



# Searching in Annotated Corpora

Sandra Kübler

`kuebler@sfs.uni-tuebingen.de`



General prerequisites for linguistic searching:

- annotation of relevant linguistic information
- translation from linguistic question into terms of annotation
- search tool (+ query language)



General prerequisites for linguistic searching:

- annotation of relevant linguistic information
- translation from linguistic question into terms of annotation
- search tool (+ query language)

Prerequisites for users:

- linguist needs to be familiar with query language
- linguist needs to be familiar with annotation style
- question must be searchable



- “Show me all occurrences of the word `hilarious` in the corpus.”



# Typical Questions



- “Show me all occurrences of the word `hilarious` in the corpus.”
- “Can the word `book` be used as a verb?”



# Typical Questions



- “Show me all occurrences of the word `hilarious` in the corpus.”
- “Can the word `book` be used as a verb?”
- “How often does `for` occur as a preposition, how often as a subordinating conjunction?”



- “Show me all occurrences of the word `hilarious` in the corpus.”
- “Can the word `book` be used as a verb?”
- “How often does `for` occur as a preposition, how often as a subordinating conjunction?”
- “What types of phrases can occur between a German NP and its relative clause?”



- “Show me all occurrences of the word `hilarious` in the corpus.”
- “Can the word `book` be used as a verb?”
- “How often does `for` occur as a preposition, how often as a subordinating conjunction?”
- “What types of phrases can occur between a German NP and its relative clause?”
- “Are there subjectless sentences in German?”



- “Show me all occurrences of the word `hilarious` in the corpus.”
- “Can the word `book` be used as a verb?”
- “How often does `for` occur as a preposition, how often as a subordinating conjunction?”
- “What types of phrases can occur between a German NP and its relative clause?”
- “Are there subjectless sentences in German?”
- “Show me all direct object ellipses.”



- “Show me all occurrences of the word `hilarious` in the corpus.”
- “Can the word `book` be used as a verb?”
- “How often does `for` occur as a preposition, how often as a subordinating conjunction?”
- “What types of phrases can occur between a German NP and its relative clause?”
- “Are there subjectless sentences in German?”
- “Show me all direct object ellipses.”
- “Show me all elliptical sentences.”



#np:[cat="NP"] & #np > [pos="ADJA"] & #np >  
[pos="NN"] & #mo:[cat="VF"] & #mo >OA #np



```
#np:[cat="NP"] & #np > [pos="ADJA"] & #np >
[pos="NN"] & #mo:[cat="VF"] & #mo >OA #np
```

Find me all trees which have a noun phrase containing an adjective and a noun, and this noun phrase is the direct object and in initial position.



- pure text
- positional annotation, e.g. POS tags, morphology, lexical information
- graph annotation, e.g. syntax



- word form
- POS tag
- preceding word(s)
- following word(s)



- word form
- POS tag
- preceding word(s)
- following word(s)

only one dimension:

linear text – only look at sequence of words and their characteristics





## Relations:

- sequence of words and their characteristics
- linear precedence not only for words but also for nodes in tree
- dominance between nodes
- immediate dominance, immediate precedence



Relations:

- sequence of words and their characteristics
- linear precedence not only for words but also for nodes in tree
- dominance between nodes
- immediate dominance, immediate precedence

more complex search!!!



Question: “Can one front PPs modifying a noun phrase in German?”



Question: “Can one front PPs modifying a noun phrase in German?”

Find all sentences that have:

- a PP in initial position
- an NP after the finite verb
- a modifier relation between the PP and the NP



Question: “Can one front PPs modifying a noun phrase in German?”

Find all sentences that have:

- a PP in initial position
- an NP after the finite verb
- a modifier relation between the PP and the NP

In terms of TüBa-D/Z:

- there is an initial field (VF) which dominates a PX
- there is an NX in the middle field (MF)
- the NX has the function  $X$  and the PX has the function  $X$ -MOD



“What can occur between a noun phrase and its relative clause in German?”



“What can occur between a noun phrase and its relative clause in German?”

Find all sentences that have:

- an NP after the finite verb
- a relative clause at the end of the sentence
- both in the same sentence



“What can occur between a noun phrase and its relative clause in German?”

Find all sentences that have:

- an NP after the finite verb
- a relative clause at the end of the sentence
- both in the same sentence

In terms of TüBa-D/Z:

- there is a middle field (MF) which dominates an NX
- there is an R-SIMPX in the final field (NF)
- there is a SIMPX node which dominates both



- there is a middle field (MF) which dominates an NX
- there is an R-SIMPX in the final field (NF)
- there is a SIMPX node which dominates both

This also retrieves trees which have adjacent NP + rel. clause.



- there is a middle field (MF) which dominates an NX
- there is an R-SIMPX in the final field (NF)
- there is a SIMPX node which dominates both

This also retrieves trees which have adjacent NP + rel. clause.

2 possibilities:

- there is another constituent in the MF
- there is a VC between MF and NF



The relation between the NX with function  $X$  and the PX with function  $X$ -MOD is difficult to specify.



The relation between the NX with function  $X$  and the PX with function  $X$ -MOD is difficult to specify.

possible solutions:

- leave this out and possibly get too many trees
- search for all different combinations of functions, e.g. OA + OA-MOD; ON + ON-MOD; ...  
a lot of work!



Question: “Give me all trees with a subject ellipsis.”



Question: “Give me all trees with a subject ellipsis.”

Find all sentences that have:

- no subject in initial field (VF)
- no subject in middle field (MF)



Question: “Give me all trees with a subject ellipsis.”

Find all sentences that have:

- no subject in initial field (VF)
- no subject in middle field (MF)

Unfortunately, not possible!!!



Question: “Give me all trees with a subject ellipsis.”

Find all sentences that have:

- no subject in initial field (VF)
- no subject in middle field (MF)

Unfortunately, not possible!!!

Can only search for existing constituents which do not have a specific label: “Give me all sentences that have a constituent in MF which is not the subject.”

Holds for all sentences that have at least one constituent in MF which is not the subject!



- **IMS Corpus Workbench**  
can search for positional annotations  
URL: <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>
- **TIGERSearch**  
can search for graph annotations; user-friendly tree drawing interface  
URL: <http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/>
- **Finite Structure Query Tool**  
powerful search tool, based on first-order logic  
URL:  
<http://tcl.sfs.uni-tuebingen.de/fsq/>



- ICECUP

developed for searching the ICE-GB corpus;  
also tree drawing interface

URL:

<http://www.ucl.ac.uk/english-usage/ice-gb/icecup.htm>

- CLARK tool

XML tool, which includes XPATH queries: very  
powerful

URL:

<http://www.bultreebank.org/clark>