Introduction to Corpus Resources, Annotation and Access

Sabine Schulte im Walde
Universität des Saarlandes

Heike Zinsmeister
Universität Tübingen

Foundational Course
18th European Summer School in Logic, Language and Information
Málaga, Spain
July 31 - August 4, 2006

Programme

<table>
<thead>
<tr>
<th>Day</th>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>07/31</td>
<td>Introduction</td>
</tr>
<tr>
<td>Tue</td>
<td>08/01</td>
<td>Tokenisation &amp; morpho-syntactic annotation</td>
</tr>
<tr>
<td>Wed</td>
<td>08/02</td>
<td>Syntactic annotation</td>
</tr>
<tr>
<td>Thu</td>
<td>08/03</td>
<td>Semantic annotation</td>
</tr>
<tr>
<td>Fri</td>
<td>08/04</td>
<td>More levels of corpus annotation Web as corpus</td>
</tr>
</tbody>
</table>

Each day: resource examples, tools, exercises

Overview

1. Empirical approach
2. What is a corpus and what is in it?
3. Standardisation efforts
4. Frequency distributions

Empirical Approach

Two approaches to linguistics

Linguistics: characterisation and explanation of linguistic observations.

- Competing approaches: rationalism vs. empiricism
- Competence (abstraction) vs. performance
- Deductive method: from the general to the specific; rules are derived from axioms and principles; verification of rules by observations
- Inductive method: from the specific to the general; rules are derived from specific observations; falsification of rules by observations.
### Empirical approach

- Describing naturally occurring language data
- Objective (reproducible) statements about language
- Quantitative analysis: common patterns in language use
- Creation of robust tools for Natural Language Processing (NLP) by applying statistical and machine learning approaches to large amounts of language data.
- Empirical turn supported by rise in processing speed of computers and their amount of storage — and the revolution in the availability of machine-readable texts (scanners (OCR devices), e-mails, the world wide web).

### Empirical resources

- Corpora: large amounts of texts
- Dictionaries and thesauri, e.g. Oxford advanced Learner’s Dictionary of current English, Roget's thesaurus
- Morphological database and analyser, e.g. U Penn’s XTAG
- Semantic hierarchies, e.g. WordNet
- Annotation tools, e.g. TreeTagger, Collins' Parser, Stanford Parser
- Processing tools, e.g. U Penn's tgrep (tregex), TIGERSearch

---

### Empirical approach: References


Online: http://bowlandfiles.lancs.ac.uk/monkey/keu/linguistics/corpus1/


---

### What is a corpus and what is in it?

---

### Definition of ‘Corpus’

- Any collection of more than one text. (McEnery & Wilson 2001)

- A large body of linguistic evidence typically composed of attested language use. (McEnery 2003)

- A collection of electronic texts built according to explicit design criteria for a specific purpose. (Altheier et al. 1992)

- A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria, in order to be used as a sample of the language. (Sinclair 1996)

---

### Attested language use

- Naturally occurring language

- Spoken language: performance errors such as slips of the tongue, hesitations, corrections due to short term memory limitations, general state of mind, alcohol level etc.

- Written language: newspapers, manuals, fiction, public speech, plays, chat-language and e-mails. Errors such as misspellings, misspelling, missing/additional words.

- Creativity of language

- Context-dependency of language, e.g. ellipsis
### Sample of a language

Corpora give only a partial description of a language
- they are incomplete
  - the Brown Corpus doesn’t include vocabulary related to the world wide web and e-mail
- they are biased
  - prominent topics in Wall Street Journal subcorpus of Penn Treebank
- they include ungrammatical sentence
  - typos, copy-and-past errors, conversion errors

Sample a corpus according to design criteria such that it is balanced and representative for a specific purpose.

→ "But knowing that your corpus is unbalanced is what counts."
   (Atkins et al., 1992: 6)

### Specific purpose: Example

- Task: developing a machine translation system for dialogues on meeting arrangements
- Creation of a corpus to assist this task (as training and testing data).
- Sampling frame:
  - telephone-based dialogues on meeting arrangements
  - different types of meetings
  - different speakers (varying features such as age, gender, acquaintance, nationality etc.)

Verbomobil corpus, TuBa-DS Treebank.

### Purpose: Reference corpus

- Task: create a representative corpus of British English
- Sampling frame:
  - 100 million words
  - 90 % written language
    - medium: book, newspaper, un-published material, ...
    - theme: informative, imaginative...
    - language level
      - information on the author and on the ‘audience’
      - samples of 40,000 words per text
  - 10 % spoken language
    - topics: educational, business, institutional, leisure ...
    - demographic parameter: age, social group, gender, region, type of interaction (monologue/dialogue...)
  - the British National Corpus (BNC)

### Corpus typology & text typology

- Classification of corpus and text types
- Contrastive parameters (adapted from Atkins et al. 92)
- Corpus typology:
  - How is the corpus data related to the original text?
  - How is it related to the language that is represented?
  - Which language(s) are represented? If more than one language, how are the subcorpora related?
  - Which period of time does it represent?
  - Does it comprise annotation? If yes, what kind of?
- Text typology:
  - In what mode was the primary data delivered?
  - In which medium was it produced / published?
  - What genre does it belong to?
  - What function does it have?

### Corpus typology: Relation to original text

- Full text:
  - Penn treebank (AE): subcorpus of Wall Street Journal editions
  - Kant Corpus (Gi): writing of philospher Immanuel Kant
- Sample:
  - Brown (AE), Limas (Gi): 500 samples of 2,000 words
  - BNC (BE): samples of max. 40,000 words/text
- Monitor:
  - Texts scanned on continuing basis; ‘filtered’ to extract data for database, but not permanently archived; data flow.
    - Wortwarte (Gi): scans 10 newspapers. ~ 1 mil. token/day, word list is archived, texts are deleted after three days.
  - Permanently growing corpus

### Corpus typology – Time & Language

- Relation to time
  - Synchronic: represents a specific period
  - Diachronic: represents language change in time
    - A Historical Corpus of the Welsh Language: 1500-1850
    - Corpus del Español: 1200-2000
- Relation to language
  - General: reference corpus
    - BNC (BE), DWDS Kerntorus (German)
  - Terminological: special corpus; specialised language
    - Technical Corpus of IULA: economics, law, medicine...
  - (Opportunistic collection (added HZ))
### Corpus typology: Language(s)
- **Language(s) of corpus**
  - English, German, Spanish, ...
- **Monolingual**
- **Bilingual or multilingual**
  - Parallel corpora: original text and its translation(s):
    - alignment: document, sentence, (multi-word, ...)
      - Europarl corpus: Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish
  - Translation corpora: also ‘comparable corpora’: original texts of same genre in different languages
    - Parole corpora: Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish

### Text typology: Text attributes
- **Mode**: written, written-to-be-read, written-to-be-spoken, spoken, spoken-to-be-written
- **Text origin**: single, several, joint, ...
- **Medium**: book, newspaper, classroom lessons, ...
- **Style**: prose, verse (blank verse, rhyme, ...)
- **Genre**: novel, short story, play, poem, essay, letter (business, personal), lecture, debate, speech, conversation, classroom lessons, advertisement, law, article, horoscope, examination, ...
- **Setting**: unclassified, education, work, leisure, public affairs, ...

### Definition & typology: References
**Corpora – online references**
- The British National Corpus (BNC) http://www.comp.lancs.ac.uk/computing/research/trip/bnc.html
- Bonner Kanti-Korpus https://www.dbp.uni-bonn.de/diit/forschkant/ (in German)
- Corpus del Español http://www.corpusdelspanol.org/

**Text typology: Text attributes**
- **Function**: (discourse force): unmarked, narrative, informative, expository, regulative/instructional, entertaining, ...
- **Topic**: general, science (biology, chemistry, ...), music (opera, pop, ...), animals (dogs, ...), ...
- **Date**: date of (first) publication, date of speech event
- **Language status**: original, translation
- **Native language(s) of author(s)**
- **...**

**Definition & typology: References**
  - online: http://www.cs.lancs.ac.uk/mooney/myle/linguisticcontents.htm section 2.
  - online: http://www.lc.cri/1/EAGLES96corpuspdf/corpuspdf.html

**DWDS Kernkorpus**
http://www.dwds.de/textbasis/kernkorpus (in German)

**Europarl Parallel Corpus**
http://people.csail.mit.edu/koehn/publications/europarl/

**A Historical Corpus of the Welsh Language 1500-1850**
http://people.pwf.carthage.edu/~ww2/howto/mwmc.html

**Limas-Korpus**
http://www.iip.uni-bonn.de/Limas/ (in German)

**Parole Corpora**
Definition & typology: References

- Penn Treebank
  http://www.clir.upenn.edu/~treebank/
- Technical Corpus of the Institut Universitari de Lingüística Aplicada Pompeu Fabra, Barcelona
  http://bwamnet.iula.upf.edu/ (in Spanish)
- VerbTree / TIM-SI
  VM1: http://www.phonetik.uni-muenchen.de/Bas/Bas/M1eng.html
  VM2: http://www.phonetik.uni-muenchen.de/Bas/Bas/M2eng.html
  TIM-SI: http://www.sfb.uni-luebeck.de/en_luebeck.shtml
- Wortwerte
  www.wortwerte.de/ (in German)

Annotation:

- The practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language.
- The end-product of this process: the linguistic symbols which are attached to, linked with, interspersed with the electronic representation of the language material itself.
- Question of granularity: how much detail should be encoded through annotation?

Annotation: Motivation

- Extracting linguistic information
  - language is ambiguous → disambiguation by annotation
  - e.g. ‘my left hand’ (JJ), ‘on your left’ (NN). ‘left early’ (VBZ). more complex grammatical phenomena
  - e.g. a direct object modified by a non-adjacent relative clause
  - I met friends in Rome, who were there for the first time.
- Re-usability
  - annotation is time-consuming and expensive.
  - automatic annotation often requires contextual (annotation) information
  - higher-level annotation often relies on lower-level annotation.
- Multi-functionality
  - same corpus used for various applications e.g. parser training, lexicography, speech synthesis, machine-aided translation, information retrieval.

Annotation types

- Morpho-syntactic information
- Lemmata
- Syntactic categories / dependencies
- Grammatical functions
- Senses
- Semantic roles
- Prosody
- Information structure: topic / focus
- Anaphora and coreference relations
- Named Entities
- Events
- Discourse structure relations
- Time
- Emotions ...

Annotation: principles of good practice

1. The raw corpus (primary data) should be recoverable.
2. Annotation should be extensible from the corpus, to be stored independently if there is a need.
3. Easy access to documentation
   (a) annotation scheme
   (b) how, where, by whom the annotation was applied
   (c) some account of the quality of annotation.
   (adapted from Leech 1997.6)

Annotation: representation

Column-based format: Brown Corpus
- Text-only version (Form A): 1963-64 on punched cards. 80 spaces/line: 70 for text, 9 for location marker, space #71 kept empty.
- Tagged version (Form C): 1979) on magnetic tapes:
  - Columns 1-30 the word or external punctuation symbol
  - Columns 31-41 the grammatical tag
  - Columns 42-52 an eleven-character location marker.

SAID VBD A01001006E1
Annotation: representation

Column-based, vertical format: LOB Corpus
A014010 AT a P
A014020 NN move
A014030 TO to
A014040 VB stop
A014050 NFP \Hr \O
A014060 NP Gaitskell

Horizontal format: LOB Corpus
A014: a NT move MN to TO stop VN \Hn NFP Gaitskell NNP

TEI-conform encoding (text-internal entity reference)
<a id="move" href="#">move</a>
→ Horizontal format is problematic if annotation is more complex.

Annotation: representation

Hierarchical structure annotation: Penn Treebank bracketing format

( ( (NP (TP (PRP (DT This)) (VP (VBZ means))) ...) )
( (NP (ATP (DT the)) (VBZ returns)) ...
( (TP (MD do)) ...
( (VP (IN great)) ...
( (PP (IN deal)) ))) )))

Tags:
- part of speech
- phrasal category
- function

Annotation: representation

XML inline representation

<sentence id="109880418:46.26" origin="Col55:318">This means</sentence>

<word form="this" pos="JJ" func="MOD"/>
<word form="mean" pos="VBD" func="SUBJ"/>
<word form="将会" pos="MD" func="MOD"/>
<word form="return" pos="VBP" func="SUBJ"/>
<word form="and" pos="CC" func="MOD"/>
<word form="great" pos="JJ" func="MOD"/>
<word form="Deal" pos="NN" func="MOD"/>

Tags:
- part of speech
- phrasal category
- function

Annotation: representation

XML stand-off representation

<sent>
<nt id="442_1" word="This" pos="DT"/>
<nt id="442_2" word="means" pos="VBD"/>
<nt id="442_3" word="will" pos="MD"/>
<nt id="442_4" word="return" pos="VBP"/>
<nt id="442_5" word="and" pos="CC"/>
<nt id="442_6" word="great" pos="JJ"/>
</sent>

Tags:
- part of speech
- phrasal category
- function

Annotation scheme

A detailed specification of the annotation
- A list of symbols used in the annotation such as terminals (e.g. parts of speech), non-terminals (e.g. syntactic category labels), and other symbols.
- A basic definition of the symbols, e.g. "U-adjective".
- A description as detailed as possible, of how the symbols are applied to text sentences, e.g.,
  - How do annotators recognize a Noun Phrase (NP) when they see one?
  - How do they distinguish NP tokens from words or word sequences which are not NPs?
Annotation scheme types

- Comprehensive grammar
  - difficult for annotators to keep track of
  - difficult to update
- Set of guidelines
  - evolving laws of precedence
  - recorded in annotator’s manual (also ‘tagging manual’)
- Reference treebank (also ‘benchmark treebank’)
- Mixed form
  - cross-referenced guidelines and examples

Annotation: principles of good practice

Additional maximes:

2. Annotation schemes made available to research community on caveat emptor principle (‘the seller cannot be held responsible for the quality of the good, unless it was warranted.’)

3. Annotation should depend on consensual or theory-neutral analyses.

4. No annotation scheme should claim authority as an absolute standard.

(Adapted from Leech 1997:6)

Exploitation of annotated corpora in NLP

- Quantitative data
- Disambiguation is a key problem in many areas such as parsing, anaphora resolution or machine translation
  - Example: CLAWS tagger (‘Constituent-Like Automatic Word-Tagging System’)
  - TAGGIT based on hand-crafted rules was used to tag the Brown Corpus — accuracy of ~77 %
  - A subset of the Brown corpus was adapted to the CLAWS tagset. Sequences of two tags were collected in a bigram matrix for calculating lexical and contextual probabilities.
  - CLAWS uses these probabilities for choosing the right tag in a given local context; e.g. LOB Corpus — accuracy of ~97 %.

Exploitation of annotated corpora in NLP

- Terminology extraction
- Evidence-based learning
- Tested for the evaluation of NLP programs

Annotation: References


Annotation: References

Online resources:

CHILDES (Child Language Data Exchange System)
http://childes.psychol.cmu.edu/

CLAWS tagger
http://www.comp.lancs.ac.uk/urnl/dews/

LOB Corpus (Lancaster-Oslo/Bergen Corpus)
http://clweb.essex.ac.uk/w3c/corpus/ling/content/corpora/list/privates/LOB/lob.html
Metadata

A corpus contains different kinds of data:

- Primary data
  * digital language data

- Annotation
  * linguistic interpretation of the primary data

- Metadata
  * contextual information about the primary data
  * documentation for subsequent users
  * key to retrieve particular types of primary data

- Meta-metadata
  * contextual information about the meta-data, e.g. who created the meta-data, when and why

(CES-conform TUSNELDA header)

<tsnslheader version="1.0" date-update="06.09.2001"
    creator="BB/BM1991\project B8" type="corpus">
    <filesdesc>
      <fileset>
        <fileset>
          <files>
            <filename/>
          </files>
        </fileset>
      </filesdesc>
    </tsnslheader>

Standardisation

Motivation

- Corpus and annotation formats vary considerably often to satisfy constraints of particular processing software.
- Community wants to share, merge and compare language resources.

→ commonality and interoperability necessary

Objectives

- Diversity in theoretical approaches.
- Existing resources may be rendered obsolete if they don’t fit the standard.

Standardisation: Architecture

Corpus in user-defined format

Mapping

Ontology of linguistic terms and concepts

Tools

Meta-data

Generic corpus data

XML-database

Open document format

Web interface
### Standardisation

- **Infrastructure**
  - Linguistic Annotation Framework (ISO TC 37 / SC 4)
  - German Sustainability project (SFB 441)
- **Ontologies of linguistic terms and concepts**
  - Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)
  - Generalized Ontology for Linguistic Description (GOLD, created in E-MELD project)
  - Data Category Registry (DCR, created in ISO TC 37 / SC 4)
- **Metadata**
  - isle Metadata Initiative (IMI)
  - Open Language Archive Community (OLAC)

### Sustainability

- **New developments in computer technology allow to capture, store, annotate and disseminate digital data.**
- **Uncritical adoption of new technologies compromises ability to preserve data.**
- **Desired: portability of digital language resources across environments, scholarly communities, domains of application and passage of time**

  [The following list is an excerpt from Bird and Simons (2003), see paper for full list.]

### Sustainability: Problem areas

- **Discovery:** the problem of finding existing resources and knowing whether they are relevant.
  - Documents are ‘hidden’ in linguists’ personal collection of computer files.
  - No publicly available metadescription.
- **Access:** unclear scope and process.
- **Citation:** URLS break, confusion of different versions of same resource.
- **Preservation**
  - Formats become obsolete.
  - Absence of supporting hardware (e.g. 5.25” floppy disks).
  - Lifespan of physical medium (digital media: 5 years).
- **Rights:** what a potential user is permitted to do with the resource.

### Sustainability: Principles of best practice

- **Content**
  - Document methods, provide original resources (e.g. recordings).
  - Map terminology and abbreviations to a common ontology of linguistic terms.
- **Format**
  - Use open formats, free tools, published proprietary formats.
  - Use Unicode for encoding
  - Use XML for markup.
- **Discovery**
  - List resources in e.g. OLAC repository.
  - Include metadata and keywords for search engine.
Sustainability: Principles of best practice

- Access
  - Publish documentation.
  - Document restrictions on and process for access.
  - Provide web access, CD/DVD and print version.
- Citation
  - Provide fixed versions.
- Preservation
  - Commit documentation and description to digital archive.
  - Refresh offline digital storage at regular intervals.
- Rights
  - Ensure that resource may be used for research purposes.

Metadata, Representation, Sustainability: References


LAF


Sustainability project


Metadata, Representation, Sustainability: Online resources

- CES: http://ics.vassar.edu/ces/
- XCES: http://www.xml-ces.org/
- DOLCE: http://www.oa-cnr.it/DOLCE/xml
- EAGLES: http://www.ldc.upenn.edu
- E-MELD: http://emeld.org
- IMDI: http://www.mpi.mpg.de
- Linguistic Data Consortium (LDC): http://www.ldc.upenn.edu
- NITE: http://nle.nis.sdu.dk
- OLAC: http://www.language-archives.org/
- TEI: http://www.tei-c.org/ (chap 5, chap 23)
- TUSNELDA: http://www.sfb441.uni-uebingen.de/tusnelda-online.html

Frequency distributions

Token and types

- Frequency information is distinctive to corpus-based methodologies.
- What is counted?
  - all the instances (tokens) of all distinct words (types) that occur in the corpus
- Example:
  Across the bridge Wayne Bridge saw the stadium. The supporters could also see him.

  → Count the tokens and the corresponding types.
  - number of tokens → corpus size \( N \)
  - number of types → vocabulary size \( L \)
**Token – Type mapping**

- Determination of tokens
  - text segmentation, ‘tokenisation’ (see tomorrow’s class)
  - include / exclude non-word items?
- Mapping of tokens to types
  - normalise upper and lower case?
  - lemmatised inflected word forms?
    - list of word/lemma correspondence
    - lemma of unknown words?
  - unknown
  - the word itself
- What is not distinguished?
  - different word senses (see Thursday’s class)

---

**Basics for lexical statistics**

<table>
<thead>
<tr>
<th>type</th>
<th>frequency</th>
<th>rank / frequency profile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Frequencies of Brown Corpus**

<table>
<thead>
<tr>
<th>top frequencies</th>
<th>million frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>rank</td>
<td>word</td>
</tr>
</tbody>
</table>
| 1  | the | 0-2 | 'It was the best 
thing since the world ended.' |
| 2  | of  | 3-4 | 'I need to get of 
my backside.' |
| 3  | and | 5-6 | 'The readers and 
the writers.' |
| 4  | to  | 7-8 | 'The truth is to 
get up.' |
| 5  | a   | 9-10 | 'The six a 
were the twenties.' |
| 6  | in  | 11-12 | 'He was in 
the car.' |
| 7  | it   | 13-14 | 'It was an 
important day.' |
| 8  | that | 15-16 | 'That was the 
biggest day.' |
| 9  | for | 17-18 | 'For the first 
time.' |
| 10 | the | 19-20 | 'The nature of 
the problem.' |

Table 4: Top and bottom of the Brown frequency list

*Baroni, prefata: 5*

**Frequencies of Brown Corpus**

**Zipf’s law**

Frequency is a non-linearly decreasing function of rank.
- It decreases more sharply among high ranks than among low ranks.
- ‘Large number of rare events’ (LNRE) distribution
- First studied by George Kingsley Zipf (1949, 1965)
- Zipf’s law predicts the frequency of a word given its rank:

\[
f(w) = \frac{C}{r(w)^a}
\]

- \(f(w)\) = frequency of word \(w\)
- \(r(w)\) = rank of word \(w\)
- \(C\) = frequency of most frequent word
- \(a\) = a constant.

Given \(C = 60,000\) and \(a = 1\):

<table>
<thead>
<tr>
<th>(f(w))</th>
<th>(f(w))</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>60,000</td>
</tr>
<tr>
<td>3</td>
<td>20,000</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\[
f(w) \approx 60,000 \times \frac{1}{x^{1.5}}
\]

\(\approx\) about 80,000 words have \(f(w)\) between 1.5 and 0.5.
Frequency: References


Introduction to lexical statistics: