

Introduction to Corpus-based Computational Semantics: *Motivation and Overview*

Alexander Koller
Columbia University

Sabine Schulte im Walde
University of Stuttgart

Introductory Course
ESSLLI 2007
Trinity College, Dublin, Ireland
August 6-10, 2007

Organisation

Introduction to Corpus-based Computational Semantics

Alexander Koller, Columbia University

Sabine Schulte im Walde, University of Stuttgart

- Course homepage:
<http://www.ims.uni-stuttgart.de/~schulte/Teaching/ESSLLI-07/>
- Reader:
<http://www.ims.uni-stuttgart.de/~schulte/Teaching/ESSLLI-07/reader.pdf>
- Further references:
<http://www.ims.uni-stuttgart.de/~schulte/Teaching/ESSLLI-07/ref.html>

Motivation



Lexical Acquisition and Lexical Resources

- **Lexical acquisition**: (automatic) definition of **linguistic information** on lexical items
- Purposes:
 - » linguistic theory
 - » lexical resources
 - » NLP tools and applications
- Sources for lexical acquisition:
 - » (hand-coded) **language resources**, such as dictionaries, thesauri, taxonomies, ontologies, etc.
 - » (annotated) **corpus data**

Corpus-based Computational Semantics

- Focus of interest: semantics
- **Corpora** → **computational semantics**:
induce semantic information from corpus data
- **Computational semantics** → **corpora**:
enrich corpus data with semantic information

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)

Corpora as Language Sample

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
 - they include **ungrammatical** sentences
 - » typos, copy-and-past errors, conversion errors, etc.
- Sample a corpus according to design criteria such that it is **balanced** and **representative** for a specific purpose.

Properties of Corpora (examples)

- *Language type*: spoken, written, usegroups, etc.
- *Language and alignment*:
monolingual, bilingual, multilingual;
parallel vs. comparable
- *Text type*: newspaper, journal, belletristic, poetry, ...
- *Domain*: finance, religion, sports, computer, children, ...
- *Homogeneity*: homogeneous vs. heterogeneous/balanced
- *Time period* and *age*
- *Size*

Corpus Annotation

- Practice of adding **interpretative, linguistic information** to an electronic corpus.
- End-product: linguistic symbols are attached to, linked with, interspersed with the electronic representation of the language material.
- **Levels of annotation**: token, part-of-speech, lemmata, syntactic functions, word senses, semantic roles, time, prosody, topic/focus, discourse relations, emotions, ...
- **Levels of granularity**: how much detail should be encoded through annotation?

Empirical Approach to Linguistics

- Goal: *characterisation and explanation of linguistic observations*
- Competing approaches: **rationalism** vs. **empiricism**
theory-based vs. data-based
competence vs. performance
- Describing **naturally occurring** language data
- **Objective** (reproducible) statements about language
- **Quantitative analysis**: common patterns in language use

Lexical Semantic Acquisition

- *Semantic bottleneck*: few, expensive resources
- Difficulty of the task: agreement on annotation decreases with the complexity and the controversy of the task
- Acquisition of lexical semantic information?
 - » exploit and combine available resources, e.g., WordNet, PropBank, dictionaries, thesauri
 - » exploit linguistic knowledge, e.g., syntax-semantics interface
 - » apply heuristics and statistics to corpus data, e.g., co-occurrence patterns

Lexical Semantic Acquisition: Why bother?

- **Word senses** in machine translation:
The judged anounced a hard sentence. →
*Der Richter verkündete einen harten *Satz.
ein hartes Urteil.*
- **World knowledge:**
Is there any fresh milk left? —
The fridge smelled awful in the morning.
- **Applications that require semantic information:**
machine translation, question answering, information
extraction, information retrieval, summarisation, etc.

Corpus-based Computational Semantics

- Resources and corpus-based approaches in computational lexical semantics
 - » **theoretical issues:**
word senses, polysemy,
semantic similarity and semantic relations, etc.
 - » **available resources:**
WordNet, dictionaries, thesauri, etc.
 - » **computational approaches** to lexical semantic acquisition

Course Overview



Course Topics and Schedule

- | | | |
|--|--------------------------|----------------------------|
| 1. Motivation | <i>Mon</i> | <i>SiW</i> |
| 2. Word senses | <i>Mon</i> | <i>SiW</i> |
| 3. Semantic similarity, relations, classes | <i>Tue</i>
<i>Wed</i> | <i>SiW/AK</i>
<i>AK</i> |
| 4. Selectional preferences, semantic roles | <i>Thu</i> | <i>SiW</i> |
| 5. World knowledge and the limit of corpus-based semantics | <i>Fri</i> | <i>AK</i> |
| 6. Summary and closing | <i>Fri</i> | <i>AK</i> |

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.

Acknowledgements

Some of the course material in the corpus introduction was adopted from course material in an ESLLI 2006 course, as prepared by our colleague **Heike Zinsmeister**, University of Tübingen, Germany.