Grammatical Functions and Parsing the German Negra Treebank

Karin Müller, Detlef Prescher and Khalil Sima'an

Language and Inference Technology Group,

University of Amsterdam

Motivation

Aim: Develop a **broad-coverage** probabilistic parser for German which recognizes constituents and **grammatical functions**.

- Method
  - Extract a symbolic context-free grammar from the Negra treebank.
  - Train probabilistic grammar versions using
    - multi-word tagging,
    - a robust non-lexicalized parsing model (the Tree-gram model, Sima'an 2000, which is closely related to DOP) and
    - subcat frames based on grammatical functions.
  - Evaluate the performance of the grammar on a test corpus.

Properties of German

- Free-er word order than English

  (1) Der Verein sucht noch ideenreiche Erwachsene mit viel Elan.

  The association still searches for creative adults full of verve.

  (2) Ideenreiche Erwachsene sucht noch der Verein mit viel Elan.

  ⋆ *Creative adults still searches for the association full of verve.*

- Richer morphology
  - Articles express case, number, gender, e.g.
    "der Verein" is nominative ⟶ subject,
    "den/dem Verein" is accusative/dative ⟶object
  - Adjectives are inflected: "den ideenreichen Erwachsenen"
  - Compounding, e.g., "Vereinsbildungsvoraussetzung"
    (*condition at which an association can be formed*)

Properties of German

- Free-er word order than English

  (1) Der Verein sucht noch ideenreiche Erwachsene mit viel Elan.

  The association still searches for creative adults full of verve.

  (2) Ideenreiche Erwachsene sucht noch der Verein mit viel Elan.

  ⋆ *Creative adults still searches for the association full of verve.*

- Richer morphology
  - Articles express case, number, gender, e.g.
    "der Verein" is nominative ⟶ subject,
    "den/dem Verein" is accusative/dative ⟶object
  - Adjectives are inflected: "den ideenreichen Erwachsenen"
  - Compounding, e.g., "Vereinsbildungsvoraussetzung"
    (*condition at which an association can be formed*)

Properties of German

- Free-er word order than English

  (1) Der Verein sucht noch ideenreiche Erwachsene mit viel Elan.

  The association still searches for creative adults full of verve.

  (2) Ideenreiche Erwachsene sucht noch der Verein mit viel Elan.

  ⋆ *Creative adults still searches for the association full of verve.*

- Richer morphology
  - Articles express case, number, gender, e.g.
    "der Verein" is nominative ⟶ subject,
    "den/dem Verein" is accusative/dative ⟶object
  - Adjectives are inflected: "den ideenreichen Erwachsenen"
  - Compounding, e.g., "Vereinsbildungsvoraussetzung"
    (*condition at which an association can be formed*)

Properties of German

- Free-er word order than English

  (1) Der Verein sucht noch ideenreiche Erwachsene mit viel Elan.

  The association still searches for creative adults full of verve.

  (2) Ideenreiche Erwachsene sucht noch der Verein mit viel Elan.

  ⋆ *Creative adults still searches for the association full of verve.*

- Richer morphology
  - Articles express case, number, gender, e.g.
    "der Verein" is nominative ⟶ subject,
    "den/dem Verein" is accusative/dative ⟶object
  - Adjectives are inflected: "den ideenreichen Erwachsenen"
  - Compounding, e.g., "Vereinsbildungsvoraussetzung"
    (*condition at which an association can be formed*)

Properties of German

- Free-er word order than English

  (1) Der Verein sucht noch ideenreiche Erwachsene mit viel Elan.

  The association still searches for creative adults full of verve.

  (2) Ideenreiche Erwachsene sucht noch der Verein mit viel Elan.

  ⋆ *Creative adults still searches for the association full of verve.*

- Richer morphology
  - Articles express case, number, gender, e.g.
    "der Verein" is nominative ⟶ subject,
    "den/dem Verein" is accusative/dative ⟶object
  - Adjectives are inflected: "den ideenreichen Erwachsenen"
  - Compounding, e.g., "Vereinsbildungsvoraussetzung"
    (*condition at which an association can be formed*)

Properties of German

- Free-er word order than English

  (1) Der Verein sucht noch ideenreiche Erwachsene mit viel Elan.

  The association still searches for creative adults full of verve.

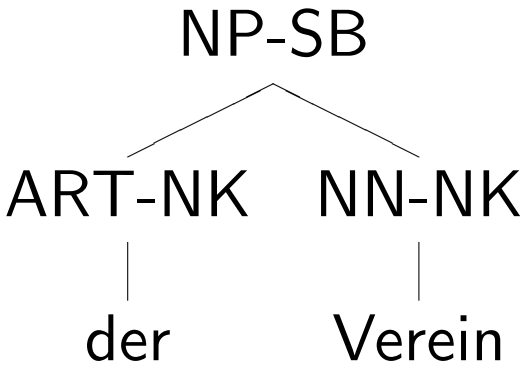
  (2) Ideenreiche Erwachsene sucht noch der Verein mit viel Elan.

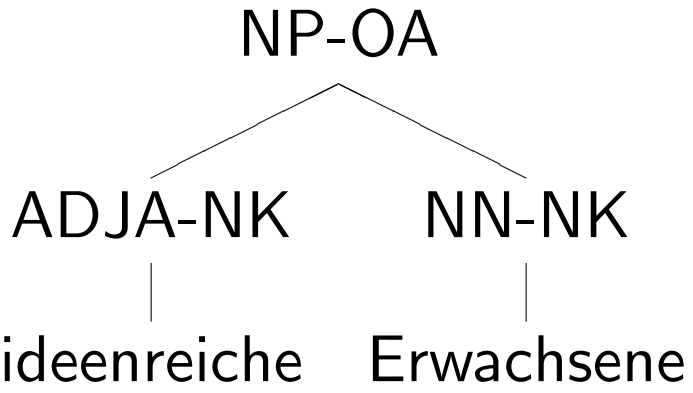
  ⋆ *Creative adults still searches for the association full of verve.*

- Richer morphology
  - Articles express case, number, gender, e.g.
    "der Verein" is nominative ⟶ subject,
    "den/dem Verein" is accusative/dative ⟶object
  - Adjectives are inflected: "den ideenreichen Erwachsenen"
  - Compounding, e.g., "Vereinsbildungsvoraussetzung"
    (*condition at which an association can be formed*)

Properties of the Negra Treebank

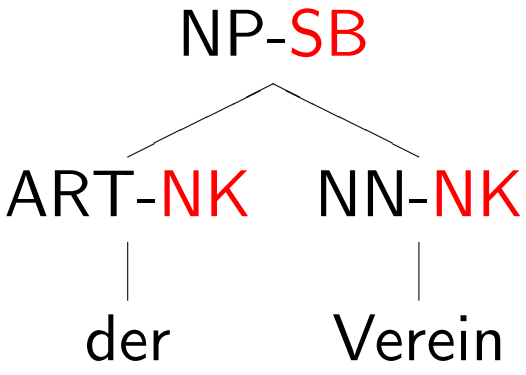
- Resource: Negra treebank, a newspaper corpus (20,000), manually annotated with syntactic structure.
- We use the following format of Negra: Nodes consist of complex labels namely constituent categories and a rich set of grammatical functions expressing syntactic relations among words

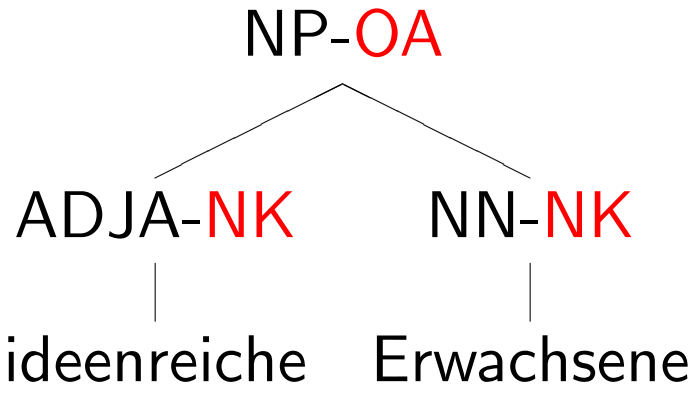




Properties of the Negra Treebank

- Resource: Negra treebank, a newspaper corpus (20,000), manually annotated with syntactic structure.
- We use the following format of Negra: Nodes consist of complex labels namely constituent categories and a rich set of grammatical functions expressing syntactic relations among words





Properties of the Negra Treebank (cont'd)

- Relatively flat structure where recursion is avoided.
  - special category for coordination (CS = CNP and CNP)
  - There is no PP $\longrightarrow$ P NP rule,
    PP $\longrightarrow$ $[\boldsymbol{mit}]_{\boldsymbol{P}}$ $[\boldsymbol{viel}]_{\boldsymbol{PIDAT}}$ $[\boldsymbol{Elan}]_{\boldsymbol{NN}}$
  - VPs do only occur if the verb subcategorizes a subordinate clause, or verb consists of auxiliary and full verb.

General Idea of Probabilistic Parsing

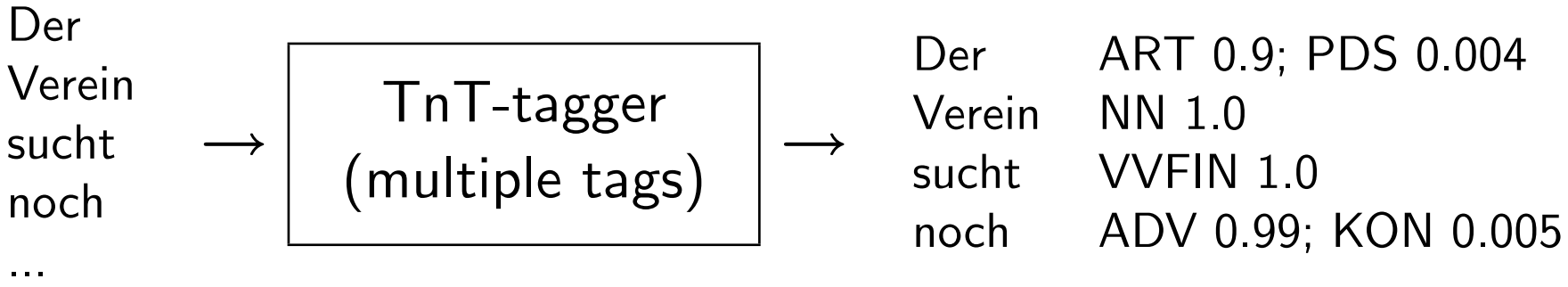
- Symbolic Parsing:
  - Parse a given sentence and assign all possible structures (syntactic trees).
- Probabilistic Parsing:
  - Use a symbolic parsing component to get all possible syntactic structures of a sentence
  - Disambiguate the different analyses by using rule probabilities, i.e. choose the most probable analysis.

Previous Research on Parsing German

- Unlexicalized models (Fissaha et al. 2003)
  - parsing with categories only (P:71.7%, Cov:**96.6%**)
  - parsing with cat. and gram. functions (P:67.5%, Cov:**97.8%**)
- Grammar transformation techniques
  - Full-parent encoding (P:54.9%, Cov:70.2%)
  - Partial-parent encoding of constituents (P:53%, Cov:80.8%)
  - Partial-parent encoding of grammatical functions (P:51.8%, Cov:91.1%)
- Lexicalization (Dubey&Keller 2003) - no increase in performance Sister-head dependencies (P:**70.9%** Cov:95.9%)

New Component: Tagging with multiple tags

- Re-train TnT-tagger (Brants 2000) on 18000 sentences of Negra. Accuracy: 96.8%
- Wordgraph makes it possible that tagging accuracy can possibly increase to 98.92% (if at least 2 tags are allowed)
- Compute $p(\texttt{pos} \longrightarrow \texttt{word})$ using TnT's output $p(\texttt{pos}|\texttt{word})$ and Bayes formula



New Component: Tree-gram model

Augment context-free rules (read-off from Negra) with

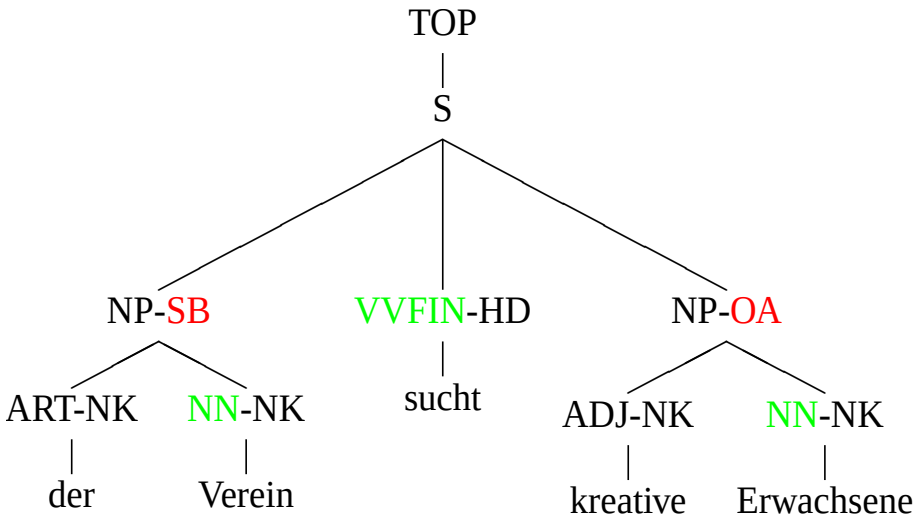
- Parent information
- Part-of-speech tags as lexical heads
- First order Markov information
- Subcat information in terms of grammatical functions

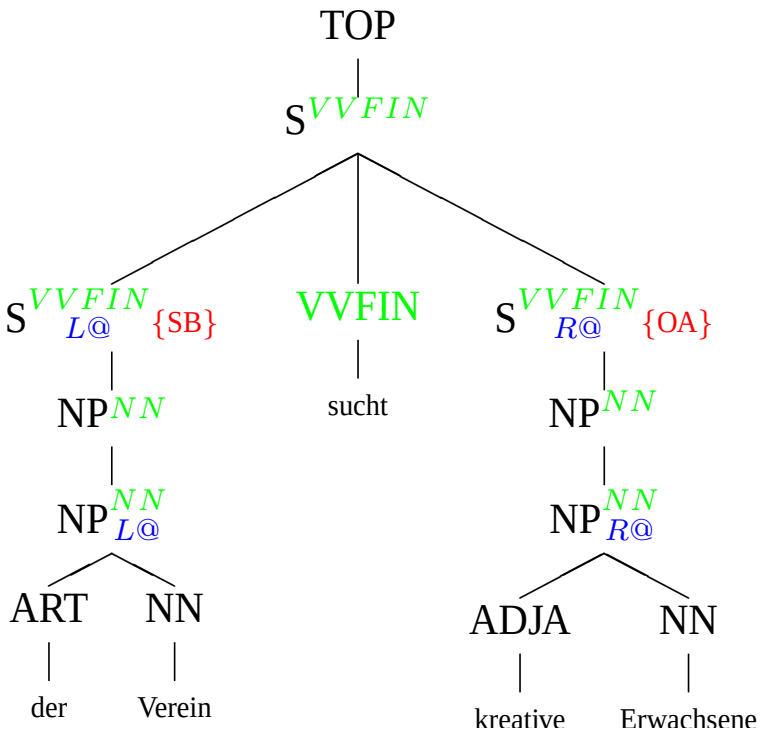
Model realization via Tree Transformation

Head of a rule: green

New nodes according to linearization (Markovian process): blue

Grammatical function: red





First Results

- Unlexicalized version of Tree-Gram parsing shows that grammatical functions are useful for parsing
- Results:

precision	72.85%
recall	71.0%
coverage	100%

- Results outperforms results reported by Dubey&Keller (2003) who used previous sister-head dependencies (prev sister cat., head word, head tag).

precision	70.9%
recall	71.3%
coverage	95.9%

Future Work

- Switch to a bigger corpus, TIGER (released 2003) comprising 40000 sentences
- Use of lexicalized models
- Use deeper tree-structures