Tools for Annotating and Searching Corpora
1: General Introduction to Corpus Annotation

Stefanie Dipper

Institute of Linguistics
Ruhr-University Bochum

Corpus Linguistics Fest (CLiF)
June 6-10, 2016
Indiana University, Bloomington
Outline

1. Corpora: general introduction

2. Annotations
A collection of texts or transcripts of spoken language

Usually (nowadays): electronically available
Corpus

Ref: Kübler and Zinsmeister (2015)

- A collection of texts or transcripts of spoken language
- Usually (nowadays): electronically available
- Aim: linguistic research
  - compare: psychologists select samples of persons for psychological investigations
  - linguists select samples of texts for linguistic investigations
A collection of texts or transcripts of spoken language

Aim: linguistic research

compare: psychologists select samples of persons for psychological investigations

linguists select samples of texts for linguistic investigations

The sample (= the corpus) should be representative in order to allow us to draw conclusions about the population (= the language as such, or speakers of a certain kind, a certain register, etc.)

representative: impossible to achieve

but see “balanced” below

often: opportunistic corpora
Goal of the course

- Suppose you want to investigate a linguistic phenomenon based on some texts that you have (= your corpus)
- What you learn in this course is
  - how to apply tools to **automatically annotate** your corpus with information that are relevant to your phenomenon
    - there are tools for a lot of different kinds of linguistic information
  - and to **search your corpus** for relevant evidence
Topics of the course

Mo  General introduction to corpus annotation
Tu  Introduction to automatic tools: how do they work?
We Application of automatic tools: which are there?
Th  Searching
Today’s topics

Broad overview of:

- Corpora
  - of different types, genres, languages, . . .
- Annotation
  - of different types and formats
  - quality assurance
- Some example corpora
Corpus components

1. Language data
2. Metadata
3. Annotations
Corpus components: 1. Language data

- ‘Primary data’: the “real” language data
  (Video/audio data, scans of manuscripts)
  - rather: transcriptions of spoken language or manuscripts, typed texts
- ‘Raw data’: not (much) processed so far
Corpus components: 2. Metadata

- Metadata = data that describes other data
- Here: data that describes the corpus as a whole
  - in contrast to annotations, which describe words and sentences
Corpus components: 2. Metadata

- Metadata = data that describes other data
- Here: data that describes the corpus *as a whole*
  - in contrast to annotations, which describe words and sentences
- Information e.g. on . . .
  - the language of the corpus
  - speakers and authors, dates
  - size and format of the corpus
  - register, domain
  - annotations (see below)
  - . . .
Corpus components: 3. Annotations

- Mainly linguistic information but sometimes extra-linguistic aspects
  - linguistic: e.g. part of speech, lemma, ...
  - extralinguistic: e.g. current game scores if the text is a baseball report

- More on annotations below
Examples of corpora: Brown Corpus (1964–67)

Brown Corpus:

- balanced selection of 1 million words of American English
Examples of corpora: Brown Corpus (1964–67)

Brown Corpus:

- balanced selection of 1 million words of American English
- selection of 500 texts (of 2,000 words each) of 15 different categories
  - press reportage (44 texts), press editorials (27), religion (17), skill and hobbies (36), learned (80), science fiction (6), humor (9), ... and more ...
Examples of corpora: Brown Corpus (1964–67)

Brown Corpus:

- balanced selection of 1 million words of American English
- selection of 500 texts (of 2,000 words each) of 15 different categories
  - press reportage (44 texts), press editorials (27), religion (17), skill and hobbies (36), learned (80), science fiction (6), humor (9), . . . and more . . .
- balanced: the sizes of the different genres are intended to be proportional to the genres that occur in “reality” = in the (written) language that the average American person is probably confronted with on average in daily life
Examples of corpora: Brown Corpus (1964–67)

Brown Corpus:

- balanced selection of **1 million words** of American English
- selection of 500 texts (of 2,000 words each) of 15 different categories
  - press reportage (44 texts), press editorials (27), religion (17), skill and hobbies (36), learned (80), science fiction (6), humor (9), . . . and more . . .

- **balanced**: the sizes of the different genres are intended to be proportional to the genres that occur in “reality” = in the (written) language that the average American person is probably confronted with on average in daily life

- freely available via NLTK ([http://www.nltk.org/nltk_data/](http://www.nltk.org/nltk_data/))

Stefanie Dipper  
Tools for annotating and searching
Examples of corpora: **Brown Corpus** (1964–67)

Brown Corpus:

- balanced selection of **1 million words** of American English
- selection of 500 texts (of 2,000 words each) of 15 different categories
  - press reportage (44 texts), press editorials (27), religion (17), skill and hobbies (36), learned (80), science fiction (6), humor (9), ... and more ...

- **balanced**: the sizes of the different genres are intended to be proportional to the genres that occur in “reality” = in the (written) language that the average American person is probably confronted with on average in daily life

- freely available via NLTK ([http://www.nltk.org/nltk_data/](http://www.nltk.org/nltk_data/))

- originally stored on punched cards
Examples of corpora: Brown Corpus (1964–67)

Ref: http://www.hit.uib.no/icame/brown/bcm.html

TELEVISION IMPULSES, SOUND WAVES, ULTRA-VIOLET RAYS, ETC**, THAT MAY 0C*
COPY THE VERY SAME SPACE, EACH SOLITARY UPON ITS OWN FREQUENCY, IS INFINITE. *SO WE MAY CONCEIVE THE COEXISTENCE OF THE INFINITE NUMBER OF UNIVERSAL, APPARENTLY MOMENTARY STATES OF MATTER, SUCCESSIVE ONE AFTER ANOTHER IN CONSCIOUSNESS, BUT PERMANENT EACH ON ITS OWN BASIC PHASE OF THE PROGRESSIVE FREQUENCIES. *THIS THEORY MAKES IT POSSIBLE FOR ANY EVENT THROUGHOUT ETERNITY TO BE CONTINUOUSLY AVAILABLE at ANY MOMENT TO CONSCIOUSNESS. *

- Only upper-case letters
- Markup within the text itself
  - end of line: position marker: 1020E1F03
  - escape symbol: ‘*’
    - *SO: capitalized So
    - ETC**: ‘**.’ indicates the period of an abbreviation
Examples of corpora: BNC (1991–94)

BNC: ‘British National Corpus’

- balanced selection of **100 million words** of British English
- written and some spoken language
- selections of texts of < 45,000 words
- annotated with parts of speech and lemmas
- freely available: [http://ota.ox.ac.uk/desc/2554](http://ota.ox.ac.uk/desc/2554) (XML format)
Examples of corpora: BNC (1991–94)

BNC: ‘British National Corpus’

- balanced selection of **100 million words** of British English
- written and some spoken language
- selections of texts of < 45,000 words
- annotated with parts of speech and lemmas
- freely available: [http://ota.ox.ac.uk/desc/2554](http://ota.ox.ac.uk/desc/2554) (XML format)
- BNC Sampler
  - 2 million words, manually validated annotations
BNCweb: Searching the BNC

Main menu

Query options
Standard query
Written restrictions
Spoken restrictions
User-specific functions
User settings
Query history
Saved queries
Categorized queries
Make/edit subcorpora
Upload external data file
Additional functions
Browse a text
Scan keywords/titles

BNCweb (CQP-Edition)

Standard Query

Bloomington

Query mode: Simple query (ignore case)
Number of hits per page: 50
Restriction: None (search whole corpus)

Simple Query Syntax help

Start Query  Reset Query
BNCweb: Searching the BNC

<table>
<thead>
<tr>
<th>No</th>
<th>Filename</th>
<th>Hits 1 to 3</th>
<th>Page 1 / 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CKW 998</td>
<td>In the meantime, with funding support from the Kress Foundation, Molly Faries of Indiana University, Bloomington, is preparing to enter her extensive archive of over 10,000 infra-red reflectograms into a database.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>EBX 1976</td>
<td>Other paintings come from the Indiana University Museum of Art in Bloomington, the Museum of Modern Art and private collections on both sides of the Atlantic.</td>
<td></td>
</tr>
</tbody>
</table>

http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php
Examples of corpora: **ANC** (since 1998)

ANC: ‘American National Corpus’

- selection of currently **22 million words** of American English
- written and some spoken language
- annotated with parts of speech, parses, named entities
- OANC: ‘Open ANC’
  - 15 million words, freely available
Examples of corpora: ANC (since 1998)

ANC: ‘American National Corpus’
- selection of currently 22 million words of American English
- written and some spoken language
- annotated with parts of speech, parses, named entities

OANC: ‘Open ANC’
- 15 million words, freely available

MASC: ‘Manually annotated sub-corpus’
- 500,000 words, freely available
- manually validated annotations of various kinds (syntactic annotations, word senses, ...)
Outline

1. Corpora: general introduction

2. Annotations
Why annotations?

Annotations in corpora facilitate linguistic research:
Why annotations?

Annotations in corpora facilitate linguistic research:

1. They *abstract* from individual cases
Why annotations?

Annotations in corpora facilitate linguistic research:

1. They **abstract** from individual cases

   - e.g. research on a specific word is facilitated by lemma annotations, which abstract from the different word forms
   - the lemma *give* stands for *give*, *gives*, *gave*, *given*, *giving*
Annotations in corpora facilitate linguistic research:

1. They **abstract** from individual cases
   - e.g. research on a specific word is facilitated by lemma annotations, which abstract from the different word forms
     - the lemma *give* stands for *give, gives, gave, given, giving*
   - e.g. research on word order is facilitated by part-of-speech annotations
     - *a man who ... VERB* vs. *a man VERB who ...*, as in *a man entered who I had not seen before*
Why annotations?

Annotations in corpora facilitate linguistic research:

1. They **abstract** from individual cases
   - e.g. research on a specific word is facilitated by lemma annotations, which abstract from the different word forms
     - the lemma *give* stands for *give, gives, gave, given, giving*
   - e.g. research on word order is facilitated by part-of-speech annotations
     - *a man who ... VERB vs. a man VERB who ...*, as in *a man entered who I had not seen before*

2. They **disambiguate** ambiguous forms
   - e.g. *walk* as a verb or noun
Why annotations?

Annotations in corpora facilitate linguistic research:

1. They **abstract** from individual cases
   - e.g. research on a specific word is facilitated by lemma annotations, which abstract from the different word forms
     - the lemma *give* stands for *give*, *gives*, *gave*, *given*, *giving*
   - e.g. research on word order is facilitated by part-of-speech annotations
     - *a man who ... VERB* vs. *a man VERB who ...*, as in *a man entered who I had not seen before*

2. They **disambiguate** ambiguous forms
   - e.g. *walk* as a verb or noun

3. They facilitate **quantitative** investigations
   - e.g. how long (#characters) is an average verb in English?
Why annotations?

Annotations in corpora facilitate linguistic research:

1. They **abstract** from individual cases
   - e.g. research on a specific word is facilitated by lemma annotations, which abstract from the different word forms
   - the lemma *give* stands for *give, gives, gave, given, giving*
   - e.g. research on word order is facilitated by part-of-speech annotations
     - *a man who ... VERB vs. a man VERB who ...*, as in *a man entered who I had not seen before*

2. They **disambiguate** ambiguous forms
   - e.g. *walk* as a verb or noun

3. They facilitate **quantitative** investigations
   - e.g. how long (#characters) is an average verb in English?

Note: annotations are **interpretations**
Kinds of annotations

Cf. the overview in Gries and Berez (To appear)

1. Annotations of words
2. Annotations of segments larger or smaller than words
3. Complex types of annotations
1 Annotations of words

- part of speech (POS), morpho-syntax
  - gave: V PastTense
1 Annotations of words

- part of speech (**POS**), morpho-syntax
  - *gave*: V PastTense
  - *gave*: VBD ( Penn Treebank tag: ‘Verb, past tense’ )
1 Annotations of words

- **part of speech (POS)**, morpho-syntax
  - *gave*: V PastTense
  - *gave*: VBD (Penn Treebank tag: ‘Verb, past tense’)

- **lemma**
1 Annotations of words

- part of speech (POS), morpho-syntax
  - *gave*: V PastTense
  - *gave*: VBD (Penn Treebank tag: ‘Verb, past tense’)

- lemma

- word senses, e.g. according to WordNet
Annotations of words

- **part of speech (POS), morpho-syntax**
  - *gave*: V PastTense
  - *gave*: VBD (Penn Treebank tag: ‘Verb, past tense’)

- **lemma**

- **word senses**, e.g. according to WordNet
  - sample senses of *appear*:

<table>
<thead>
<tr>
<th>No</th>
<th>Gloss</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>give a certain impression or have a certain outward aspect</td>
<td><em>This appears to be a very difficult problem</em></td>
</tr>
<tr>
<td>2</td>
<td>come into sight or view</td>
<td><em>He suddenly appeared at the wedding</em></td>
</tr>
<tr>
<td>3</td>
<td>be issued or published</td>
<td><em>Did your latest book appear yet?</em></td>
</tr>
</tbody>
</table>
1 Annotations of words

- **part of speech (POS), morpho-syntax**
  - *gave*: V PastTense
  - *gave*: VBD (Penn Treebank tag: ‘Verb, past tense’)

- **lemma**

- **word senses**, e.g. according to WordNet
  - sample senses of *appear*:

<table>
<thead>
<tr>
<th>No</th>
<th>Gloss</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>give a certain impression or have a certain outward aspect</td>
<td><em>This appears to be a very difficult problem</em></td>
</tr>
<tr>
<td>2</td>
<td>come into sight or view</td>
<td><em>He suddenly appeared at the wedding</em></td>
</tr>
<tr>
<td>3</td>
<td>be issued or published</td>
<td><em>Did your latest book appear yet?</em></td>
</tr>
</tbody>
</table>

- **example annotation (MASC):**
  
  *The article is a classic, probably the first on the subject to appear[3] in a scholarly journal*
2 Annotations of segments ≠ words

- Named Entities (MASC):
  
  \[\text{December 1998}\]_{\text{date}}. \text{Your contribution to } \text{[Goodwill]}_{\text{other}} \text{ will mean more than you may know.}
2 Annotations of segments ≠ words

- Named Entities (MASC):
  
  \[ \text{December 1998}_\text{date}. \text{Your contribution to } \text{Goodwill}_\text{other} \text{ will mean more than you may know.} \]

- Morphological segments (and labels):

  \[
  \begin{array}{cccccc}
  \text{My} & \text{s} & \text{Marko} & \text{poexa-l-i} & \text{avtobus-om} & \text{v} & \text{Peredelkino.} \\
  \text{1PL} & \text{COM} & \text{Marko} & \text{go-PST-PL} & \text{bus-INS} & \text{All} & \text{Peredelkino.} \\
  \text{we} & \text{with} & \text{Marko} & \text{go-PST-PL} & \text{bus-by} & \text{to} & \text{Peredelkino.} \\
  \end{array}
  \]
  
  'Marko and I went to Perdelkino by bus.'

  (Leipzig Glossing Rules, Comrie, Haspelmath, and Bickel (2008))
3 Complex types of annotations

We must strike a balance between them.
3 Complex types of annotations

We must strike a balance between them.
Annotations: quality assurance

Ref: Zinsmeister, Witt, Kübler, and Hinrichs (2008)

- **Consistency**
  - the same linguistic phenomena are annotated in the same way
  - similar or related phenomena must receive annotations that represent their similarity or relatedness if possible
Annotations: quality assurance

Ref: Zinsmeister, Witt, Kübler, and Hinrichs (2008)

- **Consistency**
  - the same linguistic phenomena are annotated in the same way
  - similar or related phenomena must receive annotations that represent their similarity or relatedness if possible

- Crucial for consistency:
  1. annotation guidelines
  2. consistency checks
  3. inter-annotator agreement
1 Annotation guidelines

- Definition of the *tagset* (= set of annotation labels)
- Annotation criteria
1 Annotation guidelines

- Definition of the **tagset** (= set of annotation labels)
- Annotation criteria

Many texts are not models of good prose, and some contain outright errors and slips of the pen. Do not be tempted to correct a tag to what it would be if the text were correct; rather, it is the incorrect word that should be tagged correctly.

Penn Treebank guidelines (Santorini 1990, p. 1)
1 Annotation guidelines

- Definition of the **tagset** (= set of annotation labels)
- Annotation criteria

*We adopt the general convention that parts of speech are defined on the basis of their syntactic distribution rather than their semantic function. This convention has several important consequences. One is that nouns in prenominal position that are being used as modifiers are tagged as nouns (NN), not as adjectives (JJ) (see Section 4-JJ or NN).*

**EXAMPLES:**
- a cotton/NN shirt
- the nearest book/NN store

Penn Treebank guidelines (Santorini 1990, p. 31)
1 Annotation guidelines

- Definition of the **tagset** (= set of annotation labels)
- Annotation criteria

```
Adjective—JJ
Hyphenated compounds that are used as modifiers are tagged as adjectives (JJ).

EXAMPLES:  happy-go-lucky/JJ
            one-of-a-kind/JJ
            run-of-the-mill/JJ
```

Penn Treebank guidelines (Santorini 1990, p. 2)
2 Consistency checks

- Search for rare constructions → they are likely to be annotation errors (Dickinson 2005)
- Multiple annotators annotate the same text → compare their results
3 Inter-annotator agreement

Agreement between multiple annotators: can be measured by different measures (Scott’s $\pi$, Cohen’s $\kappa$, Fleiss’ $\kappa$, . . .)
3 Inter-annotator agreement

- Agreement between multiple annotators: can be measured by different measures (Scott’s $\pi$, Cohen’s $\kappa$, Fleiss’ $\kappa$, . . .)

→ Come to this afternoon’s practical session to learn more about it!

- we manually annotate some sample data with the Penn Treebank Tagset
- . . . and compute our inter-annotator agreement
- using the tool WebAnno
Today’s session

- ‘Corpus’: primary data, metadata, annotations
- Annotations
  - why are they useful
  - different kinds
- Example corpora and annotations
Today’s session

- ‘Corpus’: primary data, metadata, annotations
- Annotations
  - why are they useful
  - different kinds
- Example corpora and annotations

Tomorrow:
- How can we produce such annotations?
  → We have a deeper look at some automatic tools and how they work
References I


*Corpus Linguistics and Linguistically Annotated Corpora.*
London: Bloomsbury Publishing.

*Part-of-Speech Tagging Guidelines for the Penn Treebank Project.*
University of Pennsylvania.
3rd Revision.

Linguistically annotated corpora: Quality assurance, reusability and sustainability.
In A. Lüdeling and M. Kytö (Eds.), *Corpus linguistics: An international handbook,* Volume 1, pp. 759–776. Berlin: Mouton de Gruyter.