EBERHARD KARLS UNIVERSITÄT TÜBINGEN SFB 441

SINBAD
Experimental Data in a Treebank of Suboptimal Structures

Sam Featherston, Ilona Steiner, and Wolfgang Sternefeld

SFB 441, Project A3
University of Tübingen
Germany

Talk Outline

- Introduction (WS)
- A3's generatively tagged treebank: Sinbad
- Searching Sinbad (IS)
- Doing syntax with experimentally obtained judgements (SF)
- Conclusions

A3: Suboptimal syntactic structures

- Generative perspective on marginally grammatical structures
- Introspective judgements as data type
- Treebank of generatively tagged structures
- Experimental approach to gathering judgements
- Data/theory relationship

Conventional Treebanks

- Conventional treebanks are collections of documents.
- A document is a coherent sequence of utterances.
- Examples are newspaper articles, novels, transcriptions of dialogue recordings, historical texts, etc.
- Utterances from written documents are supposed to be grammatically well-formed.

Evidence from Conventional Treebanks

- The treebank is seen as a database of grammatical sentences together with their syntactic analyses.
- The linguist queries the treebank to find instances of a particular linguistic phenomenon.
- Only positive instances can be found. A particular structure may not be found a wide range of reasons.
- This type of linguistic argumentation provides information about what speakers *do* do, not about what any speaker *can* do.

Generative Linguistics

- In the Chomskian framework, linguistic discussion uses a different paradigm: The core question is what speakers can and cannot do.
- Linguists use introspective data: they make up possible and impossible examples which reveal what is and is not possible.
- In this discourse, the data base considered as relevant is extended to include counter-examples.

Generative Linguistics

- Within the Generative paradigm, ungrammatical sentences provide negative evidence. A proposed linguistic analysis must generate the grammatical and exclude ungrammatical sentences.
- But introspective judgements of (un)grammaticality of individual examples may vary considerably. Researchers may disagree with each other.

Suboptimal Structures

- Chomsky originally assumed a dichotomy: Structures can be either grammatical or ungrammatical.
- It is sometimes recognized nowadays this binary distinction is an over-simplification.
- In the literature this is often acknowledged by admitting that judgements are only relative: a sentence marked * is worse than one without *.
- Sentences can be suboptimal in the sense that they are neither perfect nor completely ungrammatical.

Suboptimal Structures

- Although suboptimal sentences and their structures play an important role in the discussion of current topics in linguistics, the situation within theory has still not changed substantially.
- Syntacticians still often idealize data to a binary model of grammaticality.
- Little attempt is made to put the suboptimality of the data on an intersubjective, quantitative basis.
- Little attempt is made to integrate suboptimality into the model of grammar.

Example of a Suboptimal Sentence

- Perfect sentence:

  Damit hat keiner gerechnet.

  With-it has nobody reckoned.

  Nobody expected it.

- Suboptimal sentence:

  Mit gerechnet hat da keiner.

  With reckoned has it nobody.

Aims of the Project

- Provide an accessible database of controversial judgments.
- Provide the user with additional experimental evidence.
- Develop a model of grammar that accounts for suboptimality.
- Develop standards of comparison for intersubjective judgements of suboptimality

Features of the Database

- Search for data and their judgements . . .
  - queries by key-words
  - queries by structure
- Make the treebank accessible to fellow linguists.
  - Open access via the internet.
  - Powerful query mechanisms.

The Sources of our examples

- Linguistics books and journal articles.
- Experimental data.
- Current size: ca. 1100 trees.
- Intended final size: about 3000 trees.
- Examples are chosen for their importance for theoretical questions.

Design Principles of the Annotation

- Problem:
  It may well be that “a sentence has as many structures as there are theories.” (H. Haider)
- Conventional treebanks pretend to avoid the problem by claiming that their annotation is “theory-neutral”.
- This is an illusion; there is no theory-neutral syntax.
- Our annotation is explicitly generative, since it aims to serve generative linguists.

Design Principles of the Annotation

- Compromise between
  - expectations of linguistically trained user;
  - standard assumptions of Generative Grammar;
  - my own (sometimes non-standard) assumptions about the structure of German;
  - simplicity of structure;
  - making it possible to formulate queries
  - better parsability.

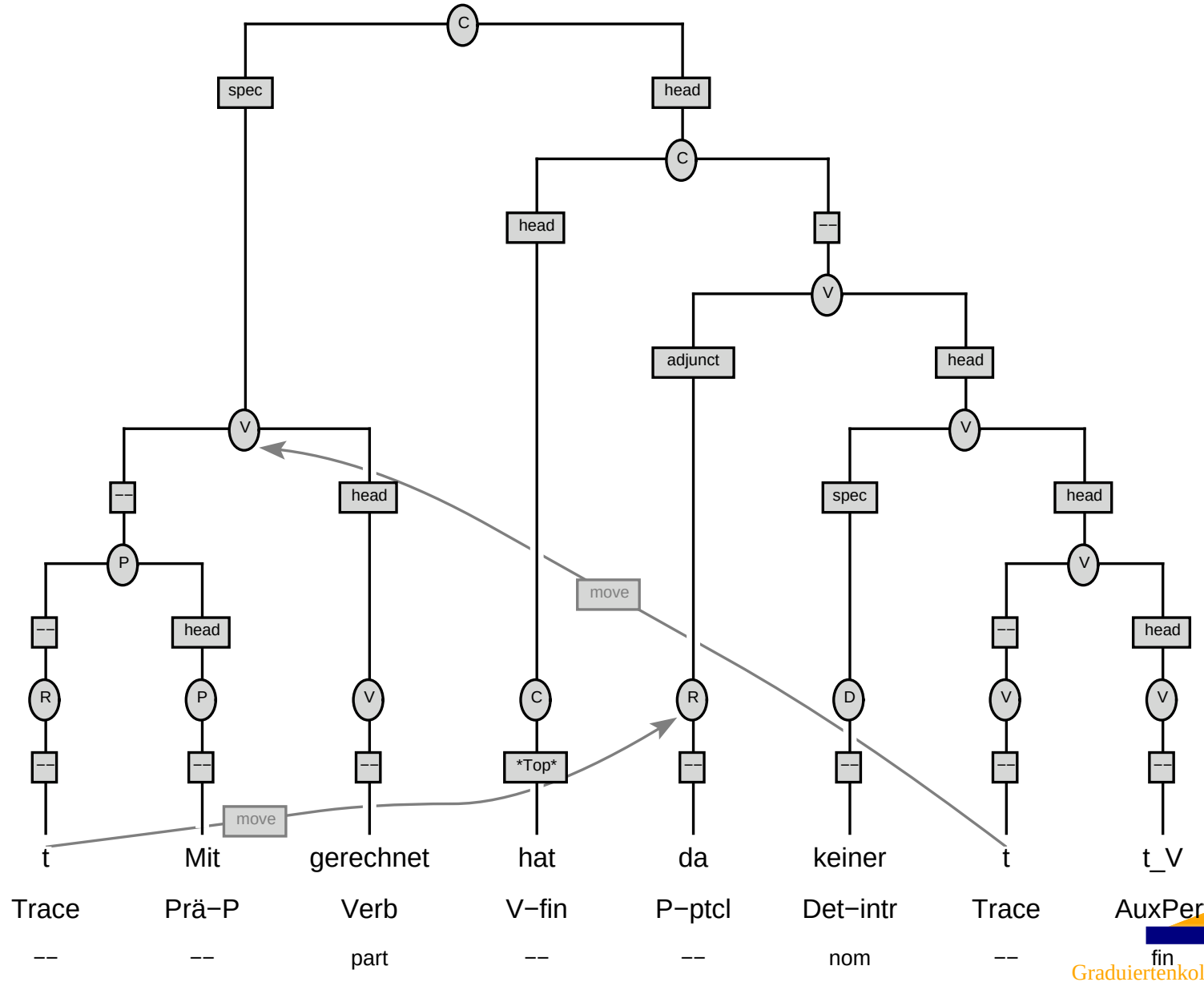
The Annotation

- Trees are binary branching.
- Explicit annotation of movement and binding.
- We allow for traces and empty categories.
  (All trees are annotated by hand.)
- Annotation is sometimes selective: it is not comprehensive but focuses on theoretically relevant features of the structure.
- We can thus use just a very small and user-friendly set of morpho-syntactic categories.

The Annotation Scheme

- Part-of-Speech tags (a relatively small set)
- Morphological information (task-oriented, incomplete)
- Syntactic categories (node labels, only seven different categories)
- Grammatical functions (edge labels: head, adjunct, complement = a minimalist X-bar theory)
- Secondary edges for movement and coreference (basically “move” and “co-ind”)
- Additional contextual features (might facilitate queries, otherwise redundant)

Example Tree



Major Categories

A the category of adjectives and adverbials

C the category of complementizers and the position of the finite verb in main clauses

D the category of determiners, including intransitive determiners like pronouns and proper names

N the category of common nouns

P the category of adpositions, i.e., pre- and postpositions

V the category of verbs

R a the rest: category for anything that does not fit into the other categories

Database of Information on Trees

Information available for each tree beyond the structure:

- Source of the example,
- Judgement of example in source,
- Set of structurally similar trees.

Summary

- First treebank for German with analyses in a GG framework.
- First treebank of suboptimal sentences with their grammaticality ratings.
- Powerful structural search facilities (more powerful than anything else on the market).
- Fully accessible via the internet.

SINBAD in the WWW

Address of SINBAD:
`http://barlach.sfb.uni-tuebingen.de/~a3/`

The next steps:

- A demonstration of the search tool fsq, developed by Stefan Kepser in project A2. (Ilona Steiner)
- An illustration of the experimental work on judgements done in A3. (Sam Featherston)