How to search in corpora for linguistically relevant data

Detmar Meurers
Ohio State University
dm@ling.osu.edu

LOT Winter School 2005, Groningen

Overview

- The "classical method" of obtaining data to verify or develop linguistic theories.
- Which role could corpus data play – and which not?
- What are the entities referred to by linguists to describe a linguistically relevant set of data?
  - word forms and parts of speech and sequences thereof
  - multiple word forms and parts of speech in specific domains
  - topological fields (*Vorfeld*, . . .)
  - constituents and their categories (NP, . . .)
  - grammatical relations (adjunct, . . .)
  - . . .
- How can one search for such entities and what kind of corpus annotation is needed for this?

The "classical method"

In the corridor of a linguistics institute the two linguists A and B meet, coffee mug in hand:

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

B: Well, something like
Über Chomsky habe ich eben ein Buch ausgeliehen.
About Chomsky have I just 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
Last week were in Düsseldorf again the newest hair fashions to
sehen. Mit kurzen Haaren hat man dieses Jahr nur drei Modelle gezeigt.
be seen. With short hair has one this year only three models shown.

I should have a closer look at such examples to see whether that adjunct generalization is as flaky as it seems.

The "classical method" (cont.)

Schütze (1996): "In the absence of anything approaching a rigorous methodology, we must seriously question whether the data gathered in this way are at all meaningful or useful to the linguistic enterprise."

Can corpus data play a role in addressing part of this problem?

Which role can corpus data play?

Searching in corpora for a linguistically relevant phenomenon can provide

- realistic data ⇒ judging grammaticality easier
- which includes (necessary) contexts and
- variation of known and unknown parameters (lexical material, syntactic construction, . . .)
  ⇒ correlations can be observed

Data from electronic corpora can

- help verify linguistic generalizations and
- serve as a broad empirical basis for the development of linguistic theories.

Which role can corpus data play—and which not?

Electronic corpora do not provide

- grammaticality judgments
  ⇒ corpus instance ≠ proof of grammaticality
- negative data
  ⇒ no corpus instance ≠ ungrammatical
- a theoretical interpretation
  ⇒ danger of uninterpreted "data cemeteries"

Can electronic corpus work do more for theoretical linguistics?

Quantitative evaluation of corpus data?

Kefer (1989): "The raised prepositional phrase usually is higher in the definiteness hierarchy than the NP A. [...] Every time that in this book we report a distinction based only on quantitative difference between two variants, the difference is statistically significant."

Touches on a number of fundamental issues:

- acceptability vs. grammaticality

  Scientific evaluation of acceptability judgments using psycholinguistic experiments and statistic methods for their evaluation (Cowart 1997)
- performance vs. competence
- core vs. periphery

One can avoid reopening these fundamental issues by using corpora only as a data supply and not for inferring quantitative judgments.

A very basic setup for corpus searches

Corpora

- 39,5 million words *Frankfurter Rundschau* (FR)
- 8,5 million words *Donaukurier* (DK)

Corpus preparation (Helmut Feldweg, Oliver Christ)

- tokenization
- tagging with ELWIS tagset (46 tags, → STTS)
- sentence segmentation

Search tool: cqp/Xkwic (Oliver Christ, Bruno Schulze)

To search for theoretically relevant examples, the characterization of a phenomenon has to be expressed in terms of

- occurrences of word forms and part of speech tags in
- direct linear sequence or
- linear sequences within a search window.

Word forms and part of speech tags

Generalization: In perfect tense constructions, AcI verbs are always realized in their substitute infinitival form (IPP). (Suchsland 1994)

(1) Er hat$_1$ ihn über die Straße gehen$_3$ **sehen**$_2$.
he has him over the street go see$_{IPP}$
'He saw him cross the street.'

Checking the generalization with a search in FR

- `("gesehen"|"gehört")` ⇒ 7982 matches
- `[tpos = "VINF"] ("gesehen"|"gehört")` ⇒ 8

(2) Nicht wenige der Anwesenden hatten das Wesen mit der Flasche
not few of.the people.present had the being with the bottle
schon zu vergangenen Anlässen singen **gehört**, so daß sich die Frage,
already at past events sing heard so that self the question
ob es dies nun kann oder nicht, schon vorher erübrigt hatte.
whether it this now can or not already before been.unnecessary had
'Many of the people present 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) Ich hab's in meiner Schulter krachen **gehört** – es hat höllisch weh getan,
I have.it in my shoulder crack heard – it has hell.like hurt done
sagte der 24jährige Kölner.
said the 24-year-old man.from.K.
'I heard it crack in my shoulders – it hurt like hell, said the 24 year old man from Cologne.'

(4) Die Frau hatte einen dumpfen Schlag sowie Münzgeld klimpern
the woman had a muted hit as.well.as coins thrumming
gehört und sofort die Polizei verständigt.
heard and immediately the police contacted
'The woman had heard a muted hit as well as thrumming coins and immediately contacted the police.'

(5) Ko Murobushi hat Tatsumi Hijikata tanzen **gesehen**.
Ko Murobushi has Tatsumi Hijikata dance seen
'Ko Murobushi has seen Tatsumi Hijikata dance.'

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

Word occurrences in domains

Exploration of a phenomenon: What kind of hypotactic chains of modal verbs in what interpretations are possible in German?

Search in the DK

How does one find a hypotactic chain of modals?

A) Restrict problem to: two occurrences in 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                                                    2053 matches
```

Expressing queries in terms of regular expressions on word forms is error prone and cumbersome ⇒ lemmatization of corpora very useful

Word occurrences in domains (cont.)

B) Additionally eliminate the following material in-between two occurrences of a modal verb in a sentence:

- commas, begin/end of direct speech
- coordinating elements

87 matches (70 actual examples)

(7) 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.'

(8) Ich **möchte** dies nicht entscheiden **müssen**.
I want this not decide must
'I do not want to have to decide this.'

(9) Montags und mittwochs **sollen** sich die Mitarbeiter voll auf die
Mondays and Wednesdays shall self the employees fully on the
Sachbearbeitung konzentrieren **können**.
paperwork concentrate be.able.to
'On Mondays and Wednesdays, the employees are supposed to be able to concentrate entirely on their paperwork.'

(10) In Eichstätt **sollen** die Kinder mehr mitbestimmen **dürfen**,
in Eichsätt shall the children more decide be.allowed.to
etwa beim Straßenbau.
for.example concerning building.of.streets
'In Eichstätt the children are supposed to be allowed to decide more, for example concerning which streets are built.'

Topological fields

Generalization: Speakers of Middle-Bavarian, South-Bavarian and Franconian use an otherwise inexistent verbal complex order when they "attempt to sound non-dialect like". (Den Besten and Edmondson 1983)

(11) daß er singen$_3$ hat$_1$ müssen$_2$
that he sing has must
'that he has had to sing'

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

Checking the generalization with a search in FR

A sequence of three immediately adjacent verbs outside of the verbal complex is rare, so we try the query:

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

189 matches (10 actual examples)

(13) der Glaube, daß jener Clan, der als nächster Mogadischu kontrolliert, sich
the belief that the clan that as next Mogadischu controls self
nach dem Vorbild der Marehan von Siad Barre genauso bereichern$_3$
after the model of.the Marehan of Siad Barre equally enrich
wird$_1$ können$_2$
will be.able
'the belief that the clan which controls Mogadischu next will be able to enrich following the model of Siad Barre'

(14) Zu dem Zeitpunkt, an dem ich mich entscheiden$_3$ hätte$_1$ müssen$_2$, war
at the time at which I me decide had have was
das Gesangsbuch wichtiger.
the hymn.book more.important
'At the time at which I would have had to decide, the hymn book was more important to me.'

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

Direct reference to topological fields

Leaving out the topological field information results in low precision, e.g., due to

1. other example patterns matching the query:

- topicalized [V V] followed by finite verb-second
- finite verb-last followed by extraposed [V V]
- [V V] followed by extraposed intransitive V
- special constructions, e.g.,

(16) Doch gelernt ist gelernt.
but learned is learned
'Things you really learned, you remember.'

(17) Von der Sowjetunion lernen heißt siegen lernen
of the Soviet Union learn means win learn
'To learn from the Soviet Union means to learn how to win.'

Direct reference to topological fields (cont.)

Other topological field notions (*Mittelfeld*, *Nachfeld*) are impossible to translate into words and POS-tags.

⇒ Annotating corpora with topological information is highly useful for linguistically motivated corpus searches. (→ structural tags, treebanks)

2. erroneous corpus annotation, e.g., tagging errors

⇒ Annotation tools (taggers, shallow parsers) usually geared towards disambiguation at any cost. For linguistic searches preferable to preserve certain ambiguities.

Constituents

Exploration of a phenomenon: Müller (1999) mentions an example with a topicalized constituent consisting of a past participle and an agentive *"von [by]"*-PP

(18) [Von Grammatikern angeführt] werden auch Fälle mit dem Partizip
of grammarians mentioned are also cases with the participle
intransitiver Verben.
intransitive verbs
'Grammarians also mention cases with the participle of intransitive verbs'

Is a "[*von*-PP passive-participle]" constituent generally available with passives?

Searching for constituents

Search in DK corpus requires approximation of

- the structure of a *von*-PP
- the *Vorfeld* as topological field before the finite verb

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

35 examples

Agentive passive (*Vorgangspassiv*)

(19) [Von ihrer 21 Monate alten Enkelin ausgesperrt] wurde Montag
by her 21 months old granddaughter lock.out was Monday
mittag eine 58jährige Hausfrau aus der Mercystraße.
noon a 58-year-old housewife from the Mercystreet
'On Monday at noon, a 58 year old house wife living on Mercystreet was locked out by her 21 month old granddaughter.'

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

Stative passive (*Zustandspassiv*)

(21) [Von den Entwicklungen auf dem Arbeitsmarkt besonders betroffen] sind
by the developments at the job-market particularly affected are
laut Arbeitsamt Ingolstadt Männer und ausländische
according-to labor-exchange Ingolstadt men and foreign
Arbeitnehmer.
employees
'Labor exchange at Ingolstadt reports that the current development of the work market particularly affected men and foreign workers.'

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

Stative passive embedded under raising verb (or idiomatic)

(23) Von Pech verfolgt scheint in dieser Saison Abwehrspieler Dietmar
By bad.luck followed seems in this season defense.player Dieter
Habermeier zu sein . . .
Habermeier to be
'This season, the defense player Dietmar Habermeier is followed by his bad luck.'

Other passive

(24) [Von einem Unbekannten verfolgt] fühlt sich ein Imker aus
by a person.unknown followed feels himself a bee-keeper from
Bad Abbach.
Bad Abbach
'A bee-keeper from Bad Abbach feels followed by a person unknown.'

Grammatical relations

Generalization Pafel (1995): "[A]rguments of the noun can be extracted, but modifiers cannot:

(25) * Mit rotem Einband habe ich ein Buch gelesen.

[. . .] Unextractability of noun modifiers is attested at least for English (Huang 1982:488; Chomsky 1986:80), Italian (Giorgi & Longobardi 1991: 62), and French (Godard 1992: 238)."

Checking the generalization with a search in FR

A) Restrict search to specific prepositions followed by a simple NP structure at the beginning of a sentence, i.e., before a finite verb:

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

1469 matches

(26) 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.'

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

Direct reference to grammatical relations

B) Use a treebank with special query tools:

- VERBMOBIL treebanks (Tübingen, OSU)
- NEGRA/TIGER treebanks & search tools (Saarbrücken, Stuttgart)

  Tree description language to specify linear order, dominance, grammatical functions, . . .

  Regarding our example: Search a PP modifier, a finite verb to the right of it and a NP modified by the PP to the right of the verb, so that the PP is only dominated by constituents which also dominate the finite verb.

  $\Rightarrow$ Also finds examples with richer internal constituent structure, e.g., coordinated NPs

(28) In Cockpit und Kabine wurden neue Gehaltsstrukturen mit
in cockpit and cabin were new salary.structures with
"marktkonformen" Anfangsgehältern vereinbart.
market.adequate starting.salaries agreed.on
'New salary structures in cockpit and cabin with starting salaries in line with real marked conditions were agreed on.'

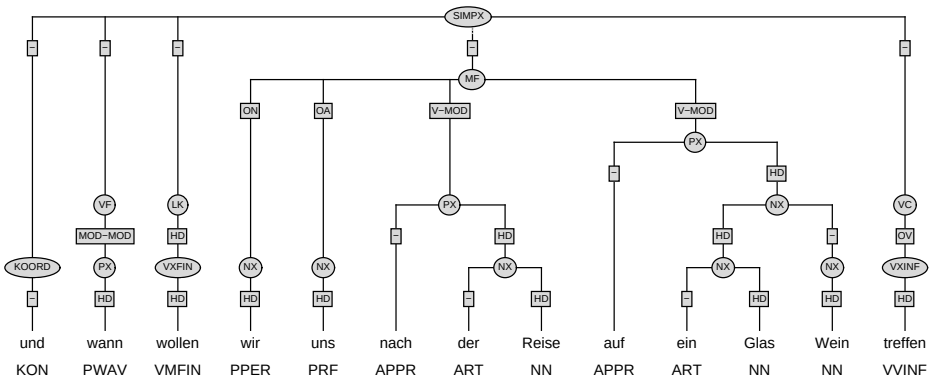
Syntactic Annotation for German

- German Verb*mobil* treebank (Hinrichs et al. 2000; Stegmann et al. 2000):
  - spoken language: dialogs in which two discourse participants negotiate business appointments.
  - 38.000 syntactic units (dialog turns)
  - flat structures based on topological fields
- Negra Treebank (Skut et al. 1997, 1998)
  - written language: *Frankfurter Rundschau*, a national newspaper
  - 20.000 sentences (350.000 tokens)
    TIGER Treebank (Brants et al. 2002): $>$ 35.000 sentences (with refined annotation)
  - flat structures as encoding of argument structure

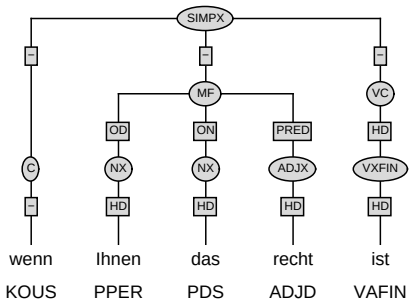
The German Verb*mobil* Treebank

- annotation consists of tree structures with node and edge labels
- tree structure:
  - encodes
    - topological field structure at top-level
    - syntactic categories
  - properties:
    - no branching edges and each daughter has one mother (some secondary edges)
    - no empty terminal nodes
- node and edge labels encode:
  - node labels
    - sentence level: turn type
    - field level: topological field names
    - phrase level: syntactic categories
    - lexical level: STTS part-of-speech (Schiller, Teufel, and Thielen 1995)
  - edge labels on phrase level: grammatical functions

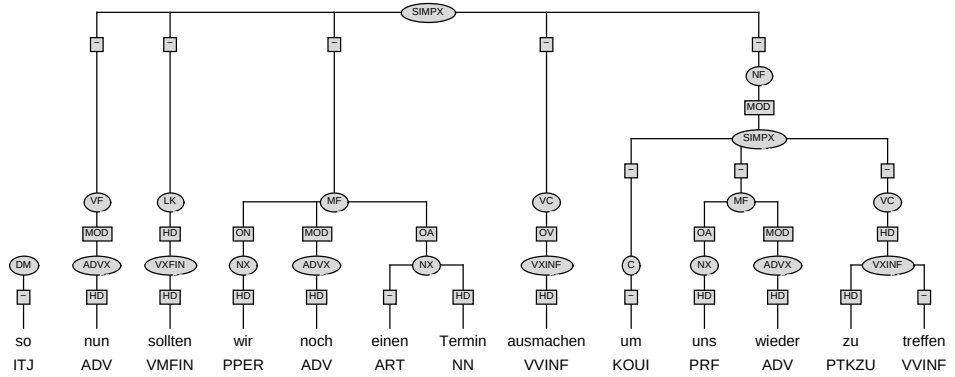
A Basic Example



A Verb-Last Example



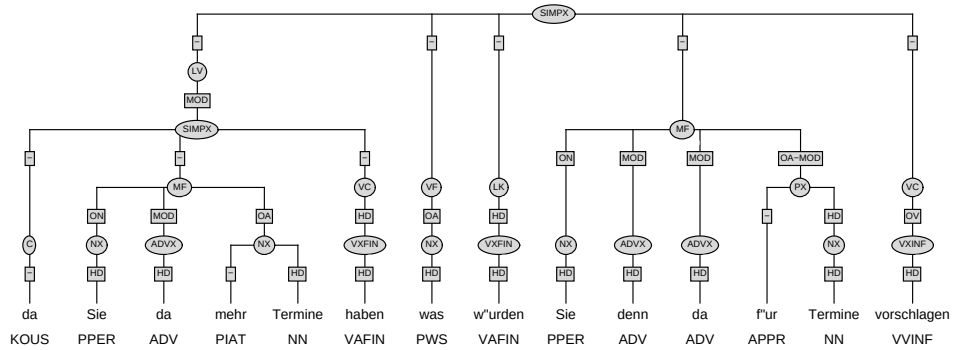
An Example with Embedding



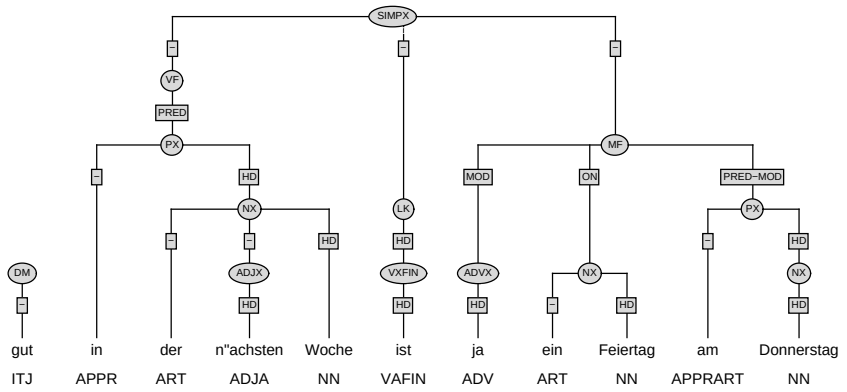
Dependencies across Categories/Fields

- remember: no crossing branches
- where dependency relations cross the border between constituents or topological fields, reference is encoded by special naming conventions for edge labels
- examples:
  - OA-MOD is a modifier of an OA occurring somewhere in the sentence
  - PRED-MOD is a modifier of a PRED

OA-MOD Example



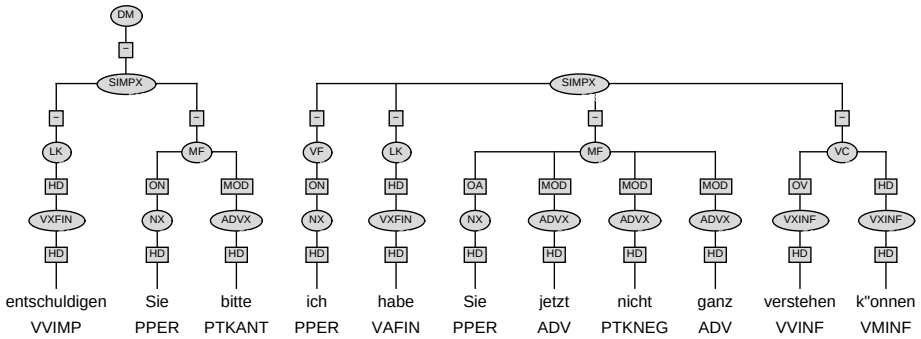
PRED-MOD Example



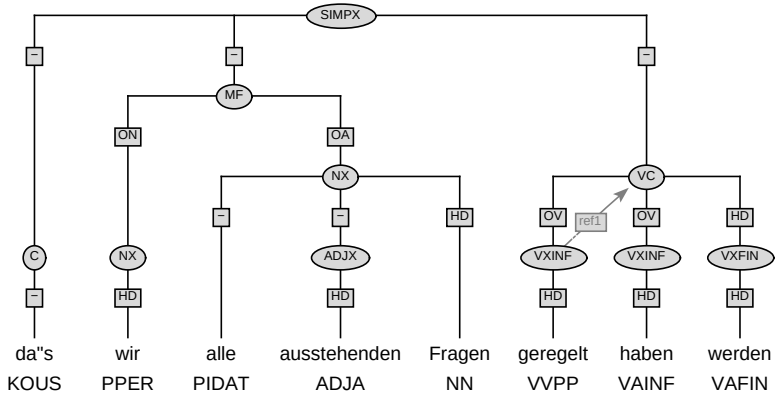
Position of Finite Verb and the Verbal Complex

- finite verb is annotated in the left bracket or verbal cluster, following surface order
- verbal complex is flat
- secondary edges mark relations in a verbal complex with more than two verbs
- relations between arguments and different verbs in a cluster not distinguished (hardly any complex non-finite constructions in Verb*mobil* data)

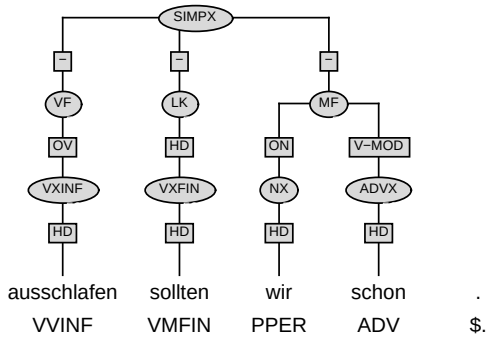
Verbal Complex with Two Verbs



Verbal Complex with Three Verbs



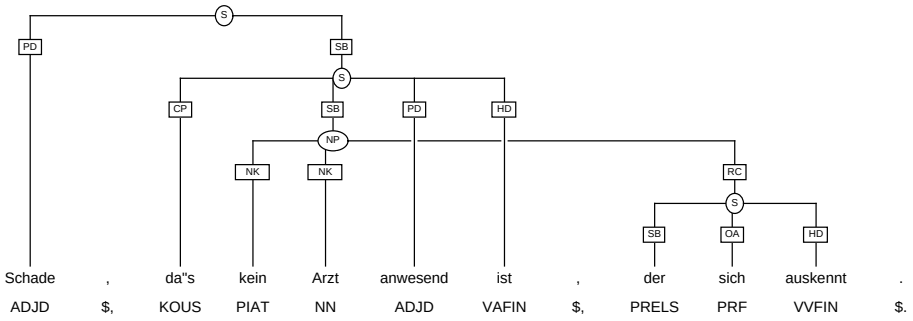
Fronting of a Non-Finite Verb



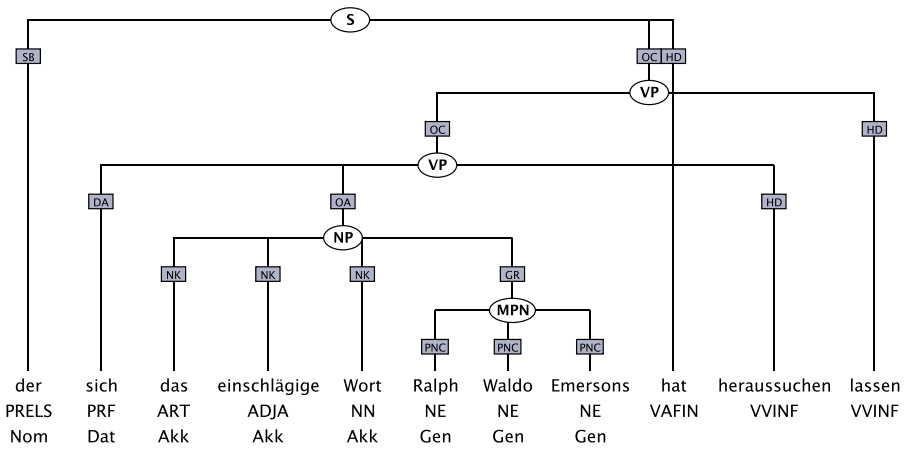
The NEGRA Treebank

- annotation consists of tree structures with node and edge labels
- tree structure:
  - encodes argument structure
  - properties:
    - branching edges used extensively
    - no empty terminal nodes
    - each daughter has one mother (but some secondary edges)
- node and edge labels encode:
  - phrase level: syntactic categories
  - lexical level: STTS part-of-speech (Schiller, Teufel, and Thielen 1995)

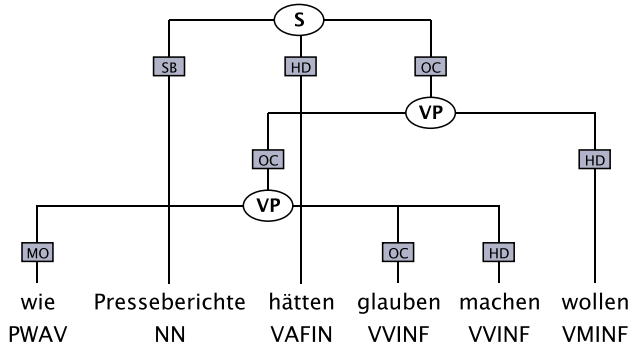
Example



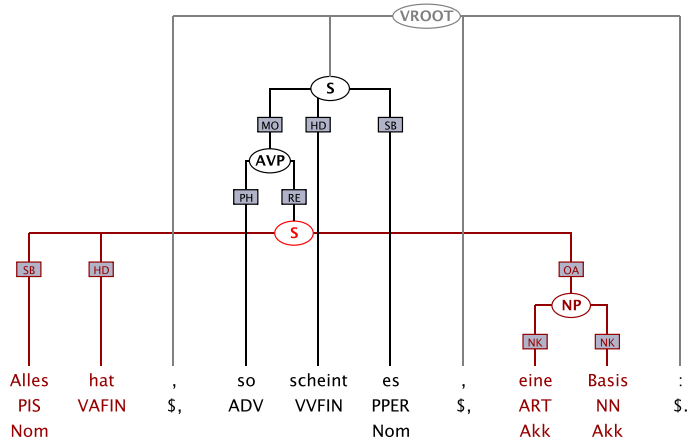
Verbal Complex (I)



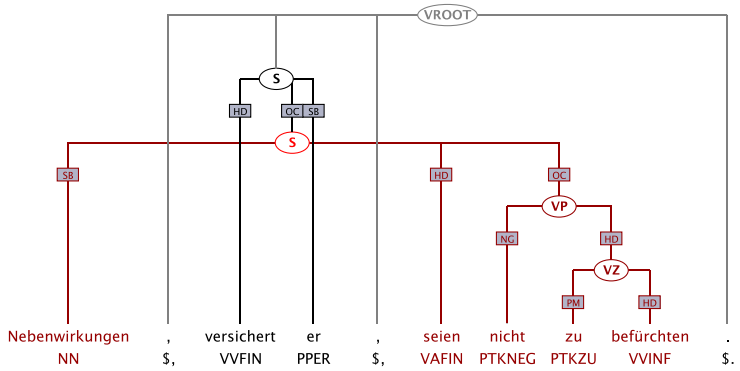
Verbal Complex (II)



Parentheticals (I)



Parentheticals (II)



How about the Syntax-Information structure interface?

- We can already search for:
  - lexical correlates of focus, e.g., focus sensitive particles (*only*, . . .)
  - word order effects of focus, e.g., topicalization in English, word order in Slavic languages
  - annotation of intonation, e.g., GTOBI labeled data as the IMS Radionewscorpus or the German VM corpus
- In the future, one can hope to be able to use
  - corpora with information structure annotation (cf. some projects in Saarbrücken and Potsdam)
  - corpora which integrate intonation, syntactic, and information structure information
  - search tools supporting search in time aligned, graph structured data (e.g., EMU, but limited expressivity)

Summary

Electronic corpora can be used to search for examples of linguistically relevant phenomena in order to

- verify generalizations or
- obtain a wide empirical basis exemplifying the phenomenon.

Corpus data are attractive since they

- exhibit a wide variation of known and unknown parameters and
- come with a context.

The linguistic terminology used to single out the relevant phenomenon needs to be reconstructed in terms of the annotation of the corpus.

Danger of selective results due to the translation of the characterization of the phenomenon to the query language.

Summary (cont.)

Depending on the task, the following levels of annotations are needed:

- basic: word forms, part of speech tags, lemmata, simple structural tags (e.g., sentence boundaries)
- topological fields and constituents: special structural tags or treebanks
- grammatical relations: treebanks

While many corpus annotations (lemmata, part of speech tags, sentence segmentation, shallow parsing chunks) can be obtained automatically, different from standard use, annotation tools for linguistic purposes should

- allow for ambiguity preserving annotation (for ambiguities which cannot be resolved with high accuracy by the efficient algorithms),
- possibly followed by more costly algorithms for ambiguity resolution.

References

Brants, Sabine, Dipper, Stefanie, Hansen, Silvia, Lezius, Wolfgang, and Smith, George (2002). The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria.

Cowart, Wayne (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage Publications.

Den Besten, Hans and Edmondson, Jerold A. (1983). The Verbal Complex in Continental West Germanic. In W. Abraham (Ed.), *On the Formal Syntax of the Westgermania*, Volume 3 of *Linguistik Aktuell*, pp. 155–216. Amsterdam: John Benjamins Publishing Co.

Hinrichs, Erhard, Bartels, Julia, Kawata, Yasuhiro, Kordoni, Valia, and Telljohann, Heike (2000). The Tübingen Treebanks for Spoken German, English, and Japanese. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Artificial Intelligence, pp. 552–576. Berlin: Springer.

Kefer, Michel (1989). *Satzgliedstellung und Satzstruktur im Deutschen*, Volume 36 of *Studien zur deutschen Grammatik*. Tübingen: Gunter Narr Verlag. (= Phd Thesis, Lüttich, Belgium, 1983).

Meurers, Walt Detmar (2000). *Lexical Generalizations in the Syntax of German Non-Finite Constructions*. Number 145 in Arbeitspapiere des SFB 340. Tübingen: Universität Tübingen. (= Ph. D. thesis, Universität Tübingen, 1999). http://ling.osu.edu/~dm/papers/diss.html.

Müller, Stefan (1999). *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche*. Number 394 in Linguistische Arbeiten. Tübingen: Max Niemeyer Verlag.

Pafel, Jürgen (1995). Kinds of Extraction from Noun Phrases. In U. Lutz and J. Pafel (Eds.), *On Extraction and Extraposition in German*, Volume 2 of *Linguistik aktuell*. Amsterdam/Philadelphia: John Benjamins Publishing Co.

Schiller, Anne, Teufel, Simone, and Thielen, Christine (1995). Guidlines für das Taggen deutscher Textcorpora mit STTS. Technical report, IMS-CL, Univ. Stuttgart and SfS, Univ. Tübingen. http://www.cogsci.ed.ac.uk/~simone/stts_guide.ps.gz.

Schütze, Carson T. (1996). *The empirical base of linguistics: grammaticality judgments and linguistic methodology*. Chicago, IL: The University of Chicago Press.

Skut, Wojciech, Brants, Thorsten, Krenn, Brigitte, and Uszkoreit, Hans (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

Skut, Wojciech, Krenn, Brigitte, Brants, Thorsten, and Uszkoreit, Hans (1997). An Annotation Scheme for Free Word Order Languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*. Washington, D.C. http://www.coli.uni-sb.de/~thorsten/publications/Skut-ea-ANLP97.ps.gz.

Stegmann, Rosmary, Telljohann, Heike, and Hinrichs, Erhard 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.

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

Thielen, Christine and Schiller, Anne (1996). Ein kleines und erweitertes Tagset fürs Deutsche. In H. Feldweg and E. W. Hinrichs (Eds.), *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, Volume 73 of *Lexicographica: Series maior*, pp. 215–226. Tübingen: Max Niemeyer Verlag.