

# **Introduction to Corpus Resources, Annotation and Access:**

## ***Introduction***

---

Sabine Schulte im Walde  
Universität Stuttgart

*Foundational Course*  
Departament de Traducció i Filologia  
Universitat Pompeu Fabra  
April 16-20, 2007

# Organisation

---

## Introduction to Corpus Resources, Annotation and Access

Sabine Schulte im Walde, Universität Stuttgart

- Lectures plus exercises
- Course homepage:  
<http://www.ims.uni-stuttgart.de/~schulte/Teaching/UPF-07/>
- Exercises and material:  
<http://www.ims.uni-stuttgart.de/~schulte/Teaching/UPF-07/ex.html>
- References:  
<http://www.ims.uni-stuttgart.de/~schulte/Teaching/UPF-07/ref.html>

# Schedule

---

- Introduction: Corpora and Annotation
- Tokenisation
- *Basic Unix Tools and Corpus Frequencies*
- Morpho-Syntactic Annotation
- Word Distributions
- *Tree Tagger and Corpus Query Processor*
- Syntactic Annotation
- *Searching Treebanks with TIGERSearch*
- Semantic Annotation
- *Semantic Annotation with SALTO*
- More Levels of Corpus Annotation
- Evaluation

# Acknowledgements

---

Most of the course material was adopted from an earlier version of this course at the European Summer School in Logic, Language and Information (**ESSLLI 2006**).

Thanks to my colleague **Heike Zinsmeister** who prepared the introduction and the lectures on syntactic annotation and the web as corpus!

# 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, 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 databases and analyser, e.g. UPenn's XTAG
- Semantic hierarchies, e.g. WordNet
- Annotation tools, e.g. TreeTagger, Stanford Parser
- Processing tools, e.g. UPenn's tgrep

**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
- 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** sentences
  - » 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.

# 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.)
- See: Verbmobil corpus, TüBa-DS Treebank.

# Design Criteria: Example

---

- British National Corpus ([reference corpus](#))
- 100 million words
- 90 % **written** language
  - time of creation: 1960-1974, 1975-1993
  - medium: book, newspaper, other publs., 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...)

# Corpus Typology: Contrastive Parameters

---

1. Full text; sample; monitor
2. Closed; open-ended
3. Synchronic; diachronic
4. General; terminological
  - » reference corpus versus special corpus
5. Monolingual; bilingual; plurilingual/multilingual
  - » parallel corpus, comparable corpus
6. Language(s) of corpus
7. (....)
9. Core; periphery

# Corpus Typology: Text Attributes

---

1. mode: written, written-to-be-read, written-to-be-spoken, spoken, spoken-to-be-written
2. text origin: single, several, joint, ...
3. medium: book, newspaper, classroom lessons

# 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
  - » automatic annotation often requires context information  
(against 'on the fly' annotation).
  - » annotation is time-consuming and expensive.
- Multi-functionality
  - » same corpus used for e.g. lexicography, speech synthesis, machine-aided translation, information retrieval.

# Annotation Levels

---

- Part-of-speech tags
- Lemmata
- Syntactic functions
- Senses
- Semantic roles
- Prosody
- Topic / Focus
- Coreference
- Named Entities
- Discourse relations
- Time
- Emotions ...

# Annotation: Principles of Good Practice

---

1. The raw corpus 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: Principles of Good Practice

---

Additional maxims:

4. 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.')
5. Annotation should depend on consensual or theory-neutral analyses.
6. No annotation scheme should claim authority as an absolute standard.

(adapted from Leech 1997:6)

# 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

# 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: tagger CLAWS
  - » 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 words/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 context (e.g. LOB Corpus) → accuracy of **~97 %**.

# 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

# 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

# From *Raw* to *Annotated* Corpus Text

---

Dunkel\swar's,\sder\sMond\sschien\shelle,\n\nschneebedeckt\sdie\sgrüne\  
sFlur,\n\nals\sein\sAuto,\sblitzeschnelle,\n\nlangsam\sum\sdie\sEcke\sfuhr.\n\nDrinne\n\nssaßen\sstehend\sLeute,\n\nschweigend\sins\sGespräch\svertieft,\n\nals\sein\stotgeschoss'ner\sHase\n\nauf\lder\sSandbank\sSchlittschuh\slie  
f.\n\nUnd\sauf\s'ner\sgrünen\sBank,\n\nindie\srot\sangestrichen\swar,\n\nsaß\sei  
n\sblondgelockter\sJüngling\n\nmit\skohlrabenschwarzem\sHaar.\n\n...

# From *Raw* to *Annotated* Corpus Text

---

Dunkel war's, der Mond schien helle,  
schneebedeckt die grüne Flur,  
als ein Auto, blitzschnelle,  
langsam um die Ecke fuhr.  
Drinne saßen stehend Leute,  
schweigend ins Gespräch vertieft,  
als ein totgeschoss'ner Hase  
auf der Sandbank Schlittschuh lief.  
Und auf 'ner grünen Bank,  
die rot angestrichen war,  
saß ein blondgelockter Jüngling  
mit kohlrabenschwarzem Haar.

...

1. Tokenisation
2. Prosody
3. Semantics

# Corpora and Annotation: References

---

- Tony McEnery and Andrew Wilson (1996): „Corpus Linguistics“. Edinburgh University Press.
- Geoffrey Leech (1997): "Introducing Corpus Annotation", In: R. Garside, G. Leech & Tony McEnery, editors: "*Corpus Annotation*". London, New York: Longman, 1-18.
- Steven Bird and Gary Simons (2003): „Seven dimensions of portability for language documentation and description.“ *Language* 79: 557-582.
- Tony McEnery (2003): "Corpus Linguistics". In: *The Oxford Handbook of Computational Linguistics*: 448-463.
- Nancy Ide (2004): „Preparation and analysis of linguistic corpora.“ In: Susan Schreibman, Ray Siemens, John Unsworth, editors: „*A Companion to Digital Humanities*.“ Blackwell.