

Text Mining for Historical Documents /  
Historische Quellen mit (computer-)linguistischen  
Methoden analysieren  
Introduction to Computational Linguistics

Caroline Sporleder and Martin Schreiber

Computational Linguistics &  
Kultur- und Mediengeschichte  
Universität des Saarlandes

Wintersemester 2009/10

22.02.2010

# What is Computational Linguistics?

## Computational Linguistics (CL) ...

“...is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty. It belongs to the cognitive sciences and overlaps with the field of artificial intelligence (AI), a branch of computer science aiming at computational models of human cognition.”

Source:

[http://www.coli.uni-saarland.de/~hansu/what\\_is\\_cl.html](http://www.coli.uni-saarland.de/~hansu/what_is_cl.html)  
(Hans Uszkoreit)

For our purposes: basically processing human/natural language with a computer (“Natural Language Processing”, NLP)

## Some Terminology

**Phonology (Phonetics):** the study of speech sounds

- **phoneme (phon):** the smallest meaning-distinguishing unit of language, e.g.  
/cat/ vs. /cut/ ⇒ “a” and “u” are phonemes
- cf. **grapheme:** smallest unit in written language, e.g. a letter (Buchstabe)
- **phoneme to grapheme conversion:** mapping phonemes to graphemes, e.g. in speech recognition

⇒ important for text-mining of audio archives

**Morphology:** the study of word structure

- **morpheme:** the smallest meaning-carrying unit of language, e.g.  
*reachable* ⇒ *reach* and *able* are morphemes
- **root:** the important bit of the word, e.g. *reach*
- **affix:** the less important stuff, e.g. *-able*  
affixes are divided into **prefixes** (stuff that comes before the root, like *mis-* in *misrepresent* (or *misunderestimate* ;-)) and **suffixes** (stuff that comes after the root, like *-able*)

⇒ important for methods dealing with non-standard orthography

**Lexicography (Lexicology):** the study of the words of a language

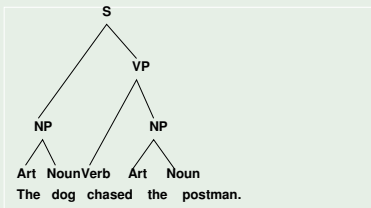
- **lexeme:** elementary unit in lexicography, “go”, “goes”, “gone” are different words but the same lexeme
- **lemma:** the base (dictionary) form of a word
- **lemmatisation:** mapping word forms to their lemmas, important for further steps of automatic analysis
- **part-of-speech:** (=Wortart), e.g., noun (Nomen, Substantiv), verb (Tu-Wort), adjective (Wie-Wort) etc.
- **part-of-speech tagging (pos tagging):** the process of automatically assigning a part-of-speech tag to a word

# Some Linguistic Terminology (4)

**Syntax:** the study of the internal (grammatical) structure of a sentence

- **syntax tree or parse tree:** an abstract representation of the internal structure of a sentence (as determined by a grammar)
- **parsing:** the process of computing sentence structure automatically
- **parser:** a tool which does parsing

## Parse tree



**Semantics:** the study of language meaning

- **word sense:** a word like *bank* has several word senses
- **word sense disambiguation:** the process of determining the word sense of a word
- **hypernym:** *flower* is a hypernym of *rose*, *animal* is a hypernym of *cat*
- **hyponym:** the inverse, i.e. *cat* is a hyponym of *animal*
- also semantic argument structure (who did what to whom?)

⇒ important for ontology construction etc.



# Automatic Text Processing

## Original Text

(Amtspresse Preußens, 1.7.1863)

Die Nachrichten aus Karlsbad über das Befinden unseres Königs lauten sehr erfreulich. Die begonnene Brunnenkur scheint dem hohen Herrn sehr wohl zu thun. Der Präsident des Staatsministeriums, Herr von Bismarck, mit welchem Se. Majestät täglich eine Zeit lang gearbeitet, hat Karlsbad jetzt wieder verlassen.

## Original Text

(Amtspressen Preußens, 1.7.1863)

Die Nachrichten aus Karlsbad über das Befinden unseres Königs lauten sehr erfreulich. Die begonnene Brunnenkur scheint dem hohen Herrn sehr wohl zu thun. Der Präsident des Staatsministeriums, Herr von Bismarck, mit welchem Se. Majestät täglich eine Zeit lang gearbeitet, hat Karlsbad jetzt wieder verlassen.

## Step 1: Tokenisation

- Where are the words in the text? What are the non-word components (punctuation etc.)?
- Where are the sentence boundaries? (sentence splitting)

# Tokenisation, isn't that easy?

## Simple solution

- words are delimited by spaces
- sentences are delimited by “.”, “!”, “?”

# Tokenisation, isn't that easy?

## Simple solution

- words are delimited by spaces
- sentences are delimited by “.”, “!”, “?”

## Yes, but ...

- ... where's the sentence boundary in:  
*Neil Budde, general manager of Yahoo! News, said: "Our expanded news search dramatically increases the consumer's ability to find events that matter to them."*
- ... how many words does *17.2.2009* consist of? What about *3.5 billion euros* ?

Tokenised

(Amtspresse Preußens, 1.7.1863)

Die Nachrichten aus Karlsbad über das Befinden unseres Königs lauten sehr erfreulich .

Die begonnene Brunnenkur scheint dem hohen Herrn sehr wohl zu thun .

Der Präsident des Staatsministeriums , Herr von Bismarck , mit welchem Se. Majestät täglich eine Zeit lang gearbeitet , hat Karlsbad jetzt wieder verlassen .

Tokenised

(Amtspresse Preußens, 1.7.1863)

Die Nachrichten aus Karlsbad über das Befinden unseres Königs lauten sehr erfreulich .

Die begonnene Brunnenkur scheint dem hohen Herrn sehr wohl zu thun .

Der Präsident des Staatsministeriums , Herr von Bismarck , mit welchem Se. Majestät täglich eine Zeit lang gearbeitet , hat Karlsbad jetzt wieder verlassen .

**Step 2:** Part-of-Speech Tagging (=Wortarten zuweisen)

- Which parts-of-speech do the words in the text have?

# Part-of-Speech Tagging, isn't that easy?

## Simple solution (if you have a dictionary)

- look up the words in a dictionary, e.g. “corner” ⇒ noun, “man” ⇒ noun, “wins” ⇒ verb, “spell” ⇒ verb



# Part-of-Speech Tagging, isn't that easy?

## Simple solution (if you have a dictionary)

- look up the words in a dictionary, e.g. “corner” ⇒ noun, “man” ⇒ noun, “wins” ⇒ verb, “spell” ⇒ verb

## Yes, but what about . . .

- *Maybe the hunters can **corner** the tiger.*
- *Steward Crowe waited on the port side until he was told to **man** the boat.*
- *Tiger Woods makes it seven **wins** in a row.*
- *Readers are still under the **spell** of Harry Potter.*

POS Tagged

(Amtspresse Preußens, 1.7.1863)

Die\_DET Nachrichten\_n aus\_prep Karlsbad\_n über\_prep das\_det  
Befinden\_n unseres\_pro Königs\_n lauten\_v sehr\_adv erfreulich\_adj  
..punct

...

POS Tagged

(Amtspresse Preußens, 1.7.1863)

Die\_DET Nachrichten\_n aus\_prep Karlsbad\_n über\_prep das\_det  
Befinden\_n unseres\_pro Königs\_n lauten\_v sehr\_adv erfreulich\_adj  
..punct  
...

## Step 3:

- what is the syntactic structure of the sentence?

# Parsing, ok this shouldn't be too difficult, should it?

## Solution

- apply your grammar rules to the sentence

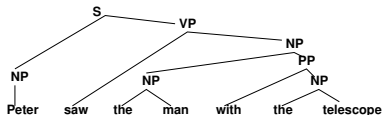
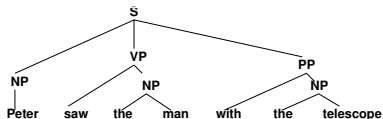
# Parsing, ok this shouldn't be too difficult, should it?

## Solution

- apply your grammar rules to the sentence

Yes, but what about ...

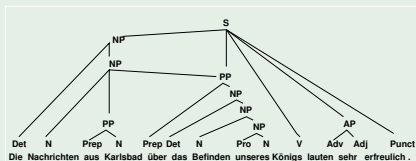
*Peter saw the man with the telescope.*



# The King on a Wellness Holiday

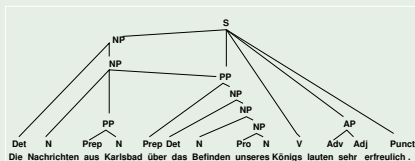
Parsed

(Amtspressen Preußens, 1.7.1863)



Parsed

(Amtspressen Preußens, 1.7.1863)



## Step 4: Semantic Analysis

- who did what where and when to whom?

# Semantic Analysis, how difficult is it?

## Solution

- build on the syntactic structure
- identify the subject, e.g. “Bismarck” in “Bismarck hat Karlsberg verlassen.”
- subject=Agent (the entity doing something)
- object=Patient (the entity to which something is done, e.g. “Karlsbad”)



# Semantic Analysis, how difficult is it?

## Solution

- build on the syntactic structure
- identify the subject, e.g. “Bismarck” in “Bismarck hat Karlsberg verlassen.”
- subject=Agent (the entity doing something)
- object=Patient (the entity to which something is done, e.g. “Karlsbad”)

## Yes, but what about ...

- *Karlsbad wurde von Bismarck verlassen.* (subject=Karlsbad, agent=Bismarck)
- *Bismarcks abrupte Abreise aus Karlsbad ...*

## Named Entity Tagging

- identify person names, locations, dates, numbers etc.

## Pronoun resolution

- Who is “he”?

## Co-reference resolution

- Do “Obama” and “the president” refer to the same person?

# Ok, so how do you do all this?

Basically two possible approaches

- manually defined rules (“if ‘corner’ follows an article like ‘the’ it is a noun”)
- use machine learning and let the program figure it out itself

- a lot of work!
- typically high precision (rules are correct) but low coverage (rules don't cover all possible eventualities)

- also a lot of work: we need manually annotated training data
- typically robust, but not necessarily always correct
- training data can be re-used but only in certain situations (domain and genre should not change), e.g.:
  - can train a system on the Wall Street Journal and apply to the New York Times
  - cannot train a system on *Der Zauberberg* and apply it to the *Amtsprelle Preußens*

When dealing with cultural heritage data this is a challenge because annotation of large amounts of data for each text type is infeasible.

⇒ need to think creatively (e.g. domain adaptation methods)