Corpus Search Tools

Sandra Kübler

Dept. of Linguistics, Indiana

Types of Information

1. words alone
2. words + lemmata
3. words + morphology
4. words + POS tags
5. words + syntactic structures
6. words + meanings

Types of Information

1. words alone
2. words + lemmata
3. words + morphology
4. words + POS tags
5. words + syntactic structures
6. words + meanings

requirements for text:

- for 1.: raw text
- for 2.-6.: annotated text

Types of Information

1. words alone
2. words + lemmata
3. words + morphology
4. words + POS tags
5. words + syntactic structures
6. words + meanings

requirements for text:

- for 1.: raw text
- for 2.-6.: annotated text

requirements for search tools:

- for 1.: concordancer
- for 2.-4.: position based query tool
- for 5.: syntactic query tool

Search Tools

- Mark Davies' online concordancer for BNC and Corpus for American English (coming soon)
  http://corpus.byu.edu/
- Scott Piao's MLCT (Multi-Lingual Toolkit)
  https://sites.google.com/site/scottpiaosite/software/mlct
  allows user to concordance text and web pages
- Steven Bird's Treebank Search
  http://nltk.ldc.upenn.edu:8080/ts/
- TIGERSearch
  http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/
  requires the import of corpora
- Stephan Kepser's Finite Structure Query Tool
  http://www.tcl-sfs.uni-tuebingen.de/fsq/
  extremely powerful, hard to use

Where corpus examples cannot help

- no negative examples
  exception: SINBAD (collection of interesting example sentences in German)
- no proof of non-existence
  corpora are limited in size, genre, etc.
- for low-frequency events, no guarantee of grammaticality

TIGERSearch

graphical search interface:

- allows to search in two-dimensional tree structures
- 2 forms of query design:
  - text mode (logical query language):

```
#n1:[cat="SIMPX"] >*
[word=("werden"|"wird")]
& #n1 >* [pos="ADJD"]
```

  - graphical mode (construct partial trees)
- searchable relations: direct dominance, dominance (transitive closure), direct linear precedence, linear precedence (transitive closure)
- limited negations

Prerequisites

- find annotated corpus – annotated with the phenomenon of interest
- in a format that TIGERSearch knows
- import corpus via TIGERRegistry
- read corpus documentation!
- start TIGERSearch
  on mac in terminal:

```
cd /Applications/TIGERSearch/lib
./runTS&
```

- load corpus: Corpus > open

Searching

- search for the words ”es gibt” (there is)
- search for sequence of POS tags: VVFIN PTKVZ
- search for sequence of POS tags: VVFIN PTKVZ but not necessarily adjacent
- search for NX with a PX modifier
- search for ADVX that is a modifier of the direct object (OA)
- search for sentences that have the direct object before the subject (ON)
- search for an initial field that has an NX or a PX as daughter

Searching (2)

- search for sentences with subject in first position
- search for sentences with subject in other positions
- search for coordinations of unlikes involving a noun phrase
- search for coordinations of unlikes where the noun phrase is NOT the first conjunct
- search for fronted verb complex
- search for discontinuous (interrupted) NX
- search for a MF directly dominated by a SIMPX or with an FKOORD in between

Types of Negation

- negated node values (not NX)
- node category is not specific label (cat != NX)
  BUT: the means look for an existing node that is not an NX
- negated edge (VF does not dominate NX)
  BUT: this means, there is a node VF, and a node NX, and they are not mother-daughter
- negated precedence (find occurrences of 'ich' that do not precede 'bin')

Limitations of TIGERSearch

- we can only search for phenomena that are present in the annotation
  ex.: the Penn tagset does not distinguish between prepositions and subordinating conjunctions

Limitations of TIGERSearch

- we can only search for phenomena that are present in the annotation
  ex.: the Penn tagset does not distinguish between prepositions and subordinating conjunctions
- we cannot search for phenomena that involve elided or deleted words, phrases, etc.

Limitations of TIGERSearch

- we can only search for phenomena that are present in the annotation
  ex.: the Penn tagset does not distinguish between prepositions and subordinating conjunctions
- we cannot search for phenomena that involve elided or deleted words, phrases, etc.
  - we cannot search for subjectless sentences, e.g. `Ihm ist kalt.` (To him is cold.)
    approximately: find all trees which do not have an NP node that has ”subject” as function label

Limitations of TIGERSearch

- we can only search for phenomena that are present in the annotation
  ex.: the Penn tagset does not distinguish between prepositions and subordinating conjunctions
- we cannot search for phenomena that involve elided or deleted words, phrases, etc.
  - we cannot search for subjectless sentences, e.g. `Ihm ist kalt.` (To him is cold.)
    approximately: find all trees which do not have an NP node that has ”subject” as function label
  - we cannot search for coordinated sentences with a subject gap in the second conjunct

Limitations of TIGERSearch

- reason: variables in TIGERSearch are existentially quantified
  i.e. they allow for searches ”there exists a node X that ...”

Limitations of TIGERSearch

- reason: variables in TIGERSearch are existentially quantified
  i.e. they allow for searches ”there exists a node X that ...”
- but: negation can only be attached to existing nodes:
  find all trees that have a node before the main verb **which is not an NP with function label subject**

Limitations of TIGERSearch

- reason: variables in TIGERSearch are existentially quantified
  i.e. they allow for searches ”there exists a node X that ...”
- but: negation can only be attached to existing nodes:
  find all trees that have a node before the main verb **which is not an NP with function label subject**
- why this restriction?
  search complexity

Steven Bird's Treebank Search

- online search tool
- restricted to corpora that are provided
- powerful search language
- extremely fast

Query Language

```
<expr> ::= <term> [<term>]*

<term> ::= <axis-operator><label> [<filter-expr>]

<filter-expr> ::= "[" <filter-element> [(AND|OR)
<filter-element>]* "]"

<filter-element> ::= [NOT] (<term> | <expr>)

<label> ::= PennTreebankLabel | word | punctuation
```

Query Language

Axis Operators

\\	Ancestor
//	Descendant
\	Parent
/	Child
–– >	Following
– >	Immediate Following
==>	Following Sibling
=>	Immediate Following Sibling
< ––	Preceding
< –	Immediate Preceding
<==	Preceding Sibling
<=	Immediate Preceding Sibling

Query examples

- search for the words ”as soon as”
- search for the POS sequence PDT DT
- search for the POS sequence PDT DT, not necessarily adjacent
- search for the word ”can” not used as a noun
- search for sentences that have a VP as the root
- search for a VP that has a PP modifier
- search for an NP that does not have an NN inside
- search for temporal NPs

Searching

- search for sentences with a fronted PP
- search for a coordinated VP
- search for coordinations of unlikes involving a noun phrase
- search for coordinations of unlikes not involving an NP but a PP
- search for a UCP that dominates an NP and a PP so that the NP precedes the PP
- search for an NP that is dominated either directly by a VP or with a UCP in between

Searching

- search for sentences with a fronted PP
  /S/PP==>NP-SBJ
- search for a coordinated VP
- search for coordinations of unlikes involving a noun phrase
- search for coordinations of unlikes not involving an NP but a PP
- search for a UCP that dominates an NP and a PP so that the NP precedes the PP
- search for an NP that is dominated either directly by a VP or with a UCP in between

Searching

- search for sentences with a fronted PP
  /S/PP==>NP-SBJ
- search for a coordinated VP
  //VP[/CC OR /$,]
- search for coordinations of unlikes involving a noun phrase
- search for coordinations of unlikes not involving an NP but a PP
- search for a UCP that dominates an NP and a PP so that the NP precedes the PP
- search for an NP that is dominated either directly by a VP or with a UCP in between

Searching

- search for sentences with a fronted PP
  /S/PP==>NP-SBJ
- search for a coordinated VP
  //VP[/CC OR /$,]
- search for coordinations of unlikes involving a noun phrase
  //UCP/NP
- search for coordinations of unlikes not involving an NP but a PP

- search for a UCP that dominates an NP and a PP so that the NP precedes the PP

- search for an NP that is dominated either directly by a VP or with a UCP in between

Searching

- search for sentences with a fronted PP
  /S/PP==>NP-SBJ
- search for a coordinated VP
  //VP[/CC OR /$,]
- search for coordinations of unlikes involving a noun phrase
  //UCP/NP
- search for coordinations of unlikes not involving an NP but a PP
  //UCP[/PP AND NOT/NP]
- search for a UCP that dominates an NP and a PP so that the NP precedes the PP

- search for an NP that is dominated either directly by a VP or with a UCP in between

Searching

- search for sentences with a fronted PP
  /S/PP==>NP-SBJ
- search for a coordinated VP
  //VP[/CC OR /$,]
- search for coordinations of unlikes involving a noun phrase
  //UCP/NP
- search for coordinations of unlikes not involving an NP but a PP
  //UCP[/PP AND NOT/NP]
- search for a UCP that dominates an NP and a PP so that the NP precedes the PP
  //UCP/NP==>PP
- search for an NP that is dominated either directly by a VP or with a UCP in between

Searching

- search for sentences with a fronted PP
  /S/PP==>NP-SBJ
- search for a coordinated VP
  //VP[/CC OR /$,]
- search for coordinations of unlikes involving a noun phrase
  //UCP/NP
- search for coordinations of unlikes not involving an NP but a PP
  //UCP[/PP AND NOT/NP]
- search for a UCP that dominates an NP and a PP so that the NP precedes the PP
  //UCP/NP==>PP
- search for an NP that is dominated either directly by a VP or with a UCP in between
  //NP[\VP OR \UCP\VP]

Differences between TIGERSearch and TS

- You can load new corpora into TIGERSearch
  but not into TS
- TIGERSearch has a relative loose definition of tree
  TS only works on real trees (no insertions, no crossing branches)
- TS can look for phrases that do NOT have a certain daughter
  TIGERSearch can only look for phrases that have a node that is not a certain phrase
  e.g. search for a VP without a PP
- TIGERSearch can have “underspecified” nodes
  TS cannot
  e.g. search for a UCP in which the NP daughter is followed by something (not an NP)

Differences betw. TIGERSearch and TS (2)

- TS makes a difference between precedence and precedence among siblings
  TIGERSearch does only in the textual search
- TIGERSearch has variables, TS does not
  $\Rightarrow$ TS cannot formulate two restrictions between two nodes
- TIGERSearch can refer to the first/last terminal daughter of a node
  TS cannot
- TIGERSearch can define the arity of a node
  TS cannot