# Introduction to Computational Linguistics

## Lexical Resources

Jan-Philipp Söhn

jp.soehn@uni-tuebingen.de

November 14th, 2007

# Introduction

A lexical entry...

# Introduction

What is missing:

- appropriate information about superordinates
- no information about coordinate items
- no information about different kinds of ...
- a lexicon is not a catalogue of general knowledge

# Introduction

What is missing:

- appropriate information about superordinates
- no information about coordinate items
- no information about different kinds of ...
- a lexicon is not a catalogue of general knowledge

# Introduction

What is missing:

- appropriate information about superordinates
- no information about coordinate items
- no information about different kinds of ...
- a lexicon is not a catalogue of general knowledge

# Introduction

What is missing:

- appropriate information about superordinates
- no information about coordinate items
- no information about different kinds of ...
- a lexicon is not a catalogue of general knowledge

# Today's Topic

- computers can be more than rapid pageturners!
- electronic lexicons are not an NLP application but a resource
- focus on syncronic organization of the lexicon
- DWDS as example of an enhanced electronic lexicon
- Princeton WordNet and GermaNet as examples for lexical-semantic wordnets

# DWDS

Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts
Berlin-Brandenburgische Akademie der Wissenschaften

see http://www.dwds.de/

# History of Wordnets

- psycholinguistics
- research on lexical component of language (word association studies) $\longrightarrow$ psycholexicology
- syncronic properties of the mental lexicon
- WordNet = dictionary based on psycholinguistic principles

# Some Basics

- Difference signifiant / signifié (de Saussure)
- different parts-of-speech: noun, verb, adjectives, adverbs, function words
- nouns are organized as topical hierarchies, verbs by entailment relations
- meanings or concepts are represented as sets of synonyms (synsets): {board, plank} and {board, committee}
- wordnet = hierarchy of concepts with relations to other concepts
- wordnets strive for completeness (WordNet: 117,660 synsets; 155,290 strings)

# Psycholinguistic Assumptions

- evidence from aphasic patients
- superordinates can serve as anaphors
- discussion about lexical memory as an inheritance system

# Distinguishing Features (1)

- Evidence for a "basic level"

- living thing – animal – mammal – lion – male lion
- BL priviledged (short response times, similar overall shape, not composite, frequent usage)

# Distinguishing Features (1)

- Evidence for a "basic level"

- living thing – animal – mammal – lion – male lion
- BL priviledged (short response times, similar overall shape, not composite, frequent usage)

# Distinguishing Features (1)

- Evidence for a "basic level"

- living thing – animal – mammal – lion – male lion
- BL priviledged (short response times, similar overall shape, not composite, frequent usage)

# Distinguishing Features (2)

- WordNet was originally designed to be used in conjunction with some online dictionary
- Coverage increased and one cannot distinguish every meaning with help of synonyms
- Thus, distinguishing features were added

# Relations between Words

- lexical relations: bidirectional relations between word meanings
  - synset-internal synonymy (end of a continuum of similarity)
  - antonymy (other end of the continuum)
- conceptual relations:
  - hyponymy, hyperonymy
  - meronymy, holonymy
  - antonymy
  - entailment
  - causation

# Exercise and Demo

# Differences between GermaNet and WordNet (1)

- GermaNet and WordNet share the same technology
- GermaNet followed linguistic design principles as opposed to Princeton's psychological motivation
- GermaNet uses artificial, i. e. non-lexicalized concepts, which have been introduced to balance the taxonomical structure more adequately
- in GermaNet, adjectives are ordered hierarchically as opposed to Princeton's grouping by the satellite approach

# Differences between GermaNet and WordNet (2)

- uniform treatment of meronymy within GermaNet, whereas WordNet has established three different pointers for Part, Member and Substance
- within GermaNet, the causation relation can be encoded between all parts of speech, not only between verbs and adjectives
- due to emphasizing the syntax-semantics-interface for disambiguation tasks we accounted for some sixty verbal subcategorisation frames. These frames are more elaborate than the WordNet frames, and, furthermore, for each verb reading we provide a typical example.

# Some more details about GermaNet

Connectivity between word classes will be improved in GermaNet by

- using existing cross-class relations ('pertains to', 'has/is attribute') more frequently
- allowing certain relations to cross word classes (verbs are allowed to 'cause to' adjectives)
- adding new cross-class relations (for selectional restrictions)

# EuroWordNet (1)

- GermaNet integrated its basic vocabulary (15,000 synsets) into EuroWordNet
- EWN is a polylingual resource comprising English, Spanish, French, Italian, German, Dutch, Czech, Estonian
- common set of "base concepts" (1000 nouns, 300 verbs) – central vocabulary
- base concepts are the (immediate) hypernyms of base level words (hammer – tool)
- international collaboration led to the founding of Global Wordnet Association in 2000

# EuroWordNet (2)

- Components of EWN:
  - Interlingual Index (unstructured list which based on WordNet synsets) which connects the concepts in the individual languages (drive–conducir–conduire–guidare–fahren–...)
  - Top ontology (semantic features, e.g. motion)
  - Domain ontology (semantic fields, e.g. traffic)

# Applications for Wordnets (1)

Word sense disambiguation

- polysemous words belong to different synsets
  {ton, laut} vs. {ton, tonerde}
- SENSEVAL since 1998

# Applications for Wordnets (2)

Information retrieval

- *Get this information out of that collection of texts!*
- semantic nets help to disambiguate query and texts and to extend the query with correlated concepts
- EuroWordNet even allows to extend the query with correlated concepts in other languages
- the recall increases

# Applications for Wordnets (3)

Semantic annotation

- semantic indexing of syntactically analyzed texts
- establish a standard input-output-format for several NLP applications such as speech recognition, parsing, and speech generation