reprographic rights agency for payments.

Derivative Works
Subscribers may reproduce tables of contents or prepare lists of articles including abstracts for internal circulation within their institutions. Permission of the Publisher is required for resale or distribution outside the institution.

Permission of the Publisher is required for all other derivative works, including compilations and translations.

Electronic Storage or Usage
Permission of the Publisher is required to store electronically any material contained in this journal, including any article or part of an article.

Except as outlined above, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the Publisher.

Address permissions requests to: Elsevier Rights Department, at the fax and e-mail addresses noted above.

Notice
No responsibility is assumed by the Publisher for any injury and/or damage to persons or property as a result of any ideas and/or instructions 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.
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 Otley Hall, 1712 Neil Avenue, Columbus, OH 43210-1288, 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.

© 2004 Elsevier B.V. All rights reserved.

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?
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.\footnote{See, for example, Ashby (1996), Schultze (1996), McEnery and Wilson (1996, Chapter 1.2), and the papers in this issue.}

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.\footnote{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.} 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.\footnote{See, for example, Johansson and Stenström (1991) and Svartvik (1992).}
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 Frankfurter Rundschau and another with 2,621,622 sentences (39,569,709 words) from the Frankfurter Rundschau.\footnote{The text of these corpora is part of the European Corpus Initiative Multilingual Text 1 CD-ROM. More information can be found at http://www.ldc.upenn.edu/Catalog/1DC94T5.html.} 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.\footnote{The corpus preparation was done by Helmut Feldweg (SST, Tübingen) and Oliver Christ (IMS, Stuttgart).} 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 Tholen and Schiller (1996).\footnote{See also http://www.sfs.uni-tuebingen.de/Elwis/sts/sts.html.} The freely available tool epyc\footnote{http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/} (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 work, the automatic 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 Suchland (1994). Suchland argues that in German perfect tense constructions, Accusativum-cum-Infinittum (AfI)
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 Suchland’s example (1).10

(1) Er hat ihn über die Straße gesehen. 
He saw him across 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 Suchland’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 cgs 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 Wezen mit der Flasche schon zu fest 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

10 In this and some of the later examples, substrates 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 past participle—despite the fact that the past participle forms 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).

oder nicht, schon vorher erbittet 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 himmelmeln 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 Komites (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 Suchland’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
of a direct specification of word forms. The theoretical issue concerns the interpretation of modal verbs in German (Kratzer, 1977, 1981; Ohischliger, 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 cop expression searching for two occurrences of such verbs within a sentence:

\[
\text{[tpos="V.*" & (word="(ge)?k\[aöo]\{n\}\{n\}.*"
| word="(ge)?w\{o\}i\{i\}\{l\}.*"
| word="(ge)?d\{s\}\{s\}\{r\}.*"
| word="(ge)?s\{o\}l\{l\}.*"
| word="(ge)?m\{a\}\{l\}\{u\}\{l\}.*"
| word="(ge)?m\{o\}\{o\}\{c\}\{c\}\{l\}.*")]
\]
\]

\text{[tpos="V.*" & (word="(ge)?k\[aöo]\{n\}\{n\}.*"
| word="(ge)?w\{o\}i\{i\}\{i\}.*"
| word="(ge)?d\{a\}\{u\}.*"
| word="(ge)?s\{o\}l\{l\}.*"
| word="(ge)?m\{a\}\{l\}\{g\}\{l\}.*")]
\]

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.

(\text{within s}). \text{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,\textsuperscript{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:

\[
\text{[tpos="V.*" & (lemma="dürfen" | lemma="können" | lemma="müssen" | lemma="sollen" | lemma="wollen")]
\]

\text{[tpos="V.*" & (lemma="dürfen" | lemma="können" | 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.\textsuperscript{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:

\[
\text{[tpos="VM.*" | tpos="VM.*"]}
\]

Searching the \textit{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, and (‘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 \texttt{[]*} in the search expressions above reduces the number of search results to 87 sentences.

\textsuperscript{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.

\textsuperscript{14} Which verbs are part of this class is a matter of definition, not deduction. One could, e.g., additionally include \textit{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 hat, müssen, 
that he sing has must

‘that he has had to sing’

(10) damit unser Lager von einer Lawine nicht getroffen, hätte, werden, können, 
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 = "VF1"] { [tpos = "V.*"] | ([tpos = "PRZU"] [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, hätte, müssen, 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 erstaunlich bezweifeln, hätte, wollen, 
not that I that seriously doubt had want

‘Not that I would have seriously wanted to doubt that.’

(13) laut der der Landesverschuldung nicht bei den Betriebskosten berücksichtigt, 
according to which the subsidy not for the operating costs considered 
hätte, werden, sollen, 
have be should

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

(14) die Ortsersanierung in Steinkirchen, die sicher 1993 abgeschlossen, werden, 
the sanitation of Steinkirchen which surely 1993 completed be 
hätte, können, 
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 Blaha (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 which cannot be 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 involves 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 topizable).14 If this turns out to be the case, it would be a

14 There are some cases which seem to be counterexamples to the general assumption that topologization in German involves a (single) constituent (Müller (2002b)). Note that the so-called partial constituent topologization phenomenon is not a counterexample: it only shows that constituency is more flexible than is commonly assumed (cf. De Keirby and Meurers, 2001).
good argument for assuming that German has a passive participle that is distinct from the homonymous past participle.  

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

‘Grammatikern 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 street-lamp attach

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

(17) [Von Baggern unklammern] 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 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\textsuperscript{17} (Skut et al., 1998), TIGER\textsuperscript{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\textsuperscript{19} annotation efforts, which limits the size of such treebanks. Current work on treebanks is reported in Hinrichs and Simov (2002) and Abellé (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 Pufel (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 red 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).\textsuperscript{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

\textsuperscript{17} http://www.coli.uni-sb.de/stfb378/negra-corpus/

\textsuperscript{18} http://www.ima.uni-stuttgart.de/projekte/TIGER/

\textsuperscript{19} A well-engineered tool supporting semi-automatic syntactic annotation is the freely available \url{annotate http://www.coli.uni-sb.de/stfb378/negra-corpus/annotate.html}.

\textsuperscript{20} We added the number, glossing and transliteration to the example.

\textsuperscript{16} See Müller (2002a, sec. 3.2) for a discussion of the different analyses of the German passive.
finite (verb-second) verb. This results in the following query for a search expression, where the structural tag <es> 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:

<es> "aus" [tpos="ART"]? [I?] [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.'

   from the 17th century sounded in dynamic differentiated play and with
   weich gestaltendem Ansatz Tanzakzente von Johann Christoph Pezelius und Michael
   soft shaped taping 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 dyna-
   mically differentiated way and with a soft tapping.'

(22) Aus der A-Jugend stolzen 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). For example, to search for example patterns such as the one in this section, Kallmeyer (2000) defines a formal

---

22 The query in terms of the German Verbnet treebank annotation searches for a "node n1 with label PX and grammatical function OA-MOD, a node n2 with label VF that dominates n1, a node n3 with label MF and a node n4 with label NX and grammatical function OA that is immediately dominated by n3."
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, Tylinan Ule, and the anonymous reviewers. I am also grateful to the OSU College of Humanities for their Seed Grant support.

References


Bieringer, W., 1983. Max Niemeyer Verlag, Tubingen.


