MGK Seminar:

Corpora in Empirical Research
(Preprocessing and Word Statistics) +

Lexical Semantics
(Tasks, Approaches and Ressources)

Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung (IMS)
Universität Stuttgart

May 8, 2015

Outline

- Corpora in empirical research
  - Corpora and annotation (reminder)
  - Preprocessing corpora
    - Tokenisation
    - Morphology
    - Part-of-speech tagging
  - Word statistics
- Lexical semantics
  - Empirical tasks and approaches
  - Resources
    - WordNet
    - FrameNet
- Lexical semantics in empirical research: Do it yourself!

Outline

Outline

Corpora in Empirical Linguistic Research

- Corpus: collection of texts
- Corpora describe naturally occurring language data.
- Corpora are the basis for empirical research in theoretical linguistics.
- Corpora allow objective (reproducible) statements about language.
- Corpora give only a partial description of a language:
  - incomplete
  - biased
  - include ungrammatical sentences

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?
- Annotation is expensive.

Corpus Evidence

- Exploration of corpora:
  - search for evidence
  - generalise over evidence
- Evidence: occurrence of sounds, characters, strings, etc.
- Quantitative analyses via "patterns" of (co-)occurrences, e.g.
  - association strength between words:
    kick the bucket; eat chocolate
  - semantic relation between words:
    flowers such as roses and tulips

Outline

Outline

Tokenisation

- Tokenisation divides the raw input character sequence of a text into sentences and the sentences into tokens.
- What is a token?
  - *words*: time / as / 40. / House / runs . . .
  - *punctuation*: ! " ,) . . .
- Simple tokeniser: Split the character sequence at whitespace positions and cut off punctuation, to obtain the token sequence.
- Problem: ambiguities, mainly caused by periods
- Errors made at this stage are very likely to cause more errors at later stages (morphology, syntax, etc.).

Tokenisation: Problems

- Major problem categories:
  - disambiguation of sentence boundaries
  - normalisation of capitalised words
  - identification of abbreviations
  - identification of multi-word expressions

- Language-dependent task:
  - Each language has different patterns.
  - The language families alphabetic vs. ideographic differ strongly. Ideographic languages provide less information (on punctuation, spaces, etc.).

- Common problem: disambiguation of periods

Tokenisation: Sentence Boundaries

- A period, an exclamation mark, or a question mark usually signals a sentence boundary.
- Other functions of periods:
  - decimal point
  - part of an abbreviation
  - end-of-sentence indicator (full-stop) and at the same time part of an abbreviation
- Examples:
  Anna went home late . Her father was angry .
  Anna came back from the U . S . A . last month .
  Anna came back from the U . S . A . She enjoyed it .
  Anna came back from the U . S . A . Continental ...

Tokenisation: Multi-Word Expressions

- Assumption: Tokens do not contain whitespace.
- Problem: Multiword expressions contain whitespace. Do they represent one or several tokens?
- Examples:
  - *Feb. 1, 2004*
  - *Daimler Chrysler AG*
  - *because of*
- For some applications it is advantageous to treat multiword expressions as a single token.

Tokenisation: Disambiguation

- Heuristics and information sources:
  - Dictionary information
  - Abbreviation lists (manual/automatic)
  - Sentence positions
  - ...
- Heuristics-based approaches:
  - Define heuristics about correspondences between a token and a set of classes.
  - Define heuristics as rules and order the rules according to their reliability.
- Classification approaches (supervised/unsupervised):
  Decision trees, neural networks, maximum entropy, etc.

Outline

Morphology

- Morphology is concerned with the inner structure of words and the formation of words from smaller units.

- The part of the word (morpheme) that carries the central meaning is called the root.

- How much and what sort of information is expressed by morphology differs widely between languages.
  Information that is expressed by syntax in one language is expressed morphologically in another one.

  Examples:

  - future tense in English vs. Spanish: *I will speak – hablaré*
  - Japanese does not mark nouns for plural.

Morphology

- Basic functions of morphology:
  - inflection: change of word form which does not change the part-of-speech category, such as conjugation (*lese, liest, las*)
  - derivation: a new word is produced by adding a morph to the base form, such as verb $\rightarrow$ adjective: *essen* $\rightarrow$ *essbar*
  - compounding: joining of two or more base forms to form a new word, such as *Kaffeefilter*
- Morphotactics: A word grammar determines the way how morphs are put together to form words.
- Morphological Parsing: A word is broken down into its component morphemes by a structured representation.

Morphological Parsing

Input	Morphological Parsed Output
cats	`cat +N +PL`
cat	`cat +N +SG`
cities	`city +N +PL`
geese	`goose +N +PL`
goose	`(goose +N +SG)` or `(goose +V)`
gooses	`goose +V +3SG`
merging	`merge +V +PRES-PART`
caught	`(catch +V +PAST-PART)` or `(catch +V +PAST)`

Computational Morphology

- Task: Take a string of characters or phonemes as input and deliver an analysis of the underlying morphemes or the morphosyntactic interpretation as output.

  incompatibilities
  in+con+patible+ity+s (morphemes)
  incompatibility+Noun+Plural (interpretation)

- Methods:
  - lexicon: full-form or lemmata
  - finite-state morphology: use regular expressions
  - machine learning of morphological structure

Stemming and Lemmatisation

- Stemming: process that strips off affixes and leaves the stem:

  cats, catlike, catty $\rightarrow$ *cat*
  rauchst $\rightarrow$ *rauch*

- Stemming only needs morphological information to determine whether two words have the same stem. Suffixes are thrown away.

- Stemming is sufficient for many applications.

- Lemmatisation: find the lemma or lexeme of the inflected form; includes disambiguation at the level of lexemes, depending on the part-of-speech:

  rauchst $\rightarrow$ *rauchen*

Outline

Part-of-Speech Tagging

- Part-of-Speech Tagging = Tagging
  underspecified (cf. *semantic tagging*) but common usage
- Task of labeling each word in a sequence of words with its appropriate part-of-speech (POS)
- Words are often ambiguous with respect to their POS:
  - *saw* → singular noun vs. past tense of the verb *see*
  - *Dinkelacker* → proper name vs. compound
- Purposes and applications (examples):
  - pre-processing step for morpho-syntactic and further analyses: lemmata, syntactic structure, etc.
  - text indexing, e.g. nouns are more useful than verbs
  - pronounciation in speech processing: *OBject* vs. *obJECT*

Part-of-Speech Tagging

- Tagging performs a limited syntactic disambiguation.
- Tagging accuracy is high (on a per-word basis): 95-98%.
  But: This corresponds to one mistake per 20-word sentence on average.
- Difficulties for English, potentially even for humans (example):
  - distinguish prepositions (IN), particles (RP) and adverbs (RB):

    Mrs./NNP Shaefer/NNP never/RB got/VBD around/**RP** to/TO joining/VBG

    All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB around/**IN** the/DT corner/NN

    Chateau/NNP Petrus/NNP costs/VBZ around/**RB** 250/CD

Part-of-Speech Tagging

- Ambiguity:
  - most English words in English are unambiguous
  - but: many common words are ambiguous (such as *can*)
  - ambiguity in Brown corpus: 11.5% vs. 40% of word types vs. tokens (DeRose, 1988)
- The syntagmatic context helps to disambiguate tags.
- Some POS sequences are common, e.g. `DET ADJ N`.
- Words are associated with dominant POS tags: The distribution of a word's POS tags is extremely uneven.
- Statistical approaches often combine syntagmatic information and lexical information on POS preferences.

Tagsets

- Tagset: set of part-of-speech tags
- The size and choice of the tagsets vary.
- Classical 8 classes (Thrax, 100 BC): noun, verb, article, participle, pronoun, preposition, adverb, conjunction
- Morphologically rich languages (such as German) need more detailed tagsets (such as gender and case).
- Criteria:
  - specifiability: degree to which humans use the tagset uniformly on the same text
  - accuracy: evaluation of output on tagged text
  - suitability for intended application

Penn Treebank Tagset

Penn Treebank Project: syntactic and semantic annotation of naturally-occuring text for linguistic structure;

Tagset:

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg prcs	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	*#*
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

Penn Treebank: Tagging Example

Under/IN
[the/DT proposal/NN]
,/,
[Delmed/NNP]
would/MD issue/VB about/IN
[123.5/CD million/CD additional/JJ Delmed/NNP common/JJ shares/NNS]
to/TO
[Fresenius/NNP]
at/IN
[an/DT average/JJ price/NN]
of/IN about/IN
[65/CD cents/NNS]
[a/DT share/NN]
,/, though/IN under/IN
[no/DT circumstances/NNS]
more/JJR than/IN
[75/CD cents/NNS]
[a/DT share/NN]
./.

STTS-Tagset

- STTS: Stuttgart-Tübingen Tag Set
- De-facto standard for German part-of-speech tagging
- Main word classes:
  1. Nouns (N)
  2. Verbs (V)
  3. Articles (ART)
  4. Adjectives (ADJ)
  5. Pronouns (P)
  6. Cardinals (CARD)
  7. Adverbs (ADV)
  8. Conjunctions (KON)
  9. Adpositions (AP)
  10. Interjections (ITJ)
  11. Particles (PTK)

Tagset Comparison: Penn Treebank vs. STTS

Penn Treebank Tagset (English) - 37 tags		STTS Tagset (German) - 54 tags	
JJ	adjective, positive	ADJA	adjective, attributive
JJR	adjective, comparative	ADJD	adjective, predicative
JJS	adjective, superlative	NN	common noun
NN	non-plural common noun	NE	proper name
NNS	plural common noun	APPR	preposition
NNP	non-plural proper name	APPRART	preposition incorporating article
NNPS	plural proper name	APPO	postposition
IN	preposition	VVFIN	base verb, finite
VB	base verb	VVIMP	base verb, imperative
VBD	base verb, past tense	VVINF	base verb, non-finite
VBG	base verb, gerund or participle I	VVIZU	base verb incorporating *zu*
VBN	base verb, participle II	VVPP	base verb, participle II
VBP	base verb, non-3rd person	PPOSS	possessive pronoun, substituting
VBZ	base verb, 3rd person	PPOSAT	possessive pronoun, attributive
POS	possessive pronoun	PRF	personal pronoun, reflexive

Part-of-Speech Tagging: Approaches

- Rule-based tagging (with hand-written rules)
- Statistical methods: HMM tagging; Maximium Entropy tagging
- Transformation-based (Brill) tagging: rules and machine learning
- Memory-based tagging

Transformation-based Tagging (TbT)

- Transformation-based Tagging: instance of the transformation-based learning approach to machine learning (Brill, 1995)

- Inspired from both

  1. rule-based taggers: based on rules that specify what tags should be assigned to what words
  2. stochastic taggers: supervised machine learning technique, in which rules are automatically induced from data

- Components:

  - specification of transformations
  - learning algorithm

Transformation-based Tagging – Example

1. Induce likelihoods of word+tag combinations from corpus:

   $P(NN|race) = 0.98$

   $P(VB|race) = 0.02$

2. Label every word with its most likely tag:

   the/DT race/**NN** for/IN outer/JJ space/NN

   is/VBZ expected/VBN to/TO race/**NN** tomorow/NN

3. Apply transformation rules:

   `Change NN to VB when the previous tag is TO.`

   expected/VBN to/TO race/NN $\rightarrow$ expected/VBN to/TO race/VB

Outline

Types and Tokens

- Tokens: total number of word instances in a corpus
  → corpus size

  $Peter_1$ $'s_2$ $father_3$ is_4 a_5 $cook_6$ $.7$
  $Peter_8$ $'s_9$ $mother_{10}$ is_{11} $also_{12}$ a_{13} $cook_{14}$ $.15$

- Types: number of distinct words in a corpus
  → vocabulary size

  $Peter_1$ $'s_2$ $father_3$ is_4 a_5 $cook_6$ $.7$
  $Peter_7$ $'s_7$ $mother_8$ is_8 $also_9$ a_9 $cook_9$ $.9$

Frequency Lists

- Frequency List: corpus types and their frequencies
- Example:

Type	Frequency
Peter	2
's	2
father	1
mother	1
is	2
also	1
a	2
cook	2
.	2

Frequency Distributions – Example

Example: part of the *deWaC* containing 448,675 tokens (beginning)

Rank	Frequency	Tokens
1	23848	,
2	18851	.
3	11907	der
4	10973	die
5	10705	und
6	5880	in
7	4276	den
8	4063	"
9	3967	zu
10	3899	von

Frequency Distributions – Example

Example: part of the *deWaC* containing 448,675 tokens (end)

Rank	Freq.	Tokens (examples)
3750-4609	10	zeitlich, wovon, Tempel, stirbt, Ordnungsmittel
4150-4610	9	samt, planen, normalerweise, kräftig, Jerusalem
4611-5244	8	EDEKA, Genuss, festgenommen, ehrenamtlich, dpa
5245-5981	7	liebt, Möhrenbrei, Kurzfassung, 700, artig
5982-6975	6	Sakristei, seufzte, Rhein, rote, Oh
6976-8442	5	Flower, effektive, Bio-Markt, betreten, CD-Rom
8443-10662	4	unscharf, Tunnel, regeln, Mabuse, BILD
10663-14501	3	Stiefvater, solidarisch, siedelten, Sex, abenteuerliche
14502-23304	2	zzgl., Wirtschaftsbosse, worum, seltsames, schälen
23305-60652	1	Zwickmühle, zweymal, zur., www.tui.com, Vortänzer

Frequency Distributions

- Properties of corpus frequency distributions across corpora
- Beginning of frequency list:
  - function words and punctuation marks
  - frequency of rank x_i is much greater than frequency of rank x_{i+1}
  - in the example: sum of frequencies from the first 10 ranks corresponds to 22% of all tokens
- End of frequency list:
  - content words, compounds, neologisms, typos, web sites
  - number of types with frequency x_i is much greater than number of types with frequency x_{i+1}
  - in the example: words with frequency 1 (hapax legomena) represent 62% of all word types; words with frequencies $1 - 10$ represent 94% of all word types
- Zipf's Law predicts the frequency of a word given its rank.

Outline

Corpora in empirical research
- Corpora and annotation (reminder)
- Preprocessing corpora
  - Tokenisation
  - Morphology
  - Part-of-speech tagging
- Word statistics

Lexical semantics
- Empirical tasks and approaches
- Resources
  - WordNet
  - FrameNet

Lexical semantics in empirical research: Do it yourself!

Lexical Semantics

- Lexical semantics is the study of how and what the words of a language denote.
- Lexical semantics involves the meaning of each individual word.
- A word sense is one of the meanings of a word.
- A word is called ambiguous if it can be interpreted in more than one way, i.e., if it has multiple senses.
- Disambiguation determines a specific sense of an ambiguous word.

Outline

Lexical Semantics: Example Tasks

- Word sense discrimination/disambiguation
- Selectional preferences and semantic roles (semantic parsing)
- Multiword expressions
- Ontological knowledge and representation

Word Senses: Cues

- Probability and prototypicality $\rightarrow$ default interpretation: corpus-related importance of word senses
- Internal text evidence (co-occurrence; context): words, morpho-syntactic embedding, etc.
- One sense per discourse
- Domain
- Real-world knowledge

Context and Co-Occurrence

- The context of a linguistic unit contains indicators for the usage and the meaning of this linguistic unit.

- Examples:
  - Character:
    PIC URE → PICTURE
    PA ER → PAPER
  - Word:
    My grandma used to a delicious cake →
    My grandma used to bake a delicious cake

Distributional Hypothesis

- Words are not combined randomly into phrases and sentences.
- The particular ways in which they go together are a rich and important source of information both about language and about the world we live in.
- Distributional Hypothesis:

  You shall know a word by the company it keeps. (Firth, 1957)

  Each language can be described in terms of a distributional structure, i.e., in terms of the occurrence of parts relative to other parts. (Harris, 1968)

Concordance

- Concordance: a word with its immediate context
- KWIC: key word in context; concordance lines
- Usage:
  - analyse key words
  - analyse word frequencies
  - compare different uses of the same word (context words; structure)
  - find and analyse collocations

Concordance – Example

1. A Christmas Carol, Chapter 1

context The cold within him froze his old features , nipped his pointed nose , shrivelled his cheek , stiffened his gait ; made his eyes red , his thin lips **blue** ; and spoke out shrewdly in his grating voice .

2. David Copperfield, Chapter 3

context Abraham in red going to sacrifice Isaac in **blue** , and Daniel in yellow cast into a den of green lions , were the most prominent of these .

3. David Copperfield, Chapter 3

context It was the completest and most desirable bedroom ever seen - in the stern of the vessel ; with a little window , where the rudder used to go through ; a little looking-glass , just the right height for me , nailed against the wall , and framed with oyster-shells ; a little bed , which there was just room enough to get into ; and a nosegay of seaweed in a **blue** mug on the table .

4. David Copperfield, Chapter 3

context Likewise by a most beautiful little girl (or I thought her so) with a necklace of **blue** beads on , who would n't let me kiss her when I offered to , but ran away and hid herself .

5. David Copperfield, Chapter 5

context Then , we had more tasks until tea , which Mr. Mell drank out of a **blue** teacup , and I out of a tin pot .

6. David Copperfield, Chapter 6

context How well I recollect our sitting there , talking in whispers ; or their talking , and my respectfully listening , I ought rather to say ; the moonlight falling a little way into the room , through the window , painting a pale window on the floor , and the greater part of us in shadow , except when Steerforth dipped a match into a phosphorus-box , when he wanted to look for anything on the board , and shed a **blue** glare over us that was gone directly !

7. David Copperfield, Chapter 7

context He was taken ill in the night - quite prostrate he was - in consequence of Crab ; and after being drugged with black draughts and **blue** pills , to an extent which Demple (whose father was a doctor) said was enough to undermine a horse 's constitution , received a caning and six chapters of Greek Testament for refusing to confess .

8. David Copperfield, Chapter 10

context All within was the same , down to the seaweed in the **blue** mug in my bedroom .

9. David Copperfield, Chapter 10

context But when she drew nearer , and I saw her **blue** eyes looking bluer , and her dimpled face looking brighter , and her whole self prettier and gayer , a curious feeling came over me that made me pretend not to know her , and pass by as if I were looking at something a long way off .

Outline

Lexical Semantics: Resources

- Manual vs. automatic resources
- Types of resources:
  - dictionary
  - thesaurus
  - encyclopaedia
  - ontology
  - taxonomy
  - classification
  - ...
- Examples:
  - WordNet
  - FrameNet

Outline

WordNet

- Online lexical reference system
- Design inspired by psycholinguistic theories of human lexical memory.
- English nouns, verbs, adjectives and adverbs are organised into synonym sets (synsets).
- Each synset represents one underlying lexical concept.
- Different (paradigmatic) relations link the synonym sets.
- WordNet was developed by Princeton University, under the direction of George A. Miller.
- WordNets now exist for many languages.

WordNet Synsets

- Synsets are sets of synonymous words.
- Polysemous words appear in multiple synsets.
- Examples:
  - noun *coffee*:

    {coffee, java}
    {coffee, coffee tree}
    {coffee bean, coffee berry, coffee}
    {chocolate, coffee, deep brown, umber, burnt umber}
  - adjective *cold*:

    {cold} adjective example
    {aloof, cold}
    {cold, dry, uncordial}
    {cold, unaffectionate, uncaring}

Synset Description

- Synset number (= offset)
- List of words
- Relation pointers to other synsets
- Glosses:
  coffee – beverage consisting of an infusion of ground coffee beans
- Examples:
  coffee – “he ordered a cup of coffee”
- Subcategorisation frames

WordNet Relations

Within synsets:

- synonymy, such as {coffee, java}

Between synsets / parts of synsets:

- antonymy: opposition, such as {cold}–{hot}
- hypernymy/hyponymy: is-a relation,
  such as {coffee, java}–{beverage, drink, potable}
- meronymy/holonymy: part-of relation,
  such as {coffee bean, coffee berry, coffee}–{coffee, coffee tree}

Morphology:

- compounds: arabian coffee, coffee break, coffee table

Outline

FrameNet

- Frame-semantic descriptions for English verbs, nouns, and adjectives
- Aim: document the range of semantic and syntactic combinations (valences) of each word in each of its senses
- Result: lexical database with
  - descriptions of the semantic frames
  - a representation of the valences for target words
  - a collection of annotated corpus attestations

FrameNet Vocabulary

- Frame semantics, developed by Charles Fillmore:
  - a theory that relates linguistic semantics to encyclopaedic knowledge
  - describes the meaning of a word (sense) by characterising the essential background knowledge that is necessary to understand the word/sentence
- Frame: conceptual structure modelling prototypical situations
- Frame element: frame-evoking word or expression
- Frame roles: participants and properties of the situation

FrameNet: Example Frames

- apply heat: common situation involving a cook, some food, and a heating instrument;
  elements: *bake, blanch, boil, broil, brown, simmer*, etc.

- change position on a scale: situation involving the change of an item's position on a scale (the attribute) from a starting point (initial value) to an end point (final value);
  elements: *decline, decrease, gain, rise*, etc.

- damaging: an agent affects a patient in such a way that the patient (or some subregion of the patient) ends up in a non-canonical state;
  elements: *damage, sabotage, scratch, tear, vandalise*, etc.

FrameNet: Example Annotations

- verbs:
  [*Cook* Matilde] fried [*Food* the catfish] [*Heating instrument* in a heavy iron skillet].
  [*Item* Colgate's stock] rose [*Difference* $3.64] [*Final value* to $49.94].

- noun:
  ... the reduction [*Item* of debt levels] [*Final value* to $25] [*Initial value* from $2066]

- adjective:
  [*Sleeper* They] were asleep [*Duration* for hours].

Outline

- Corpora in empirical research
  - Corpora and annotation (reminder)
  - Preprocessing corpora
    - Tokenisation
    - Morphology
    - Part-of-speech tagging
  - Word statistics
- Lexical semantics
  - Empirical tasks and approaches
  - Resources
    - WordNet
    - FrameNet

Lexical semantics in empirical research: Do it yourself!

Check out Word Senses and Frames

1. Identify an (ambiguous) noun, verb and/or adjective you are interested in.

2. Look into the words' concordances using the IMS Open Corpus Workbench online demo.

3. Check out the words' senses and related words using WordNet and WordNet Search.

4. Check out the words' semantic frames using English FrameNet.