Introduction to Corpus Linguistics

Sandra Kübler

kuebler@sfs.uni-tuebingen.de

Seminar für Sprachwissenschaft

University of Tübingen

Overview of the Course

- introduction to corpus lingustics (Mon)
- tokenization, lemmatization, named entity annotation (Mon)
- POS tagging (Tue)
- treebanking (Wed)
- chunk parsing, parsing (Thu)
- searching in annotated corpora (Fri)
- parallel corpora (Fri)

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

What is a Corpus?

CORPUS: (1) A collection of texts, especically if complete and self-contained: the corpus of Anglo-Saxon verse. (2) In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database. Currently, computer corpora may store many millions of running words, whose features can be analyzed by means of tagging (the addition of identifying and classifying tags to words and other formations) and the use of concordancing programs. Corpus linguistics studies data in any such corpus ...

(from *The Oxford Companion to the English Language*, ed. McArthur & McArthur, 1992)

Why Are Electronic Corpora Useful?

- as a collection of examples for linguists
- as a data resource for lexicographers
- as instruction material for language teachers and learners
- as training material for natural language processing applications
  - training of speech recognizers
  - training of statistical part-of-speech taggers and parsers
  - training of example-based and statistical machine translation systems

Examples for Linguists

Give examples for English noun phrases ...

Examples for Linguists

Give examples for English noun phrases ...

examples from the Penn treebank:

- `USX's transition from Big Steel to Big Oil`
- `Pittsburgh instead of New York or Findlay, Ohio, Marathon 's home`
- `his concern about boosting shareholder value`
- `the modest goal of becoming tax manager by the age of 46`

Examples for Linguists (2)

- a move that, in effect, raised the cost of a $7.19 billion Icahn bid by about $3 billion
- an undistinguished college student who dabbled in zoology until he concluded that he couldn’t stand cutting up frogs
- the sale of the reserves of Texas Oil & Gas, which was acquired three years ago and hasn’t posted any significant operating profits since
- not just its reserves of about 1.2 trillion cubic feet of natural gas and 28 million barrels of oil but also its pipeline, gas-gathering and contract-drilling operations

Data for Lexicographers

How many senses does the word `line` have?

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Data for Lexicographers

How many senses does the word `line` have?
14 (according to Webster's New Encyclopedic Dictionary, 1994):

1. a comparatively strong slender cord
2. a cord, wire, or tape used in measuring and leveling
3. piping for conveying a fluid
4. a row of words, letters, numbers or symbols that are written, printed, or displayed
5. something that is distinct, elongated, and narrow
6. a state of agreement (bring ideas into line)
7. a course of conduct, action, or thought (a political line)
8. limit, restraint (overstep the line of good taste) ...

Instruction for Language Learning

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

How do you say in English: `think about` or `think on`?

Instruction for Language Learning

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

How do you say in English: `think about` or `think on`?

If in doubt, ask google:

36.300.000 hits for `think about`

738.000 hits for `think on`

Types of Corpora

- mono-lingual versus multi-lingual corpora
- special-purpose, domain-specific corpora versus general-purpose, large-scale corpora
- spoken language corpora versus collections of written text
- ad-hoc corpus collections versus balanced, representative corpora
- raw text versus marked-up documents
- unannotated versus annotated corpora
- WWW as a corpus

Bad Start for Corpus Linguistics

Noam Chomsky (1957) *Syntactic Structures*:

- p. 15: "…it is obvious that the set of grammatical sentences cannot be identified with any particular corpus of utterances …
  …a grammar mirrors the behavior of the speaker, who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of new sentences."

Bad Start for Corpus Linguistics (2)

Noam Chomsky (1957) *Syntactic Structures*:

- p. 16/17: "...ones's ability to produce and recognize grammatical utterances is not based on notions of statistical approximations or the like.
  ...If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list; there appears to be no particular relation between the order of approximations and grammaticalness."

English Corpora

- **Brown Corpus**: 1 million words of written American English texts from various genres, dating from 1961
- **Lancaster-Oslo-Bergen (LOB) Corpus**: 1 million words of written British English texts, dating from 1961. Genres are parallel to the Brown Corpus.
- **British National Corpus**: 100 mio. words of written and spoken language, balanced corpus of current British English
- **Internation Corpus of English (ICE)**: national or regional varieties of English; one million word collections of contemporary spoken and written English (Great Britain, USA, Australia, South Africa, Canada, Hong Kong, India, etc.)

Non-English Corpora

- **IPI PAN Polish Corpus**: 300 mio. words
- **Czech National Corpus**: 100 mio. words
- **Hungarian National Corpus**: 80 mio. words
- **Croatian National Corpus**: 30 mio. words
- **Hellenic National Corpus**: 20 mio. words
- **METU Turkish Corpus**: 10 mio. words
- . . .

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

Parallel Corpora

- **MULTEXT-East**: for Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Resian, Romanian, Russian, Slovene, and Serbian. For most languages: Orwell's 1984.
- **Hansard Corpus**: from the official records (Hansards) of the 36th Canadian Parliament [1997-2000], 3 mio. words
- **Europarl**: extracted from the proceedings of the European Parliament; includes versions in 11 European languages: Romanic (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish. Ca. 20 mio. words.

The Need for Corpus Mark-Up

Annotation guidelines are needed in order to facilitate the accessibility and reusability of corpus resources.

Minimal information:

- authorship of the source document
- authorship of the annotated document
- language of the document
- character set and character encoding used in the corpus

Leech's Seven Maxims of Annotation

1. It should be possible to remove the annotation from an annotated corpus in order to revert to the raw corpus.
2. It should be possible to extract the annotations by themselves from the text. This is the flip side of maxim 1. Taking points 1. and 2. together, the annotated corpus should allow the maximum flexibility for manipulation by the user.
3. The annotation scheme should be based on guidelines which are available to the end user.
4. It should be made clear how and by whom the annotation was carried out.

Leech's Seven Maxims (2)

5. The end user should be made aware that the corpus annotation is not infallible, but simply a potentially useful tool.
6. Annotation schemes should be based as far as possible on widely agreed and theory-neutral principles.
7. No annotation scheme has the a priori right to be considered as a standard. Standards emerge through practical consensus.

Text Encoding Initiative (TEI)

- project sponsored by the Association for Computational Linguistics, the Association for Literary and Linguistic Computing, and the Association for Computers in the Humanities
- encoding guidelines
- link: `http://www.tei-c.org`
- define how documents should be marked-up with the mark-up language SGML (or more recently XML)

XML

- XML: Extensible Markup Language
- similar to HTML
- has no fixed “semantics”: user defines what tags mean
- recognized as international ISO standard
- formally verifiable via document type definitions (DTD)
- tools available for editing, displaying, querying

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

XML – Example

```
<CATALOG>
  <CD>
   <TITLE>Empire Burlesque</TITLE>
   <ARTIST>Bob Dylan</ARTIST>
   <COUNTRY>USA</COUNTRY>
   <COMPANY>Columbia</COMPANY>
   <PRICE>10.90</PRICE>
   <YEAR>1985</YEAR>
  </CD>
  <CD>
    <TITLE>Greatest Hits</TITLE>
    <ARTIST>Dolly Parton</ARTIST>
    <COUNTRY>USA</COUNTRY>
    <COMPANY>RCA</COMPANY>
    <PRICE>9.90</PRICE>
    <YEAR>1982</YEAR>
  </CD>
</CATALOG>
```

TEI Guildelines

Each text that is conform with the TEI guidelines consists of two parts– a header and the text itself.

The header contains information such as:

- author, title, and date
- the edition or publisher used in creating the machine-readable text
- information about the encoding practices adopted

XML Annotated Text

```
<text>
<body>
<div type=BODY>
<div type="Q">
<head>Subject: The staffing in the Commission of the European Communities
</head>
<p>Can the Commission say:</p>
<p>1. how many temporary officials are working at the Commission?</p>
<p>2. who they are and what criteria were used in selecting them?</p>
</div>
<div type="R">
<head>Answer given by <name type=PERSON><abbr rend=TAIL-SUPER>Mr</ABBR>
Cardoso e Cunha</name> on behalf of the Commission <date>(22 September
1992)</date></head>
<p>1 and 2. The Commission will send tables showing the number of temporary
staff working for the Commission directly to the Honourable Member and to
Parliament's Secretariat.</p>
</div></div></body></text>
```

Levels of Linguistic Annotation

- morphological annotation (e.g. inflection, derivation, compounding)
- morpho-syntactic annotation: part-of-speech (POS) tagging
- syntactic annotation (e.g. named entities, phrasal chunking, full syntactic analysis)
- semantic annotation (e.g. word-sense disambiguation, anaphora and coreference resolution, information structure)
- discourse annotation (e.g. dialog turns, speech acts)

Why Do We Need Annotation?

- for training NLP tools
- for finding examples
  - what is the plural form of `fish`?
  - which nouns can occur as bare nouns, without a determiner?
  - are there subjectless sentences in German? – Yes, e.g. `Mir ist kalt.` (To me is cold.)
  - is it possible in English to have something between a noun and its modifying relative clause?

Step by Step Annotation

- tokenization
- lemmatization / morphological analysis
- part-of-speech tagging
- named-entity recognition
- partial parsing
- full syntactic parsing
- semantic and discourse processing

Preprocessing the Text: Tokenization

Tokenization refers to the annotation step of dividing the input text into units called *tokens*.

Each tokens consists of either:

- a morpho-syntactic word
- a punctuation mark or a special character (e.g. &, @, %)
- a number

Tokenization – Example

before tokenization:

```
Milton wrote "Paradise Lost." Then his
wife dies and he wrote "Paradise
Regained."
```

Tokenization – Example

before tokenization:

```
Milton wrote "Paradise Lost." Then his
wife dies and he wrote "Paradise
Regained."
```

after tokenization:

```
Milton wrote " Paradise Lost .  " Then
his wife dies and he wrote " Paradise
Regained .  "
```

Why is Tokenization Non-Trivial?

- disambiguation of punctuation
  e.g. period can occur inside cardinal numbers, after ordinals, after abbreviations, at end of sentences
- recognition of complex words
  - compounds, e.g. `bank transfer fee`, `US-company`
  - mergers, e.g. `don't`, `England's`, French: `t'aime`
  - multiwords, e.g. complex prepositions `provided that`, `in spite of`

Lemmatization

- refers to the process of relating individual word forms to their citation form (lemma) by means of morphological analysis
  e.g. `stopped` $\Rightarrow$ `stop`
- provides a means to distinguish between the total number of word tokens and distinct lemmata that occur in a corpus
  e.g. helps to find all occurrences of `buy`
- is indispensable for highly inflectional languages which have a large number of distinct word forms for a given lemma

Lemmatization – German Example


```
wie       wie       +Adv+Wh+#lex+COWIE
wie       wie       +Conj+Coord+#lex+COWIE
wie       wie       +Conj+Subord+#lex+COWIE

sie       sie       +Pron+Pers+3P+Pl+Fem+Nom+#lex+PERSPRO
sie       sie       +Pron+Pers+3P+Sg+Fem+Nom+#lex+PERSPRO

offenbar  offenbaren  +Verb+Imp+2P+Sg+#lex+VVFIN
offenbar  offenbar    +Adj+Pos+Pred+#lex+ADJD

gedacht   gedenken    +Verb+PPast+#lex+VVPP
gedacht   dachen      +Verb+PPast+#lex+VVPP
gedacht   denken      +Verb+PPast+#lex+VVPP

hat       haben     +Verb+Indc+3P+Sg+Pres+#lex+VAFIN
```

Tools for Lemmatization

- **XEROX Morphological Analyzer**: comprehensive morphological analyzers for many languages including English, French, Dutch, German, Hungarian, Italian, Portuguese, Czech, Danish, Finnish, Norwegian, Polish, Russian, Turkish.
  link: `http://www.xrce.xerox.com/competencies/content-analysis/toolhome.en.html`
- **Lingsoft**: morphological anylyzers for English, Danish, German, Swedish, and Finnish
  link: `http://www.lingsoft.fi/demos.html`

Morphological Analysis

Xerox:

```
half half+Adj
half half+Adv
half half+Noun+Sg
```

Lingsoft:

```
"<half>"
"half" <Quant> DET PRE SG/PL @QN>
"half" <NonMod> <Quant> PRON SG/PL
"half" N NOM SG
"half" ADV
```