

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

- Mon 07/31 **Introduction**
- Tue 08/01 **Tokenisation & morpho-syntactic annotation**
- Wed 08/02 **Syntactic annotation**
- Thu 08/03 **Semantic annotation**
- Fri 08/04 **More levels of corpus annotation**
Web as corpus

each day: resource examples, tools, exercises

Introduction

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 databasis and analyser, e.g. UPenn's XTAG
- Semantic hierarchies, e.g. WordNet
- Annotation tools, e.g. TreeTagger, Collins' Parser, Stanford Parser
- Processing tools, e.g. UPenn's tgrep (tregex), TIGERSearch

Empirical approach: References

Tony McEnery (2003): "Corpus Linguistics" In "The Oxford Handbook of Computational Linguistics", pp. 448-463. Oxford University Press.

Tony McEnery and Andrew Wilson (2001): "Corpus Linguistics". 2nd edition. Edinburgh University Press, chapter 1.

Online:
<http://bowlandfiles.lancs.ac.uk/monkey/ihe/linguistics/corpus1/>

Lothar Lemnitzer and Heike Zinsmeister (2006): "Korpuslinguistik. Eine Einführung". Narr, Tübingen, chapter 2 (in German).

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**. (Atkins 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, misediting, 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.)
- Verbmobil corpus, TüBa-DS Treebank.

Purpose: Reference corpus

- Task: create a **representative corpus of British English**
 - Sampling frame:**
 - » 100 million words
 - » 90 % **written** language
 - time of creation: 1960-1974, 1975-1993
 - 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
 - topic: 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) is/are represented? If more than one language, how are the subcorpora related?
 - » Which period of time does it present?
 - » 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 Korpus (G): writing of philosopher Immanuel Kant
- Sample**
 - Brown (AE), Limas (G): 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 (G): scans 10 newspapers, ~ 1 mill. token/day, word list is archived, texts are deleted after three days.
 - Permanently growing corpus
 - Bank of England (E): in 2005: 450 million words.

Corpus typology – Time & Language

Relation to time

- Synchronic:** represents a specific period
 - Brown Corpus (AE): texts published in 1961.
- 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 Kerncorpus (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, ...

Text typology: Text attributes

- **Function** (illocutionary force): unmarked, narrative, informative, expository, regulatory/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

- Sue Atkins, Jeremy Clear and Nicholas Ostler (1992): "Corpus Design Criteria". In *Literary and Linguistic Computing*, 7 (1), 1-16.
- Tony McEnery (2003): "Corpus Linguistics" In: Ruslan Mitkov, editor "The Oxford Handbook of Computational Linguistics", pp. 448-463. Oxford University Press.
- Tony McEnery and Andrew Wilson (2001): "Corpus Linguistics". 2nd edition. Edinburgh University Press, chapter 2.
[online](http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm): <http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>: section 2.
- John Sinclair (1996) "EAGLES. Preliminary recommendations on Corpus Typology". EAGLES Document EAG-TCWG-CTYP/P
[online](http://www.ilc.cnr.it/EAGLES96/corpus/typ/corpus.html): <http://www.ilc.cnr.it/EAGLES96/corpus/typ/corpus.html>.
- John Sinclair (2005) "Corpus and Text - Basic Principles" In: Martin Wynne, editor "Developing Linguistic Corpora: a Guide to Good Practice". Oxford: Oxbow Books: 1-16. Available [online](http://ahds.ac.uk/linguistic-corpora/) from <http://ahds.ac.uk/linguistic-corpora/> [Accessed 2006-07-26].

Definition & typology: References

Corpora – online references

- The Bank of English (COBUILD - 'Collins Birmingham University International Language Database' & HarperCollins Publishers):
<http://www.titania.bham.ac.uk/docs/about.htm>
- The British National Corpus (BNC)
<http://www.comp.lancs.ac.uk/computing/research/ucrel/bnc.html>
- Bonner Kant-Korpus
<http://www.ikp.uni-bonn.de/dt/forsch/kant/> (in German)
- Brown Corpus
http://en.wikipedia.org/wiki/Brown_Corpus
- Corpus del Español
<http://www.corpusdelespanol.org/>

Definition & typology: References

- DWDS Kerncorpus
<http://www.dwds.de/textbasis/kerncorpus> (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.cam.ac.uk/dwew2/hcwl/menu.htm>
- Limas-Korpus
<http://www.ikp.uni-bonn.de/Limas/> (in German)
- Parole Corpora
<http://www.elda.org/catalogue/en/text/doc/parole.html>

Definition & typology: References

Penn Treebank
<http://www.cis.upenn.edu/~treebank/>

Technical Corpus of the Institut Universitari de Lingüística Aplicada
Pompeu Fabra, Barcelona
<http://bwananet.iula.upf.edu/> (in Spanish)

Verbmobil / TüBa-DS
VM1: <http://www.phonetik.uni-muenchen.de/Bas/BasVM1eng.html>
VM2: <http://www.phonetik.uni-muenchen.de/Bas/BasVM2eng.html>
TüBa-DS: http://www.sfs.uni-tuebingen.de/en_tuebads.shtml

Wortwarte
www.wortwarte.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), 'I *left* early' (VBD).
 - » 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 **extricable** 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.

```
TELEVISION IMPULSES, SOUND WAVES, ULTRA-VIOLET RAYS, ETC*., THAT MAY 1020E1F03  
OC* 1025E1F03  
COPY THE VERY SAME SPACE, EACH SOLITARY UPON ITS OWN FREQUENCY, IS INF 1030E1F03  
INITE. *SO WE MAY CONCEIVE THE COEXISTENCE OF THE INFINITE NUMBER OF U 1040E1F03
```

- **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      NPT \OMr       \O
A014060      NP  Gaitskell
```

Horizontal format: LOB Corpus

```
A014 ^ a_AT move_NN to_TO stop_VB \OMr_NPT Gaitskell_NP
```

TEI-conform encoding (text-internal 'entity reference')

```
a&at; move&nn;
```

→ Horizontal format is problematic if annotation is more complex.

Annotation: representation

Multi-layered (potentially overlapping) annotation: CHAT format, CHILDES database

```
@Situation: Ross giggling and laughing
```

```
...
```

```
*FAT: would you huh ?
```

```
%snd:"boys07a2" 611973 613253
```

```
%mor: v:aux|would pro|you co|huh ?
```

```
*CHI: I do that !
```

```
%snd:"boys07a2" 613253 614822
```

```
%mor: pro|I v|dO pro:dem|that !
```

```
%par: yelled happily ...
```

(adapted from boys07a-in.cha, MacWhinney)

CHI = Ross Target_Child (1;4.11)

FAT = Brian Father

%snd = information on digitised audio file, time in milliseconds

%mor = morpho-syntactic information

%par = information on paralinguistic behaviour

Annotation: representation

Hierarchical structure annotation: Penn Treebank bracketing format

```
( (S
  (NP-SBJ (DT This) )
  (VP (VBZ means)
    (SBAR (-NONE- 0)
      (S
        (NP-SBJ (DT the) (NNS returns) )
        (VP (MD can)
          (VP (VB vary)
            (NP (DT a) (JJ great) (NN deal) )))))
      (. .) ))
  (source: wsj0728.mrg)
```

tags:

part of speech, phrasal category, function

Annotation: representation

XML inline representation

```
<sentence editor="korder" date="1998080418:46:20"
  origin="cd15e1.export" comment="%%
  &lt;g001acln1_015_AAJ_150000_E&gt;";>
  <node cat="S" func="--" parent="0">
    <node cat="NP" func="SBJ">
      <word form="I" pos="PP" func="HD"/>
    </node>
    <word form="have" pos="VBP" func="HD"/>
    <node cat="VP" func="COMP">
      <word form="to" pos="TO" func="HD"/>
      <node cat="VP" func="COMP">
        <word form="go" pos="VB" func="HD"/>
        <node cat="PS" func="ADJ">
          <word form="to" pos="IN" func="HD"/>
          <node cat="NP" func="COMP">
            <word form="Berlin" pos="NP" func="--"/>
          </node> ...
        </sentence>
        (TüBa-D/S, cd15, sentence 22)
```

part of speech, phrasal category, function

Annotation: representation

XML stand-off representation

```
<terminals>
  <terminal id="s42_1" word="I" pos="PP" />
  <terminal id="s42_2" word="have" pos="VBP" />
  <terminal id="s42_3" word="to" pos="TO" />
  <terminal id="s42_4" word="go" pos="VB" />
  <terminal id="s42_5" word="to" pos="IN" />
  <terminal id="s42_6" word="Berlin" pos="NP" /> ...
</terminals>
<nonterminals>
  <nt id="s42_500" cat="NP">
    <edge label="HD" idref="s42_1" />
  </nt>
  <nt id="s42_508" cat="S">
    <edge label="SBJ" idref="s42_500" />
    <edge label="HD" idref="s42_2" />
    <edge label="COMP" idref="s42_507" />
  </nt> ...
</nonterminals>
(TüBa-D/S, cd15, sentence 22)
```

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. 'JJ=adjective'.

• A description as detailed as possible, of how the symbols are applied to text sentences, e.g.,

» How do annotators recognise 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 maxims:

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-Likelihood 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
- testbed for the evaluation of NLP programs

Annotation: References

Geoffrey Leech (1997): "Introducing Corpus Annotation", In: R. Garside, G. Leech & T. McEnery, editors: "Corpus Annotation". Longman, London, New York, 1-18.

Geoffrey Leech and Elisabeth Eyes (1997): "Syntactic Annotation: Treebanks". In: R. Garside, G. Leech & T. McEnery, editors: "Corpus Annotation". Longman, London, New York, 34-52.

Tony McEnery (2003): "Corpus Linguistics". In: Rusan Mitkov, editor: "The Oxford Handbook of Computational Linguistics: 448-463.

CHAT:
Brian MacWhinney (2000): "The CHILDES Project: Tool for Analyzing Talk". 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates.

TAGGIT tagger:
Green, B. and Rubín, G. (1971): "Automated Grammatical Tagging of English". Department of Linguistics, Brown University.

Annotation: References

Online resources:

CHILDES (Child Language Data Exchange System)
<http://childes.psy.cmu.edu/>

CLAWS tagger
<http://www.comp.lancs.ac.uk/ucrel/claws/>

LOB Corpus (Lancaster-Oslo/Bergen Corpus)
http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html

Metadata

A corpus contains different kind 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

(Meta-)Metadata: CES header

<cesHeader> "electronic title page" prefixed to every text, or to the corpus as a whole.

type CORPUS TEXT*

creator of the header

version of the header

status revision status of the header: NEW* UPDATE

date.created date on which the header content was created.

date.updated

<fileDesc> full bibliographic description

<encodingDesc> relationship between an electronic text and the source or sources from which it was derived.

<profileDesc> various aspects of a text, e.g. language used, the situation and date of its production, the participants and their setting, and a descriptive classification for it.

<revisionDesc> summarizes the revision history for a file.

CES-conform TUSNELDA header

```
<tusneldaheader version="1.0" date.updated="06.09.2001"
date.created="06.09.2001" creator="SFB441/project B8" type="corpus">
<filedesc>
<titlestmt>
<h.title>Novosadski korpus of Spoken Language 1980/2000</h.title>
<respstmt>
<resptype>annotation</resptype>
<respname>Gabi Fulir/Slavica Stevanovi&ccacute;</respname>
</respstmt>
</titlestmt>
<editionstmt version="1"></editionstmt>
<extent>
<wordcount>25037</wordcount>
<tokencount>30145</tokencount>
<charactercount>109040</charactercount>
<bytecount units="kb">389.1</bytecount>
</extent>
.....
</tusneldaheader>
```

CES-conform TUSNELDA header

```
<tusneldaheader version="1.0" date.updated="06.09.2001"
date.created="06.09.2001" creator="SFB441/project B8" type="corpus">
<filedesc>
<titlestmt>
<h.title>Novosadski korpus of Spoken Language 1980/2000</h.title>
<respstmt>
<resptype>annotation</resptype>
<respname>Gabi Fulir/Slavica Stevanovi&ccacute;</respname>
</respstmt>
</titlestmt>
<editionstmt version="1"></editionstmt>
<extent>
<wordcount>25037</wordcount>
<tokencount>30145</tokencount>
<charactercount>109040</charactercount>
<bytecount units="kb">389.1</bytecount>
</extent>
.....
</tusneldaheader>
```

Standardisation

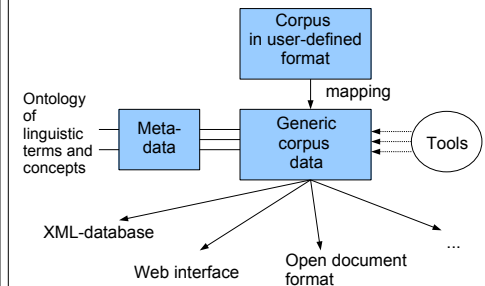
Motivation

- 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

Objections

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

Standardisation: Architecture



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 (IMDI)
 - Open Language Archive Community (OLAC)

Standardisation

Generic corpus data

- **Content specification:** tags and attributes
 - Text Encoding Initiative (TEI)
 - XML Corpus Encoding Standard (XCES, initiated by EAGLES)
- **Data model:** logical structure of linguistic annotation
- **Hierarchical:** syntactic annotation in treebanks
 - TUSNELDA
- **Timeline-based:** phonetic transcription and further annotation
 - Annotation graphs
- **Combined**
 - Nite Object Model (NOM)

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

- **Content:** information content of the resource
 - Coverage: unbalanced, low recording quality
 - Terminology: ambiguous / unknown terms
- **Format:** electronic representation
 - Openness: proprietary format often restricted to specific hardware / operating system
 - Encoding: idiosyncratic representation of characters
 - Markup: idiosyncratic, instable representation of information about character string: e.g. font changes, punctuation

chien n dog.
chien: [n] dog.

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

- Steven Bird & Gary Simons (2003) "Seven dimensions of portability for language documentation and description". *Language* 79(3):557-582.
- Steven Bird & Mark Liberman (2001). "A formal framework for linguistic annotation". *Speech Communication* 33(1,2): 23-60.
- Burnard, Lou (2005) "Metadata for Corpus Work" In: Martin Wynne, editor: "Developing Linguistic Corpora: a Guide to Good Practice". Oxford: Oxbow Books: 30-46. Available online from <http://ahds.ac.uk/linguistic-corpora/> [Accessed 2006-07-26].
- LAF:
Nancy Ide, and Laurent Romary (2004) "International standard for a linguistic annotation framework" In *Journal of Natural Language Engineering*, 10.3-4, 211-225.
- Sustainability project:
Christian Chiarcos, Erhard Hinrichs, Timm Lehmborg, Georg Rehm, Thomas Schmidt, and Andreas Witt. to appear. Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. In Paper accepted at the E-MELD workshop 2006, Ypsilanti.

Metadata, Representation, Sustainability: References

Online resources

- CES: <http://cs.vassar.edu/CES/>
- XCES: <http://www.xml-ces.org/>
- DOLCE: <http://www.loa-cnr.it/DOLCE.html>
- EAGLES: <http://www.ilc.cnr.it/EAGLES/home.html>
- E-MELD: <http://emeld.org>
- IMDI: <http://www.mpi.nl/IMDI>
- ISO/TC37/SC4: <http://www.tc37sc4.org>
- Linguistic Data Consortium (LDC): <http://www ldc.upenn.edu>
- NITE: <http://nite.nis.sdu.dk>
- OLAC: <http://www.language-archives.org/>
- TEI: <http://www.tei-c.org/> (chap 5, chap 23)
- TUSNELDA: <http://www.sfb441.uni-tuebingen.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 V

| tokens | types | types: case | types: case & lemma |
|------------|------------|-------------|---------------------|
| Across | Across | across | across |
| the | the | the | the |
| bridge | bridge | bridge | bridge |
| Wayne | Wayne | Wayne | Wayne |
| Bridge | Bridge | | |
| saw | saw | saw | see |
| the | | | |
| stadium | stadium | stadium | stadium |
| , | | | |
| The | The | | |
| supporters | supporters | supporters | supporter |
| could | could | could | can |
| also | also | also | also |
| see | see | see | |
| him | him | him | he |
| , | | | |
| 16 (14) | 13 | 11 | 10 |

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?
 - » lemmatise 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

frequency list

| type | f | type | f |
|--------|---|-----------|---|
| across | 1 | see | 2 |
| also | 1 | supporter | 1 |
| bridge | 2 | the | 3 |
| can | 1 | Wayne | 1 |
| he | 1 | | |

rank / frequency profile

| r | f | r | f |
|---|---|---|---|
| 1 | 3 | 6 | 1 |
| 2 | 2 | 7 | 1 |
| 3 | 2 | 8 | 1 |
| 4 | 1 | 9 | 1 |
| 5 | 1 | | |

frequency spectrum

| f | V(f) |
|---|------|
| 1 | 6 |
| 2 | 2 |
| 3 | 1 |

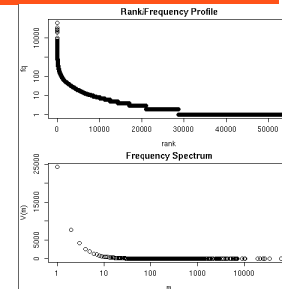
Frequencies of Brown Corpus

| top frequencies | | | bottom frequencies | | | |
|-----------------|----------------|------|--------------------|-------|----------------|---------------------------------|
| rank | f _q | word | rank | range | f _q | randomly selected examples |
| 1 | 62642 | the | 7967-8522 | 10 | | recordings undergone privileges |
| 2 | 35971 | of | 8523-9236 | 9 | | Leonard indulge creativity |
| 3 | 27831 | and | 9237-10042 | 8 | | unnatural Lolotte authenticity |
| 4 | 25608 | to | 10043-11185 | 7 | | diffraction Augusta postpone |
| 5 | 21883 | a | 11186-12510 | 6 | | uniformly throttle agglutinin |
| 6 | 19474 | in | 12511-14369 | 5 | | Bud Councilman immoral |
| 7 | 10292 | that | 14370-16938 | 4 | | verification gleamed groin |
| 8 | 10026 | is | 16939-21076 | 3 | | Princes nonspecifically Arger |
| 9 | 9887 | was | 21077-28701 | 2 | | blitz pertinence arson |
| 10 | 8811 | for | 28702-53076 | 1 | | Salaries Evensen parentheses |

Table 4: Top and bottom of the Brown frequency list

(Baroni, prefinal: 5)

Frequencies of Brown Corpus



(Baroni, prefinal: 7)

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 C = frequency of most frequent word
 $r(w)$ = rank of word w a = a constant

Zipf's law

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

Given $C = 60,000$ and $a = 1$:

| $r(w)$ | $f(w)$ | $r(w)$ | $f(w)$ |
|--------|-----------------------|--------|-------------------------|
| 2 | $60,000 / 2 = 30,000$ | 100 | $60,000 / 100 = 600.00$ |
| 3 | $60,000 / 3 = 20,000$ | 101 | $60,000 / 101 = 594.06$ |
| ... | ... | 102 | $60,000 / 102 = 588.23$ |

→ about 80,000 words have $f(w)$ between 1.5 and 0.5

Frequency: References

- Marco Baroni (to appear) "Distributions in text". In: Anke Lüdeling and Merja Kytö, editors, "Corpus linguistics: An International Handbook", Berlin: Mouton de Gruyter.
 - George Kingsley Zipf (1949) "Human behaviour and the principle of least effort". Cambridge MA: Addison-Wesley.
 - George Kingsley Zipf (1965) "The pseudo-biology of language". Cambridge MA: MIT Press.
- [Introduction to lexical statistics:](#)
- Harald Baayen (2001) "Word frequency distributions" Dordrecht:Kluwer