On Corpora and Theoretical Linguistics

Detmar Meurers
Universität Tübingen

DoSciLa – Doctoriales en Sciences du Langage: “La Langue dans tous ses états”
Université Paris Diderot, Paris 7
March 18, 2011

Overview

Preliminaries

- The “classical method” of obtaining data to verify or develop linguistic theories.
- Which role can corpus data play – and which not?

Bridging the gap

- What are the entities referred to by linguists to describe a linguistically relevant set of data?
  - word forms and parts of speech
  - sequences thereof, multiple occurrences in domains
  - topological fields (*Vorfeld*, . . . ; cf. fronted, extraposed)
  - 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 out of NPs in German?

B: Well, something like “*Über Chomsky habe ich eben ein Buch ausgeliehen. [About Chomsky have I just now a book borrowed.]*” sounds ok.

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 is well known you can never extract adjuncts from NPs.

A: Really? How do you explain it’s good in the following context then: “There was a big hair fashion show in Düsseldorf. *Mit kurzen Haaren hat man dieses Jahr nur drei Modelle gezeigt. [With short hair has one this year only three models shown.]*”

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.”

Which role can corpus data play in addressing this problem?

Which role can corpus data play?

Searching in corpora for linguistically relevant phenomena can provide

- realistic data ⇒ judging grammaticality easier
- which includes (optional or obligatory) 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.

. . . – and which not?

Electronic corpora do not provide:

- grammaticality judgments
  ⇒ corpus instance ≠ proof of grammaticality
- negative data
  ⇒ no corpus instance found ≠ ungrammatical
  but: statistical analysis on the right corpus can establish underuse or unavailability of a pattern
- a research question or theoretical interpretation
  ⇒ danger of uninterpreted "data cemeteries"

Obtaining vs. interpreting/evaluating the data

- It is useful to clearly separate:
  - how to obtain examples, e.g., from corpora
  - how to evaluate/interpret examples
- Corpus data can be evaluated in different ways, including introspective grammaticality judgements.
- Relationship of data and theory is complex:
  - acceptability vs. grammaticality
  - performance vs. competence
  - core vs. periphery
- Focus of talk is on obtaining data.
  - How can one access example classes which are of relevance for a particular linguistic research question?
  - How can one formulate effective queries and what annotation does it require?

  = Translating from the characterization of the phenomenon of interest to the data and its annotations in the corpus.

Corpus querying: Precision of search

- **Precision** of search for a pattern of interest:
  - Of the results to the query, how many represent the learner language patterns searched for?
  - False positives can result in two ways:
    - Term used for query also characterizes patterns other than the ones we are interested in.
    - Some of the annotations the query refers to are incorrect.
- Requirements on precision of search
  - for qualitative analysis: Needs to be high enough to find relevant examples among the false positives.
  - for quantitative analysis: For reliable results, very high precision is required, in particular where specific rare language phenomena are concerned (and as known from "Zipf's curse", most things occur rarely).

Corpus querying: Recall of search

- By **recall** of search we are referring to:
  - How many of the intended examples that in principle are in the corpus are in fact found by the query?
- Requirements on recall of search
  - for qualitative analysis: Any results found are useful, but danger of partial blindness if example subclasses are not captured by query approximating target phenomenon.
  - for quantitative analysis: Maximizing recall is crucial for reliable quantitative results.

⇒ Where the query characterizing the target phenomenon is expressed in terms of the annotation, quality and consistency of the annotation is important.
  - take into account annotation error rate and inter-annotator agreement levels reported for annotation scheme

A 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 (Oliver Christ, Bruno Schulze)

To search for the relevant examples, a phenomenon has to be characterized 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 to be tested

In perfect tense constructions, AcI verbs are always realized in their substitute infinitival form (IPP). (Suchsland 1994, see discussion in Meurers 2000)

(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 matches

Word forms and part of speech tags: some results

(2) Ich hab's in meiner Schulter krachen **gehört** ...
I have it in my shoulder crack heard
'I heard it crack in my shoulders'

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

(4) 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
already at past events sing heard so that self the
Frage, ob es dies nun kann oder nicht, schon vorher erübrigt hatte.
ques. wh. it this now can or not already before unnec. 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.'

Word occurrences in domains

Exploration of a phenomenon:

What hypotactic chains of modal verbs in what interpretations are possible in German?

Search in the DK

How can one find modal verbs selecting another modal?

1. 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.)

2. 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)

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

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

Topological Fields

Generalization (den Besten & Edmondson 1983):

> Speakers of Middle-Bavarian, South-Bavarian and Franconian use an otherwise inexistent verbal complex order when they "attempt to sound non-dialect like".

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

Topological Fields

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)

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

Direct reference to topological fields

Problem 1: Low precision since topological field information not directly encoded.

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.,

(9) 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.)

2. erroneous corpus annotation

   ⇒ Annotation tools usually geared towards disambiguation at any cost. For linguistic searches it may be preferable to preserve certain ambiguities.

Problem 2: Other topological notions (*Mittelfeld*, extraposition) are impossible to translate into words and POS-tags.

⇒ Annotating corpora with topological information would be highly useful for linguistic corpus searches.

Constituents

Exploration of a phenomenon:

Müller (1999) noticed fronted constituents consisting of a past participle and an agentive *"von [by]"*-PP

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

Research qustion: Can a *by*-phrase generally be part of a fronted passive?

Constituents (cont.)

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`

  results in 35 examples, including a range of agentive, stative, and other passives

Grammatical relations

Generalization by Pafel (1995):

"Arguments of the noun can be extracted, but modifiers cannot:

(11) * Mit rotem Einband habe ich ein Buch gelesen.
with red cover have I a book read

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)."

Grammatical relations (cont.)

Checking the generalization with a search in FR

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

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

1469 matches

(12) 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.'

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

Direct reference to grammatical relations

B) Use a treebank with special query tools, e.g., the German TIGER treebanks & search tools (Brants et al. 2004)
- Use tree description language to specify linear order, dominance, grammatical functions, ...

⇒ Finds examples with richer internal constituent structure, e.g., coordinated NPs

(14) In Cockpit und Kabine wurden neue Gehaltsstrukturen
in cockpit and cabin were new salary structures
mit "marktkonformen" Anfangsgehältern vereinbart.
with 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.'

Caveat: The more elaborate a query, the stronger its dependence on the specifics and quality of the annotation.

A treebank-based look at subjacency

(15) [$_{NP}$ Many books [$_{PP}$ with [stories t_1]] t_2] were sold [that I wanted to read].

- Baltin (1981) & Chomsky (1986, p. 40): Relative clause cannot be related to t_1, since Subjacency excludes crossing of more than one barrier.
- This is a standard assumption also assumed to apply to German (e.g., Grewendorf 1988, p. 281).
- But Müller (1999, 2004) argues that extraposition may cross arbitrarily many NP boundaries:

  (16) Karl hat mir [ein Bild [einer Frau $_{-i}$]]
  Karl has me a picture of.a woman
  gegeben, [die schon lange tot ist]$_i$.
  given who already long dead is

- Can corpora shed light on this and the factors involved?

A Query Based on the TIGER Treebank

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

1. Search for an NP node (#xp),
2. that immediately dominates a relative clause (RC).
3. #xp is immediately dominated by an NP node.
4. #xp is discontinuous,
   that is, the object clause is usually extraposed.

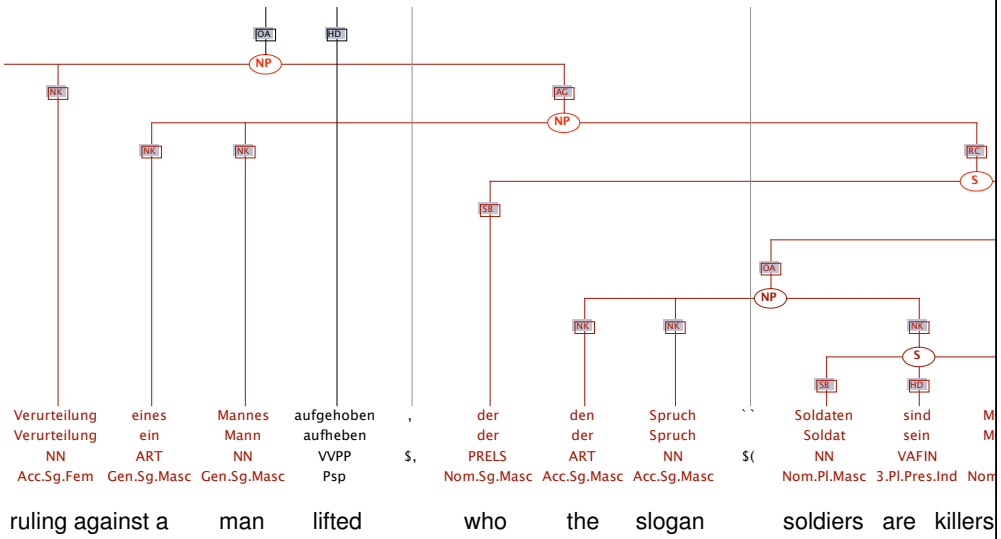
We get 523 hits in the TIGER corpus (40018 sentences), many of which violate the supposed subjacency constraint.

One of the examples from the TIGER Corpus

(17) Bereits vor einem Jahr hatte eine aus drei Verfassungsrichtern
already back one year had one of three judges
bestehende Kammer [die Verurteilung [eines Mannes]]
consisting court the ruling against a man
aufgehoben, der den Spruch "Soldaten sind Mörder" auf seinen
lifted who the slogan soldiers are killers on his
Wagen geklebt hatte.
car glued had

'A year ago, three judges had already repealed [a ruling against [a man]], who had glued the slogan "soldiers are killers" to his car.'

One of the examples from the TIGER Corpus



Zipf's Curse

- For some other phenomena we searched for in treebanks we found: nothing!
- Many examples relevant for linguistics occur infrequently and require huge corpora for successful searches.
- A focused search for relevant examples usually requires annotated corpora.
- Corpora with rich annotation of high quality so far involve manual annotation or correction, limiting their size.
- Shallow or statistical parsing can provide larger annotated corpora; annotation coarseness and quality problematic.
- Syntactic annotation by its nature is based on theoretical assumptions → difficult or impossible to find examples contradicting those assumptions.

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 a phenomenon
- Corpus data are attractive since they
  - exhibit a wide variation of known and unknown parameters
  - are (often) accessible with relevant context
- The use of corpus data does not commit to a particular way of evaluating or interpreting the data.
- Linguistic terminology characterizing the phenomena needs to be reconstructed in terms of annotation.
- Querying of corpora is one of a range of complementary empirical methods, not a cure-all:
  - some phenomena or combination of factors too rare
  - exploration using corpora nicely combinable with psycholinguistic experiments zooming in on specifics

Interested in more detail? Here we go:

- Detmar Meurers (2005): "On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German". *Lingua* 115 (11).
- Detmar Meurers (2007): "Advancing Linguistics Between the Extremes: Some thoughts on Geoffrey Sampson's Grammar without Grammaticality" *Corpus Linguistics and Linguistic Theory* 3(1).
- Detmar Meurers and Stefan Müller (2009): "Corpora and Syntax". Chapter 42 in Lüdeling, A. and Kytö, M.: *Corpus Linguistics*. Handbooks of Linguistics and Communication Science (HSK), Volume 2. Berlin: Mouton de Gruyter. 920–933.

References

Baltin, M. (1981). Strict Bounding. In C. L. Baker & J. J. McCarthy (eds.), *The Logical Problem of Language Acquisition*, Cambridge: Massachusetts, London: England: The MIT Press.

Brants, S., S. Dipper et al. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation* 2(4), 597–620.

Chomsky, N. (1973). Conditions on Transformations. In S. R. Anderson & P. Kiparski (eds.), *A Festschrift for Morris Halle*, New York: Holt, Rinehart & Winston, pp. 232–286.

Chomsky, N. (1986). *Barriers*, vol. 13 of *Linguistic Inquiry Monographs*. Cambridge: Massachusetts, London: England: The MIT Press.

Chomsky, N. (1993). *Lectures on Government and Binding – The Pisa Lectures*. No. 9 in Studies in Generative Grammar. Berlin, New York: Mouton de Gruyter, 7 ed.

den Besten, H. & J. A. Edmondson (1983). The Verbal Complex in Continental West Germanic. In W. Abraham (ed.), *On the Formal Syntax of the Westgermania*, Amsterdam/Philadelphia: Benjamins, pp. 155–216. (Linguistik Aktuell 3).

Fanselow, G. (1991). Minimale Syntax. *Groninger Arbeiten zur Germanistischen Linguistik* 32.

Grewendorf, G. (1988). *Aspekte der deutschen Syntax. Eine Rektions"=Bindungs"=Analyse*. No. 33 in Studien zur deutschen Grammatik. Tübingen: Gunter Narr Verlag.

Haider, H. (1991). *Fakultativ kohärente Infinitivkonstruktionen im Deutschen*. Arbeitspapiere des SFB 340 No. 17, IBM Deutschland GmbH, Heidelberg.

Haider, H. (1996). Downright Down to the Right. In U. Lutz & J. Pafel (eds.), *On Extraction and Extraposition in German*, Amsterdam: Benjamins, no. 11 in Linguistik Aktuell / Linguistics Today, pp. 245–271.

Jacobson, P. (1987). Phrase Structure, Grammatical Relations, and Discontinuous Constituents. In G. J. Huck & A. E. Ojeda (eds.), *Discontinuous Constituency*, New York: Academic Press, vol. 20 of *Syntax and Semantics*, pp. 27–69.

Meurers, W. D. (2000). *Lexical Generalizations in the Syntax of German Non-Finite Constructions*. Arbeitspapiere des SFB 340 No. 145, Eberhard-Karls-Universität, Tübingen. PhD Thesis. http://www.ling.ohio-state.edu/~dm/papers/diss.html. 19.08.2002.

Müller, S. (1999). *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche*. No. 394 in Linguistische Arbeiten. Tübingen: Max Niemeyer Verlag.

Müller, S. (1999). *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche*. No. 394 in Linguistische Arbeiten. Tübingen: Max Niemeyer Verlag. http://www.cl.uni-bremen.de/~stefan/Pub/hpsg.html. 17.03.2011.

Müller, S. (2004). Complex NPs, Subjacency, and Extraposition. *Snippets* 8, 10–11. http://www.cl.uni-bremen.de/~stefan/Pub/subjazenz.html. 17.03.2011.

Pafel, J. (1995). Kinds of Extraction from Noun Phrases. In U. Lutz & J. Pafel (eds.), *On Extraction and Extraposition in German*, Amsterdam/Philadelphia: Benjamins, vol. 2 of *Linguistik aktuell*.

Rohrer, C. (1996). Fakultativ kohärente Infinitkonstruktionen im Deutschen und deren Behandlung in der Lexikalisch Funktionalen Grammatik. In G. Harras &

M. Bierwisch (eds.), *Wenn die Semantik arbeitet. Klaus Baumg"artner zum 65. Geburtstag*, Tübingen: Max Niemeyer Verlag, pp. 89–108.

Ross, J. R. (1967). Constraints on Variables in Syntax. Ph.D. thesis, MIT. Reproduced by the Indiana University Linguistics Club.

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

Suchsland, P. (1994). "Äußere" und "innere" Aspekte von Infiniteinbettungen im Deutschen. In A. Steube & G. Zybatow (eds.), *Zur Satzwertigkeit von Infinitiven und Small clauses*, Tübingen: Max Niemeyer Verlag, no. 315 in Linguistische Arbeiten.

von Stechow, A. & W. Sternefeld (1988). *Bausteine syntaktischen Wissens. Ein Lehrbuch der Generativen Grammatik*. Opladen/Wiesbaden: Westdeutscher Verlag.