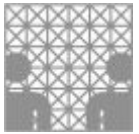


Language Understanding

nach Biemann, NLP-Vorlesung

Analyse-Levels

Phonologie

Segmentierung

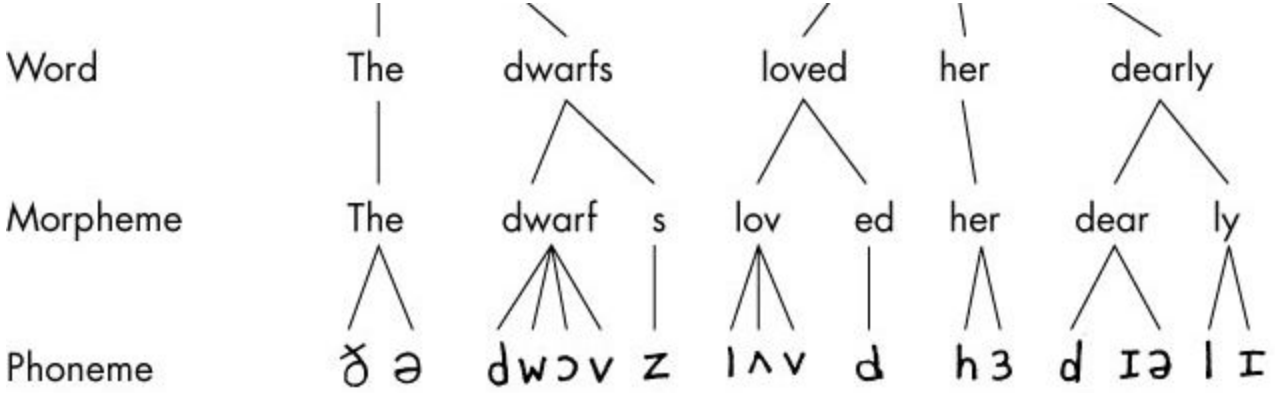
Morphologie

Syntax (POS-Tagging, Parsing)

(Semantik)

Phonologie

Phoneme (Groome, 2006):



-> Homophone (knight/night, wreck a nice beach/?)

Segmentation

- Worte und Sätze, einige Definitionen
  - Leerzeichen an beiden Seiten? *(Bauer, 1988)*
  - Nur Bindestriche und Apostrophe im Wort? *(Kučera and Francis, 1967)*
  - Sätze enden mit Punkten? *(Grefenstette und Tapanainen, 1994)*
- Tokenisierung: Input in geordnete Tokens umwandeln (“tokenizer”)
  - John likes Mary and Mary likes John. → {"John", "likes", "Mary", "and", "Mary", "likes", "John", "."}
  - Er sagte: “Ich finde, 42€ sind zu wenig.” → …
  - Mehrdeutig: Punkte, Leerzeichen (U.S.A., Frankfurt a.M.), Kommas (1,3), Apostrophe (geht’s?), Bindestriche (Seiten 2-4)

Morphologie

- Wortform und -formation
- Morpheme sind die kleinsten bedeutungstragenden Einheiten:
  - Katzen: Katz (Nomen) + en (Plural)

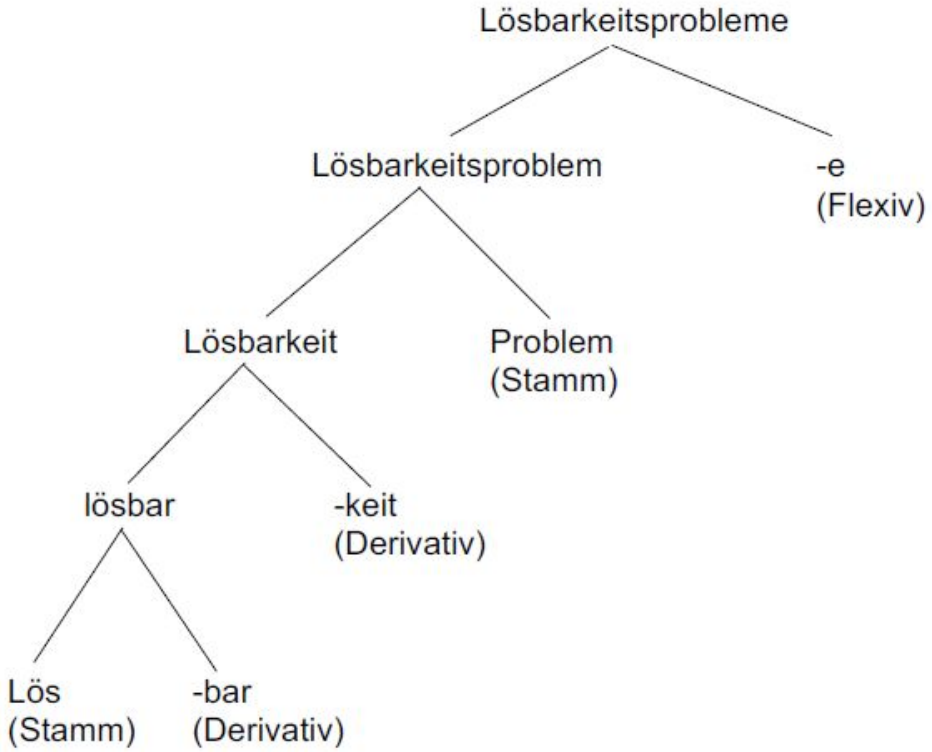
uygarlaştıramadıklarımızdanmışsınızcasına

uygar	laş	tır	ama	dık	lar	ımız	dan	mış	sınız	casına
civilized	+BEC	+CAUS	+NABL	+PART	+PL	+P1PL	+ABL	+PAST	+2PL	+AsIf

"(behaving) as if you are among those whom we could not civilize"

- Stamm (Katz), Suffixe (Katz-en), Präfixe (un-wahr), Infixe (fan-bloody-tastic), Zirkumfixe (ge-tag-t)

Morphologische Analyse: *Lösbarkeitsprobleme*



Morphologische Normalisierung

- Eine kanonische Repräsentation morphologisch verwandter Wortformen:
  - Stemming
  - Lemmatisierung
- Stemming: sitzen → sitz
  - Saw?
- Lemmatisierung: Rückführung von flektierter zu unflektierter Form
  - gesagt -> sagen, ging-> gehen

Syntax

- Regularität und Einschränkung der Anordnung von Wörtern *(Manning & Schütze, 2003, p. 93)*
- POS-Tagging: Part-Of-Speech

  - “einen” → ?

Syntax

- Regularität und Einschränkung der Anordnung von Wörtern *(Manning & Schütze, 2003, p. 93)*
- POS-Tagging: Part-Of-Speech

- "einen" → **VVINF**: *Er kann das Volk einen.*
- "einen" → **VVFIN**: *Sie einen Deutschland.*
- "einen" → **ART**: *Er hat einen Apfel.*

Parts Of Speech

- Englisch: **traditionell** 8 POS-Tags:
  - N: Noun
    - chair, bandwidth, pacing
  - V: Verb
    - study, debate, munch
  - ADJ: Adjective
    - purple, tall, ridiculous
  - ADV: Adverb
    - unfortunately, slowly
  - P: Preposition
    - of, by, to
  - PRO: Pronoun
    - I, me, mine
  - DET: Determiner
    - the, a, that, those

Language	Tagset Size
English	139
Czech	970
Estonian	476
Hungarian	401
Romanian	486
Slovene	1033

Hajic, 2000

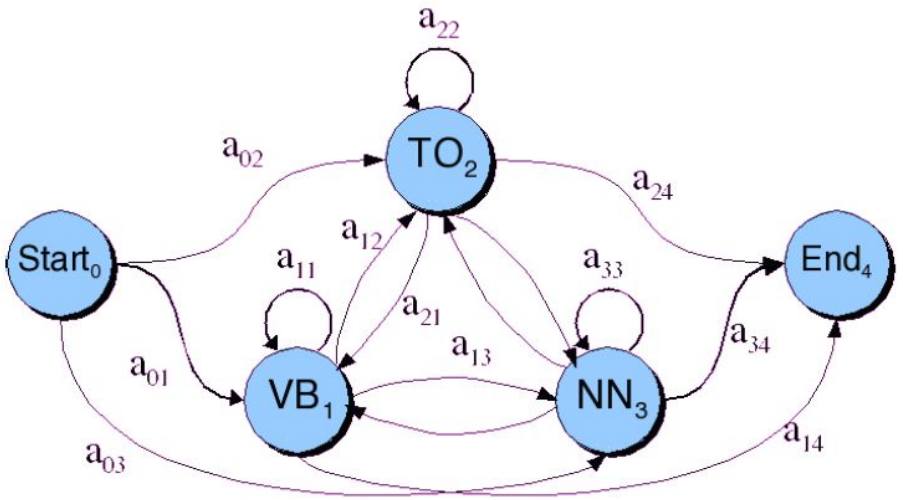
Parts Of Speech

- Informationen über:
  - Wort, mögliche Nachbarn
  - Betonung
  - Bedeutung (Ball)
  - Lemmatisierung
- Nicht eindeutig (back: N, ADJ, ADV, V)

POS Tagging

- POS-Tag L für Wort W: $L_{\max} = (l^1_{\max}, l^2_{\max}, \ldots l^T_{\max}) = \underset{L}{\operatorname{argmax}} \; P(L \mid W)$

- State transitions:



POS Tagging

- POS-Tag L für Wort W: $L_{max} = (l^1_{max}, l^2_{max}, \ldots l^T_{max}) = \underset{L}{\text{argmax}} \; P(L \mid W)$

- State transitions: aus Brown-Corpus, 87 tags, ohne Smoothing

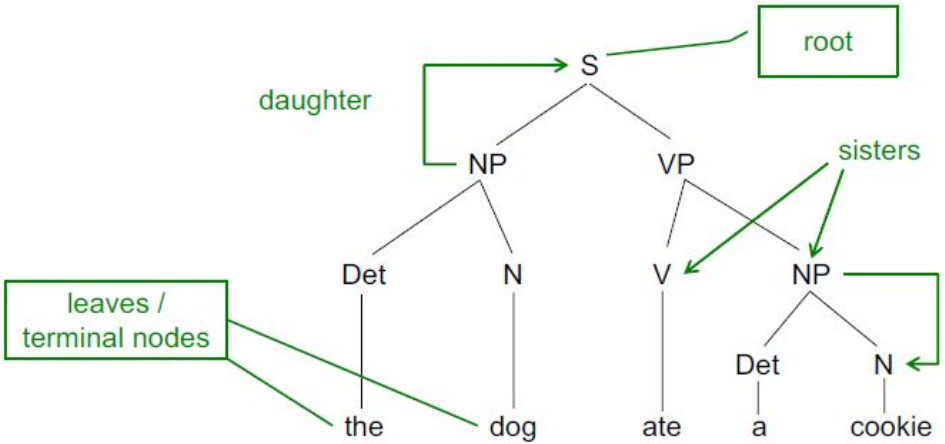
	VB	TO	NN	PPSS
<s>	.019	.0043	.041	.067
VB	.0038	.035	.047	.0070
TO	.83	0	.00047	0
NN	.0040	.016	.087	.0045
PPSS	.23	.00079	.0012	.00014

Phrasen

- Sequenzen verwandter Wörter
  - sie, der Hund, der große Hund
  - Identifizierbar durch Ersetzung (Ich habe einen Hund → Ich habe ihn)
- Kopf: legt syntaktischen Typ der Phrase fest
- Modifikatoren behandeln ein anderes Wort oder Phrase
  - Pre-, Postmodifieres
- Phrasen-Arten:
  - Prepositional phrase: in love
  - Noun phrase: blaues Wunder
  - Verb phrase: Kuchen essen
  - Adjectival phrase: voller Wasser
  - Adverbial phrase: sehr vorsichtig

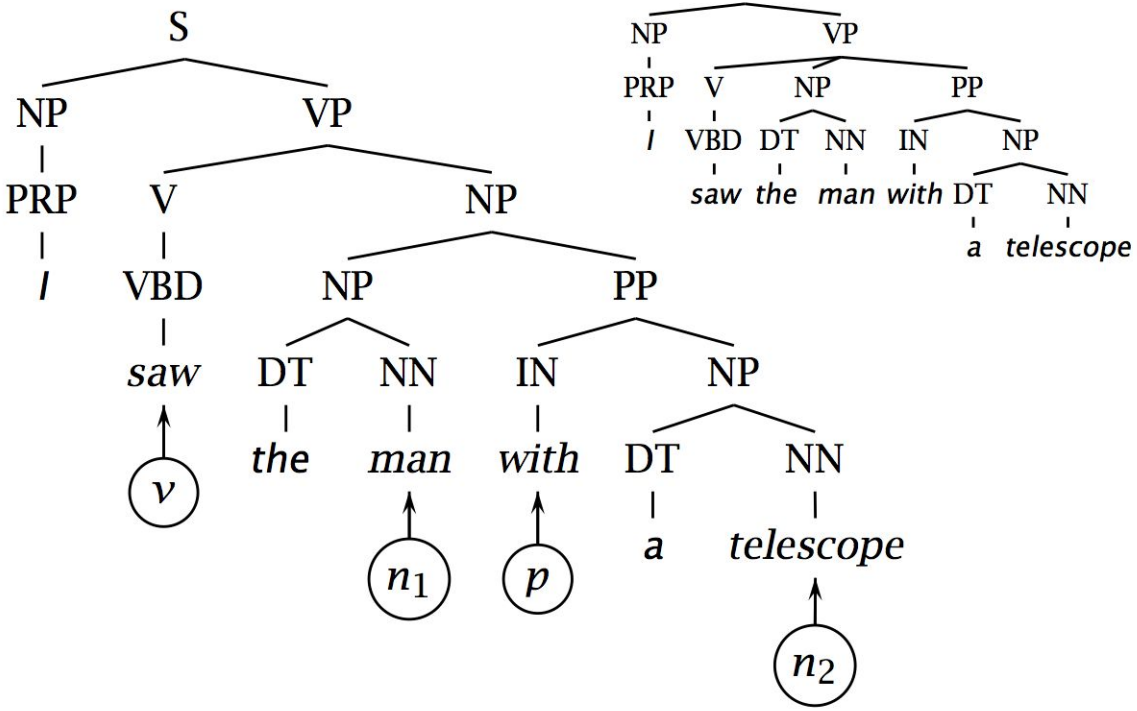
Grammatiken

- Erlaubte Strukturen einer Sprache durch Regeln (terminale- und non-terminale Symbole)
- Parsing: Modellierung durch Kontextfreie Grammatiken



Ambiguities

- Attachment ambiguity: I shot an elephant in my pyjamas
- Coordination ambiguity: Große Äpfel und Orangen
- “Garden Path sentences”: The old man the boat
  - The horse raced past the barn fell.
    The man whistling tunes pianos.
    The cotton clothing is made of grows in Mississippi.
    The complex houses married and single soldiers and their families.
    The author wrote the novel was likely to be a best-seller.
    The tomcat curled up on the cushion seemed friendly.
    The man returned to his house was happy.
    The government plans to raise taxes were defeated.
    The sour drink from the ocean.
    - Wikipedia


S
NP
VP
PRP
V
NP
I
VBD
NP
PP
saw
DT
NN
IN
NP
the
man
with
DT
NN
v
a
telescope
n_1
p
n_2


pretty little girl 's school

Semantik

- **Bedeutung** eines Wortes, Satzes, Dokuments…

Hund

UFO

Liebe

Semantik (2)

Ball

Löffel

Mutter

Semantik (2)

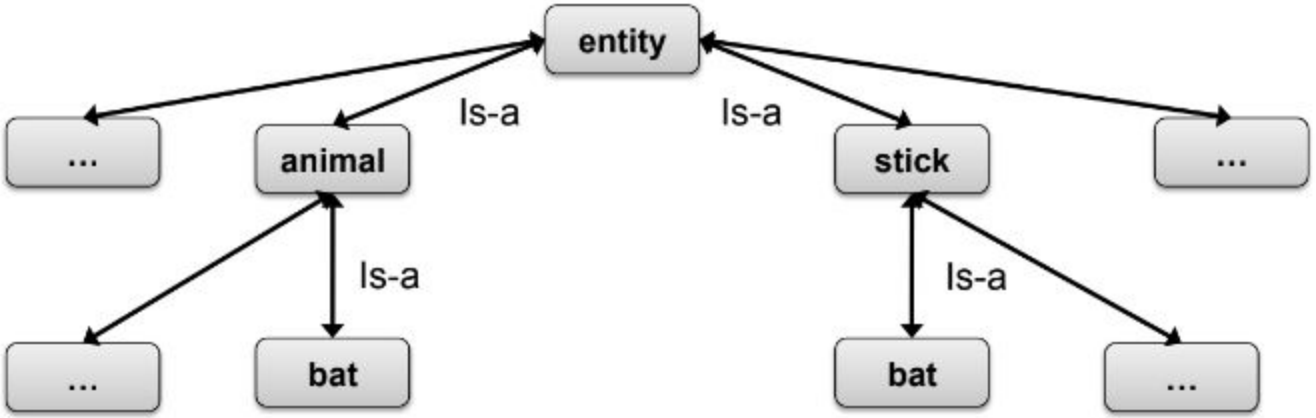
Ball

Löffel

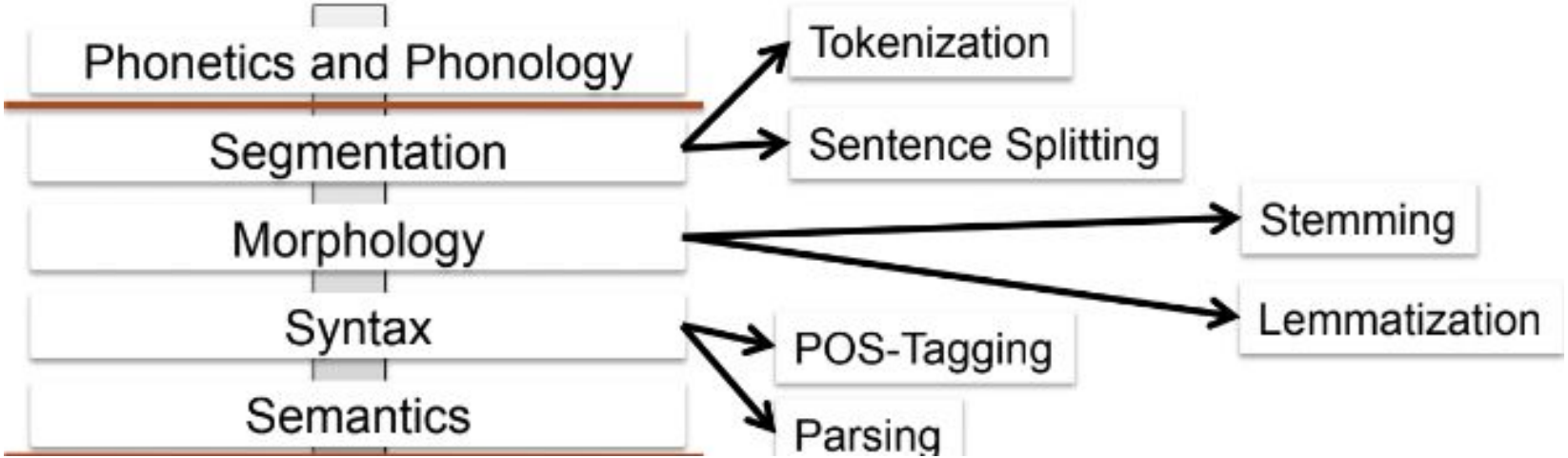
Mutter

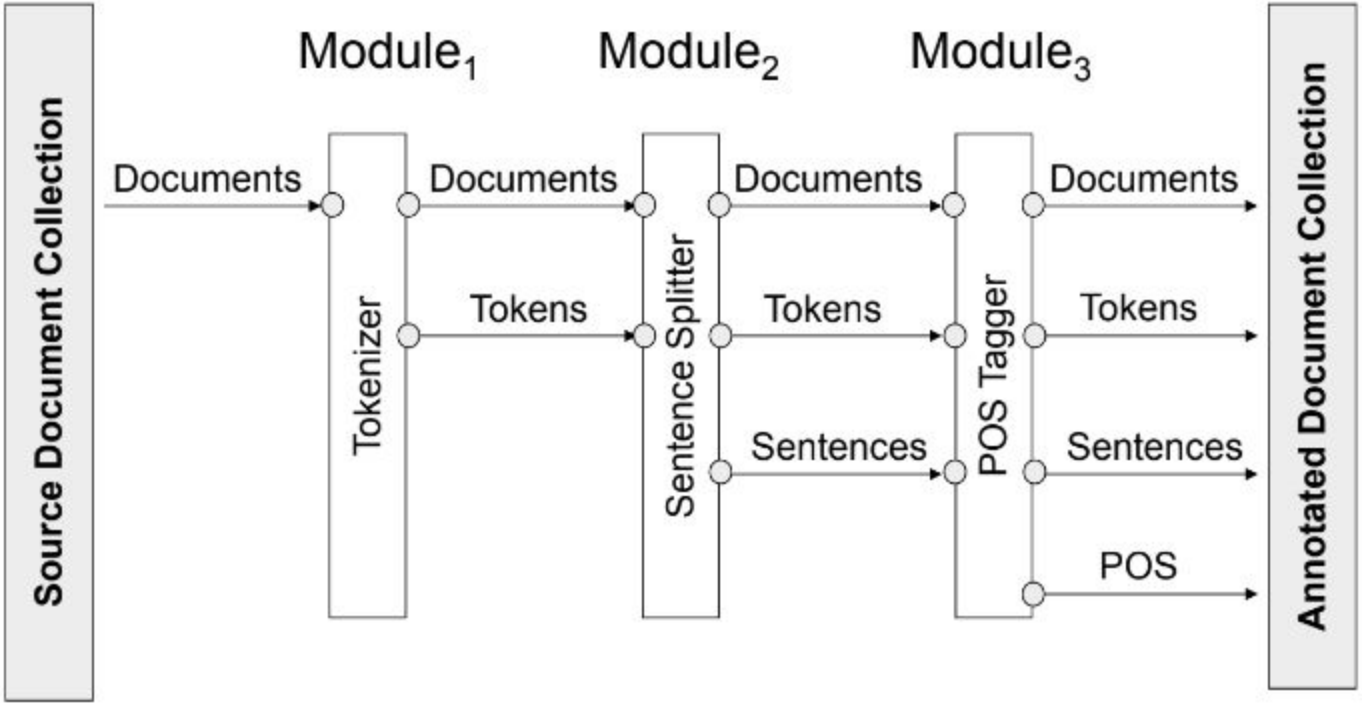
Mehrdeutigkeit

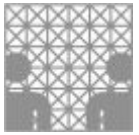
- Homonyme:
  - Eigentliche Homonyme: gleiche Flexion (der Ball, die Bälle, des Balls…)
  - Uneigentliche Homonyme: (die Bank, die Bänke/Banken)
- Homographie (Hochzeit), Homophonie (Bug/buk)
- Lexikalische Mehrdeutigkeit:
  - He hit the ball with the **bat**
  - **Time flies like an arrow.**
    - (as an imperative) Measure the speed of flies like you would measure that of an arrow—i.e. (You should) time flies as you would time an arrow.
    - (imperative) Measure the speed of flies in the way an arrow would—i.e. (You should) time flies in the same manner that an arrow would time them.
    - (imperative) Measure the speed of flies with qualities resembling those of arrows—i.e. (You should) time those flies that are like an arrow.
    - (declarative) Time moves in a way an arrow would.
      (declarative, i.e. neutrally stating a proposition) Certain flying insects, called "time flies", enjoy an arrow.


entity
ls-a
ls-a
...
animal
stick
...
ls-a
ls-a
...
bat
bat
...

Verarbeitung




Source Document Collection
Module1
Module2
Module3
Documents
Tokenizer
Documents
Tokens
Sentence Splitter
Documents
Tokens
Sentences
POS Tagger
Documents
Tokens
Sentences
POS
Annotated Document Collection



Danke/PTKANT!/$.

Fragen/NNS?/$.