Achim Stein: Diachronic Corpora
Aston Corpus Summer School 2011

Achim Stein
achim.stein@ling.uni-stuttgart.de

Institut für Linguistik/Romanistik
Universität Stuttgart

2nd of August, 2011

Installation of query tools

- CorpusSearch
- TIGERSearch

Syntax and text corpora

- Syntax models
- Formats of syntactic annotation

Queries

Install CorpusSearch

- For the latest version of CorpusSearch and the documentation see: http://corpussearch.sourceforge.net/
- For this course, we provide a CorpusSearch with some scripts which make things easier. On your classroom machines, please copy the folder X:\LSS\cs to your desktop, then open it.

Install CorpusSearch at home

At home, download the folder as a zip file (for Windows *or* Mac) from my homepage: http://www.uni-stuttgart.de/lingrom/stein/ (search for "Aston" or go to →Ressourcen...Talks)

Run CorpusSearch

- Click on `start-command-window.bat`.
  This is a shortcut for opening the command line window ("terminal"). You can also launch it in *Programmes-Accessories*.
- In the terminal, run your first search by typing:

```
cswin query.txt mandeville-sample.psd
```

The original CorpusSearch command line

If you don't use the cswin.bat file, type the whole command

1. your query file must have the suffix `.q`, e.g. `query.q`
2. you must type the following line:

```
java -classpath CS_2.003.jar csearch/CorpusSearch query.q
*.psd
```

3. your output file will have the suffix `.out`, e.g. `query.out`

File handling and editing

- The folder `cs` already contains a simple query file: `query.txt`. Click on it to edit it with the default text editor.
- If you run `cswin...`, the output file will be `cs-out.txt`.
  Careful: each query will **overwrite** the previous output file! Rename it if you want to keep it.
  - Useful: typing the first letter(s), then TAB, will expand to the matching file name, and the arrow keys allow you to return to previous commands of the session (up), which you can than edit like a line of text.

Improving the environment...

- Free alternatives to the standard text editors are *Crimson Editor* (Windows) or *Textwrangler* (Mac).
- If you want a Unix-like terminal for Windows, install *Cygwin*. It has many commands useful for text corpus manipulation.

TIGERSearch

- On your classroom machines
  - Launch TIGERSearch in the folder `X:\LSS\ts\bin` by clicking on the file `TIGERSearch.exe` (*not* on the icon file with the nice tiger)
  - In the tree of corpora in the left part of the TIGERSearch window, open `DemoCorpora-English-PPCME2Sampler` with a double click.
  - Click on the `Explore corpus` icon, lean back, and browse through the sentence structures using the `Next/Previous` buttons.
- Download TIGERSearch (University of Stuttgart)

Install TIGERSearch at home

Download the installation package (for Windows, Mac, Linux) from: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/oldindex.shtml

Syntactic relations

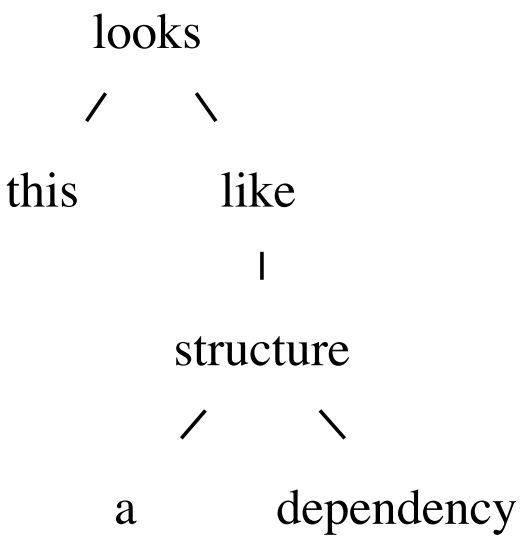
Syntactic relations between words can be expressed in two ways:

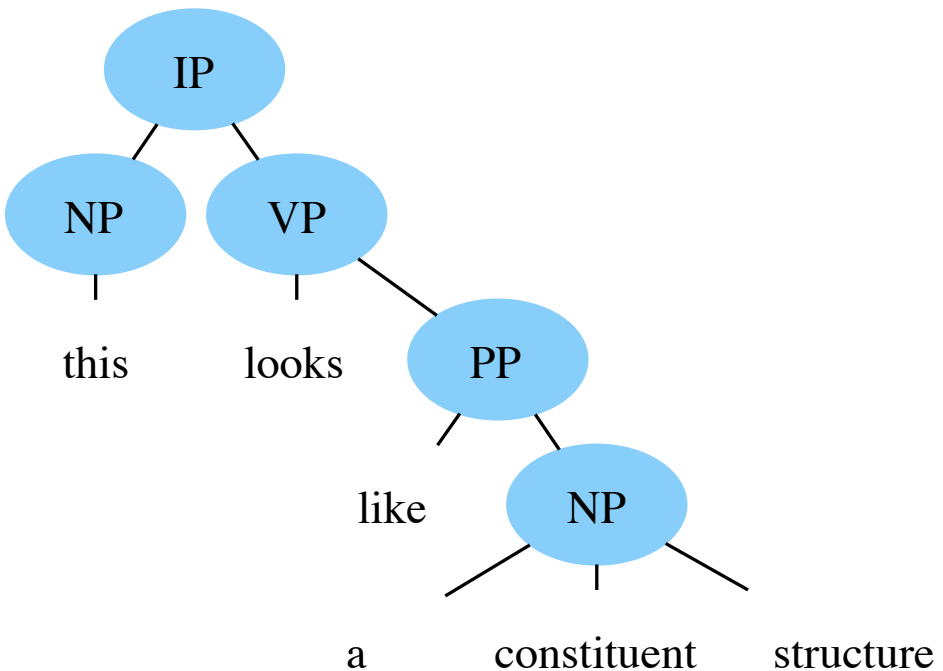
- Dependency
  - On which word depends a given word?
  - Tree with lines between words.
  - Grammatical functions can be attached as arc labels.
  - see Tesnière (1965)
- Constituency
  - Which words belong together (form a group)?
  - Tree with lines between constituents, words are terminal nodes ("leaves").
  - Grammatical functions are configurations in the structure.
  - see Bloomfield (1933)

Syntactic relations as tree graphs

Terminology

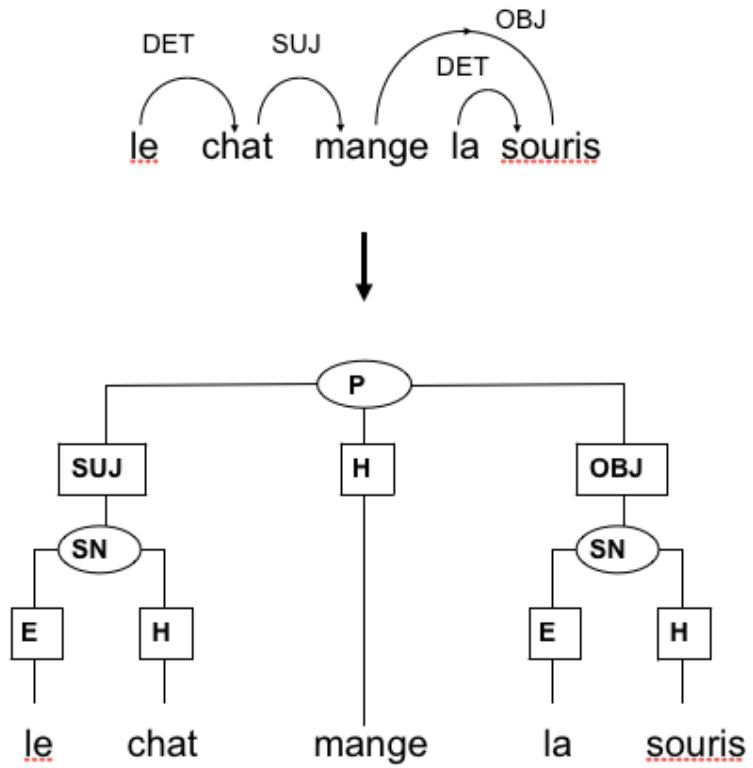
A **tree** (graph) is composed of **nodes** (terminal, non-terminal) and **arcs** (lines, labelled).





Translating syntactic graphs

- Dependency graphs can be translated into constituency graphs (and vice versa)
- In the example (Bourigault et al., 2005):
  - relations (subject etc.) are nodes
  - types of dependencies are arc labels



Syntactic annotation formats

- Tools for idiosyncratic formats (non XML, no standard)
  - CorpusSearch (University of Pennsylvania, UPENN)
  - The PENN format is widely used for english corpora: YCOE, PPCME, EME etc.
- Internal format: bracketed structures

```
(, .)
(CONJP (CONJ and)
    (IP-SUB (NP-SBJ *con*)
      (BEP ys)
      (NP-OB1 (NP (D +te) (ADJS nobleste) (N +ting))
        (, /)
        (CONJP (CONJ an)
          (NP (D +te) (ADJS he+geste) (N ssep+te)))
        (CP-REL (WNP-5 0)
          (C +tet)
          (IP-SUB (NP-SBJ *T*-5)
            (MD may)
            (BE by)))))))))))))
                    (E_S .)) (ID CMAYENBI,92.1797))
```

Syntactic annotation formats

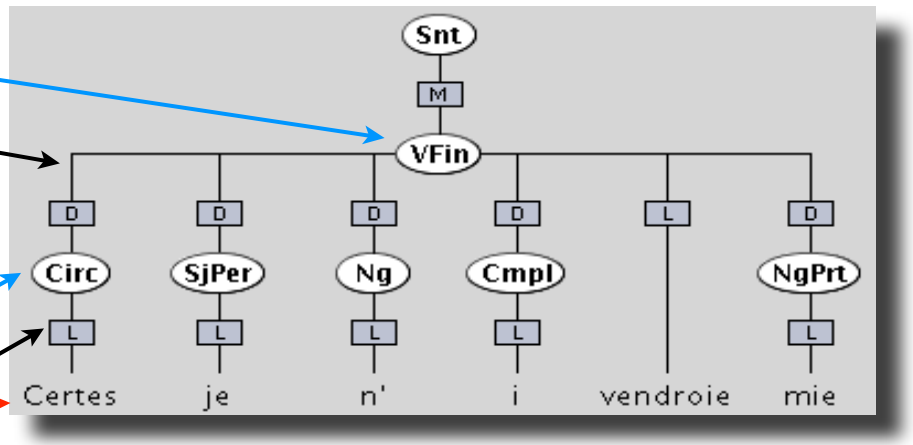
- Tools for XML-formatted corpora
  - TigerSearch / Tiger XML (IMS, University of Stuttgart): http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/oldindex.shtml
  - ANNIS / PAULA XML (Universität Potsdam) http://www.sfb632.uni-potsdam.de/~d1/paula/doc/ exchange format for linguistic annotations
- Clear tendency towards XML formats. XML-based software has import filters für other formats:
  - PAULA has filters for TIGERSearch (and others)
  - TIGERSearch has filters for PENN corpora (and others)

Syntactic Structures in TIGER-XML / TIGERSearch

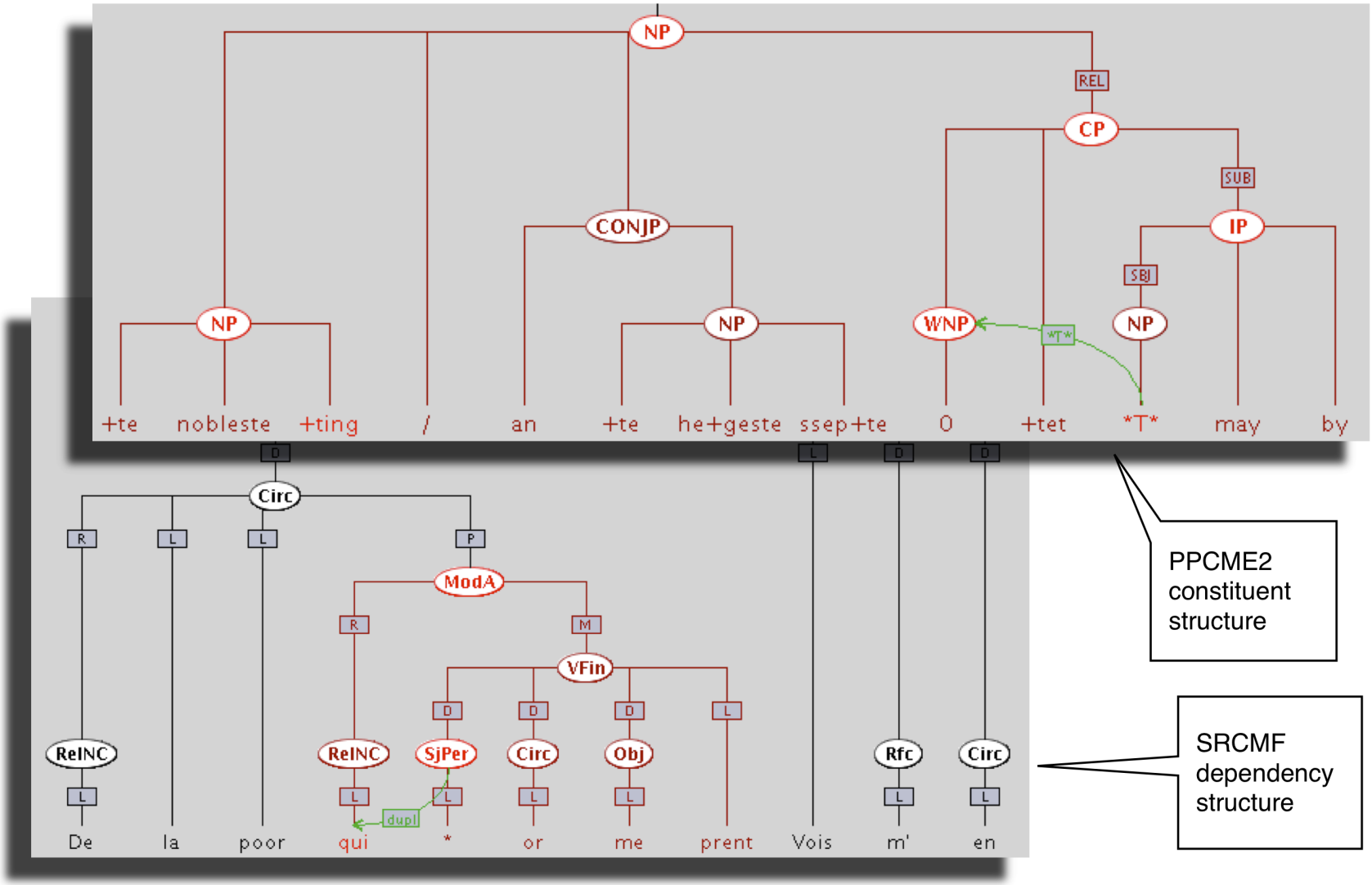
terminale Knoten

nicht terminale Knoten

```
<s id='beroul_pb:1_lb:19_1263221020.72'>
  <graph root='_1263221020.72'>
    <terminals>
      <t word="Certes" id="w26_00095" pos="ADV" lemma="certes"/>
      <t word="je" id="w26_00097" pos="PRO_pers" lemma="je"/>
      <t word="n'" id="w26_00098" pos="PRO_clit" lemma="ne"/>
      <t word="i" id="w26_00099" pos="PRO_clit" lemma="y"/>
      <t word="vendroie" id="w26_00100" pos="VER" lemma="venir"/>
      <t word="mie" id="w26_00101" pos="ADV" lemma="mie"/>
    </terminals>
    <nonterminals>
      <nt id="_452409.15" cat="Ng">
        <edge label='L' idref='w26_00098'/>
      </nt>
      <nt id="_221023.93" cat="VFin">
        <edge label='D' idref='_452410.38'/>
        <edge label='D' idref='_452418.9'/>
        <edge label='D' idref='_452406.05'/>
        <edge label='D' idref='_452407.65'/>
        <edge label='D' idref='_452409.15'/>
        <edge label='L' idref='w26_00100'/>
      </nt>
      <nt id="_452410.38" cat="NgPrt">
        <edge label='L' idref='w26_00101'/>
      </nt>
      <nt id="_452418.9" cat="Cmpl">
        <edge label='L' idref='w26_00099'/>
      </nt>
      <nt id="_452406.05" cat="Circ">
        <edge label='L' idref='w26_00095'/>
      </nt>
      <nt id="_221020.72" cat="Snt">
        <edge label='M' idref='_221023.93'/>
      </nt>
      <nt id="_452407.65" cat="SjPer">
        <edge label='L' idref='w26_00097'/>
      </nt>
    </nonterminals>
  </graph>
</s>
```



dependency and constituency (in TIGERSearch)



Diachronic French corpora

- *Nouveau Corpus d'Amsterdam* (NCA, 3,3 mio words, 9th-13th c., part of speech annotated, lemmatised, Stein et al., 2006).
  http://www.uni-stuttgart.de/lingrom/stein/corpus/
- *Base de Français Médiéval* (BFM, 70 texts, 3 mio words, 9th-15th c., 26 texts online, Guillot et al., 2007).
  http://bfm.ens-lyon.fr/

Syntactic Reference Corpus of Medieval French

Texts of NCA and BFM will be published with syntactic annotation in the SRCMF project.

- *Les voies du français* (MCVF, 2,5 mio words, Old French to 18th c., PENN-style syntactic annotation, Martineau, 2008)
  http://www.voies.uottawa.ca

query time...

The TIGERSearch query language

`[ ]`	each node is enclosed by `[ ]`
`[pos="P"]`	attribute `pos` has value `P`
`#p:[pos="P"]`	We can name nodes using `#name:[ ]`
`[pos="P"] . [pos="N"]`	`.` (dot) means 'precedes'
`[pos="NP"] > [pos="N"]`	`>` means 'dominates'

The TIGERSearch query language

```
#mother:[ ] > #p:[pos="P"]
& #mother > [cat="NP"]
```

any mother node dominates a preposition, and the same mother dominates noun phrase

Bloomfield, L. (1933). *Language*. Holt, New York.

Bourigault, D., Fabre, C., Frérot, C., Jacques, M.-P., and Ozdowska, S. (2005). Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles, Dourdan, France*.

Guillot, C., Marchello-Nizia, C., and Lavrentiev, A. (2007). La base de français médiéval (bfm) : états et perspectives. In Kunstmann, P. and Stein, A., editors, *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*. Steiner, Stuttgart.

Martineau, F. (2008). Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus*, 7.

Stein, A. et al., editors (2006). *Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca 1150-1350), établi par Anthonij Dees (Amsterdam 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-D. Gleßgen*. Institut für Linguistik/Romanistik, Stuttgart.

Tesnière, L. (1965). *Éléments de syntaxe structurale*. Klincksieck, Paris, 2 edition.