Syntax and Textcorpora
An Introduction to
TigerXML and TigerSearch

Achim Stein

Institut für Linguistik/Romanistik
Universität Stuttgart

29.10.2010
Texts on the Computer
- Plain text formats
- Non-Text Formats
- Annotated Text

Syntax and Textcorpora
- Syntactical Theories
- Corpus Annotation
- Corpus Queries

Practice
Plain Text (ASCII)

One character = one Byte
but which one?
One of the first standard:
ASCII (American Standard
Code for Information Inter-
change)

<table>
<thead>
<tr>
<th>binary</th>
<th>ASCII</th>
<th>char</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000001</td>
<td>65</td>
<td>A</td>
</tr>
<tr>
<td>1000010</td>
<td>66</td>
<td>B</td>
</tr>
<tr>
<td>1000011</td>
<td>67</td>
<td>C</td>
</tr>
</tbody>
</table>

plain text
A word on Word

▶ One character ≠ one byte

▶ Word has created an industrial pseudo-standard: wide-spread, but not publicly available (undocumented)

▶ Documents can be read only with specific software (as with many other commercial office software)
HTML (Hyper Text Markup Language)

Readable annotations, which are distinct from the text and in a consistent format.

```html
<b>bold</b> ⇒ bold
<font color=”blue”>blue</font> ⇒ blue
```

- HTML elements are pre-defined and standardised.
- opening code, e.g. `<font>`
  may include attribute-value-pairs, e.g. `color=”black”`
- closing code, e.g. `</font>`
The easiest way to see HTML code is to use your internet browser: menu “show source code”, “Seitenquelltext anzeigen” or similar.

HTML source code view
XML

- XML has been developed starting 1998 as an Extensible Markup Language.
  - A “document type definition” (DTD) at the top of each file (or in a separate file) defines the codes and the structure of the text.
  - XML is more general than HTML.

**Hint:**
Modern software often uses XML files in an unvisible way, using different file suffixes, e.g. *.odt, *.docx.
⇒ Rename to *.zip, open the archive, and have a look!
XML-structured text (from the TEI homepage)

<anthology>
  <poem>
    <heading>The SICK ROSE</heading>
    <stanza>
      <line>O Rose thou art sick.</line>
      <line>The invisible worm,</line>
      <line>That flies in the night</line>
      <line>In the howling storm:</line>
    </stanza>
    <stanza>
      <line>Has found out thy bed</line>
      <line>Of crimson joy:</line>
      <line>And his dark secret love</line>
      <line>Does thy life destroy.</line>
    </stanza>
  </poem>
</anthology>
XML: document type definition (DTD)

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>

<!DOCTYPE intern []
<!-- "anthology" must contain at least one "poem" -->
<!ELEMENT anthology (poem+)>
<!-- "poem" has a "heading" and at least one "stanza" -->
<!-- "poem" has an attribute "nr" with any kind of value -->
<!ELEMENT poem (heading,stanza+)>
<!ATTLIST poem nr CDATA "">
<!-- text is between "heading" elements -->
<!ELEMENT heading (#PCDATA)>
<!ELEMENT stanza (line+)>
<!ELEMENT line (#PCDATA)>
]>
```
Levels of corpus data

1. Primary data: (natural) language data
   ▶ our research interest
   ▶ problem: how to code the data?

2. Meta data: information about language
   ▶ important for all kinds of data
   ▶ documentation, bibliographical information etc.

3. Annotation
   ▶ Information on different linguistic levels
     ▶ part of speech
     ▶ syntactic categories: phrases, dependencies
     ▶ semantics: anaphores, semantic classes

vgl. e.g. Lemnitzer and Zinsmeister (2006)
Achim Stein: Syntax and Textcorpora

Texts on the Computer

Annotated Text

File Edit Options Buffers Tools SGML Help

<subcorpus id="abreja" titleDees="L’abrejance de l’ordre de chevalerie. v. 1-2266" editionDees="éd. U. Robert."
clientRegionDees="86 (Doubles)" datePVR="1836" dateNoyenne="1290" codeRegional="78" coefficientRegional="86"
_tcp="vers" ponctuation="outil" mots="1250"
sigleDEAF="n1" auteurDees="n1" dateDeuvrre="n1" dateManuscrit="n1">
<s line="1">
<word pos="PRE" deespos="301" taggerpos="PRE" lemma="par" src="S">per</word>
<word pos="ADJ:obj:masc:sg" deespos="022" taggerpos="ADJ" lemma="bon" src="+CIT">bon</word>
<word pos="ADV" deespos="311" taggerpos="ADV" lemma="ici" src="+IT">ici</word>
<word pos="VER:pres:3.sg" deespos="513x" taggerpos="VER" lemma="commercier" src="+IM">comerce</word>
<word pos="PON" deespos="int" note="ajout" taggerpos="PON" lemma="."></word>
<word pos="PREDET:en:obj:masc:sg" deespos="132:2" taggerpos="PROCON" lemma="o3loq" src="+IZ">u</word>
<word pos="NOM:obj:masc:sg" deespos="002" taggerpos="ADV" lemma="ne1" src="+Z">non</word>
<word pos="PRE" deespos="301" taggerpos="PRE" lemma="de" src="+S">de</word>
<word pos="NOM:obj:masc:sg" deespos="002" taggerpos="NOM" lemma="due" src="+IT">due</word>
<word pos="DET:def:suj:femi:sg" deespos="105" taggerpos="DET:def" lemma="le" src="S">li</word>
<word pos="NOM:suj:femi:sg" deespos="005" taggerpos="NOM" lemma="abrejance" src="+T">abrejance</word>
<word pos="PRE" deespos="301" taggerpos="PRE" lemma="de" src="+S">de</word>
<word pos="DET:def:obj:masc:sg" deespos="102" taggerpos="DET:def" lemma="le" src="S">l</word>
<word pos="NOM:obj:masc:sg" deespos="002" taggerpos="NOM" lemma="odorialdre" src="+IT+CIT">ordre</word>
<word pos="PRE" deespos="301" taggerpos="PRE" lemma="de" src="+S">de</word>
<word pos="NOM:obj:femi:sg" deespos="006" taggerpos="NOM" lemma="chevalerie" src="+IT">chevalerie</word>
<word pos="ADV" deespos="311" taggerpos="ADV" lemma="coment" src="+I">coment</word>
<word pos="VER:pres:3:sg" deespos="513x" taggerpos="VER" lemma="devoirIdire" src="+CML+M">doit</word>
<word pos="VER:inf:i" deespos="592" taggerpos="VER" lemma="entr’estrelestrere" src="+M+IT+estre">est</word>
<word pos="VER:pp:sg" deespos="582" taggerpos="VER" lemma="establir" src="+Z">establir</word>
<word pos="PON" deespos="int" note="ajout" taggerpos="PON" lemma="."></word>
</s>
<s line="2">
<word pos="VER:pp:sg" deespos="582" taggerpos="VER" lemma="faire" src="+CML">faite</word>
<word pos="VER:imp:3:sg" deespos="559x" taggerpos="VER" lemma="estrelestrere" src="+IITC">fu</word>
<word pos="PRE" deespos="301" taggerpos="PRE" lemma="par" src="+S">per</word>
<word pos="ADJ:obj:masc:sg" deespos="022" taggerpos="ADJ" lemma="noble1\n\n\nnoble" src="+ITC">\n\nnoble</word>
<word pos="NOM:obj:masc:sg" deespos="002" taggerpos="NOM" lemma="omo" src="+IT">home</word>
</s>
</subcorpus>

XML format of the *Nouveau Corpus d’Amsterdam* (Old French)
Texts on the Computer
Plain text formats
Non-Text Formats
Annotated Text

Syntax and Textcorpora
Syntactical Theories
Corpus Annotation
Corpus Queries

Practice
Syntactic Relations

- Dependency
  - On which word depends a given word?
  - Tree with arcs (branches) between words.
  - Grammatical functions (subject etc.) can be labels of the arcs.
  - vgl. Tesnière (1965)

- Constituency
  - Which words form a constituent?
  - Tree with arcs between constituents and words (at terminal level, leaves).
  - Grammatical functions are derived from the structure (subject = left daughter of IP etc.)
  - vgl. Bloomfield (1933)
**Tree Graphs**

**Terminologie**

A tree (graph) consists of **nodes** (terminal, non-terminal) and **arcs** (labelled or not).

```
  looks
   / \        / \        / \  
this like this looks
   \ /     \ /     \ /  
structure a dependency
```

```
  IP
   / \        / \  
 NP VP PP
   /   \  /   \  
 this looks like
   /   \   /   \  
 a     a constituent structure
```
Translate between syntactical representations

- Dependency graphs can be translated to constituency graphs (and vice versa)
- In the example (Bourigault et al., 2005):
  - relations (subjects etc.) are nodes
  - types of dependency are labels of arcs
Problems in all representations

- Are functional categories top nodes / heads?
  - constituency: is the NP a DP, the clause an IP etc.?
  - dependency (analog): is the top node of the nominal phrase the article or the noun?

- Modification, e.g. adverbials, attributes
  - dependency: a different type of dependency?
  - constituency: adjunction, to which node?

- Non configurational aspects of language (e.g. movement)

(1) *Syntax* I love.
(2) *Chomsky*, I hate him.
Texts on the Computer
- Plain text formats
- Non-Text Formats
- Annotated Text

Syntax and Textcorpora
- Syntactical Theories
- Corpus Annotation
- Corpus Queries

Practice
Relevance of syntactically annotated corpora

- *Treebanks* are used to test and train NLP software.
- Resources:
  - Penn Treebanks for Englisch (e.g. PPCME for Middle English)
  - Prague Dependency Treebank für Czech
  - NEGRA Treebank for German
  - ...
  - SRCMF for Old French:
    *Syntactic Reference Corpus of Medieval French*
Syntactic Annotation

Questions:

1. How to introduce structures?
   - by hand
   - automatic: rule-based, probabilistic

2. How to code structures?

3. How to query structures?
Manual annotation

- Beispiel: Notabene (SRCMF project, available on sourceforge)
**Parsing**

- Grammar- and lexikon-based, e.g. YAP Schmid (2000)
  - HPSG-like constituency rules plus lexicon
  - Generates all possible structures for an input sentence.
Parsing

- Probabilistic methods
  - Partial *chunk parsing* or complete analysis
  - Training on pre-analysed *treebanks*

(3) Der hunter sees the gazelle
    [with the binoculars].

- Ambiguity and disambiguation
  - Selection of the correct analysis: manual or probabilistic
YAP-Parser: manual disambiguation
Texts on the Computer
Plain text formats
Non-Text Formats
Annotated Text

Syntax and Textcorpora
Syntactical Theories
Corpus Annotation
Corpus Queries

Practice
Syntactic corpus queries

- Software for non-XML formats:
  - CorpusSearch (University of Pennsylvania, UPENN)
    http://corpussearch.sourceforge.net
  - PENN-Format for many English corpora: YCOE, PPCME, EME etc.

- Internal format: bracketed text with labels

```
(, .)
(CONJP (CONJ and)
  (IP-SUB (NP-SBJ *con*)
   (BEP ys)
   (NP-OBJ (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 corpus queries

- Software for XML formats:
  - TigerSearch / Tiger XML (IMS, 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

- No standards, but a clear shift towards XML-based formats. Filters for other wide-spread formats.
  - PAULA has filters for TigerSearch etc.
  - TigerSearch has filters for PENN structures etc.
Syntactic structures in TigerXML

```xml
<s id='beroul_pb:l1b:19_1263221020.72'>
  <graph root='1263221020.72'>
    <terminals>
      <t word='Certes' id='w26_000095' pos='ADV' lemma='certes'/>
      <t word='je' id='w26_000097' pos='PRO_pers' lemma='je'/>
      <t word='n' id='w26_000098' pos='PRO_clit' lemma='ne'/>
      <t word='vendr' id='w26_00100' pos='VER' lemma='vendr'/>
      <t word='m' id='w26_00101' pos='ADV' lemma='m'/>
    </terminals>
    <nonterminals>
      <nt id='_452409.15' cat='Ng'>
        <edge label='L' idref='w26_000098'/>
      </nt>
      <nt id='_221023.93' cat='VFin'>
        <edge label='D' idref='452409.38'/>
      </nt>
      <nt id='_452410.38' cat='NgPrt'>
        <edge label='L' idref='w26_000101'/>
      </nt>
      <nt id='_452418.9' cat='Cmpl'>
        <edge label='L' idref='w26_000099'/>
      </nt>
      <nt id='_452406.05' cat='Circ'>
        <edge label='L' idref='w26_000095'/>
      </nt>
      <nt id='_221020.72' cat='Snt'>
        <edge label='M' idref='221023.93'/>
      </nt>
    </nonterminals>
  </graph>
</s>
```
Dependency and Constituency (in TigerSearch)
Texts on the Computer
Plain text formats
Non-Text Formats
Annotated Text

Syntax and Textcorpora
Syntactical Theories
Corpus Annotation
Corpus Queries

Practice
Installation of TigerSearch

1. Install Tiger on your system (here and now)
2. First queries
3. Short Tiger intro (in German) and other materials:
4. Work through chapter III (Query language) of the manual (help icon or as pdf in the TigerSearch/doc subfolder)
Windows Users

- Install a Unix shell (Bash):
  http://www.cygwin.com/: download installer
  install, accept all options (requires internet access)

- Install Perl:
  http://www.activestate.com/activeperl/downloads


