Einführung in die Computerlinguistik

Morphologie III

Hinrich Schütze & Robert Zangenfeind

Centrum für Informations- und Sprachverarbeitung, LMU München

7.12.2015

Take-away

- Wortarten / Lexikalische Kategorien
- Schwierigkeiten bei der Zuordnung zu Wortarten
- Anwendungen

Einleitung

Wozu Wortarten (POS)?

- Viele syntaktische Eigenschaften sind identisch für (große) Klassen von Wörtern
- Regeln gelten nur für bestimmte Kategorien von Lexemen
- Kategorisierung der Lexeme nötig $\rightarrow$ Generalisierungen werden möglich
- Wichtig für Anwendungen in CL
- z.B. Mehrdeutigkeit von Schlagzeilen:

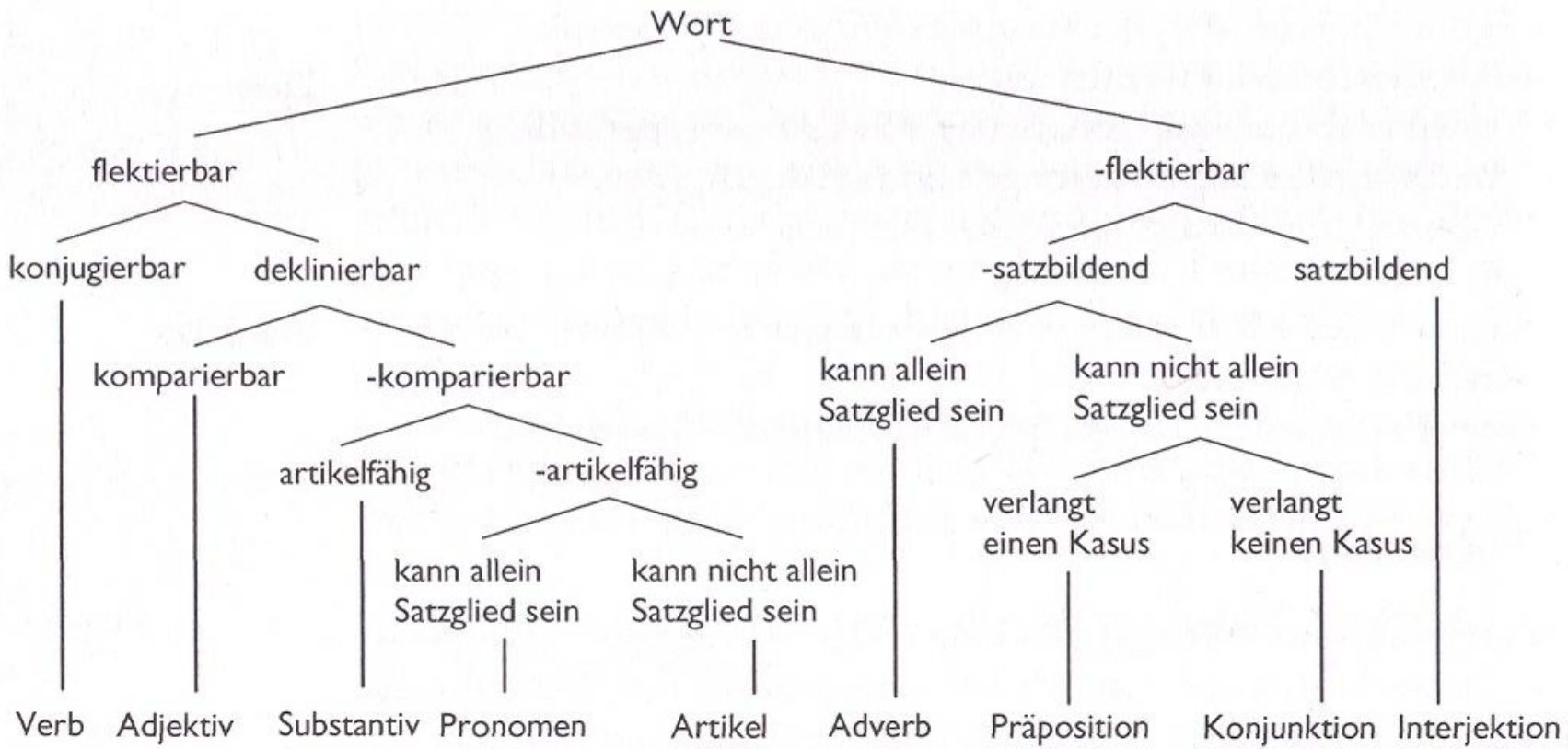
(1) Immer überlegen!
(2) Die verspielte Revolution
(3) Laden stört

Kriterien zur Klassifizierung

- Lexeme bilden z.T. offene Listen → nicht aufzählbar
- vgl. dagegen: grammatische Morpheme bilden geschlossene Listen
- Linguistische Kriterien sind nötig zur Klassifizierung:
  - (i) morphologische Kriterien nutzen Flektierbarkeit (funktioniert nicht für alle Lexeme)
  - (ii) syntaktisch distributionelle Kriterien (→ Kontext des Wortes), z.B. Konjunktionen verknüpfen Wörter, Phrasen, Sätze

→ eine mögliche Klassifizierung:

Klassifizierung von Heringer (2009:64)



Klassifizierung: Verben

- konjugierbar
- Vollverben, Hilfsverben, Modalverben, Funktionsverben (Stützverben)

Klassifizierung: Nomen

- deklinierbar
- festes Genus
- Subkategorisierung: Substantive – Pronomen

Klassifizierung: Adjektive

- deklinierbar (nicht bei prädikativer Verwendung!)
- außerdem:
  - z.B. *einerlei*, *futsch*, *quitt* (nicht deklinierbar)
  - z.B. *lila*, *orange*, *rosa* (ungern dekliniert)
  - *sonstig*, *ständig* (keine prädikative Verwendung)
  - viele komparierbar

Klassifizierung: Determinierer

- geschlossene Liste
- definite Artikel (def) (z.B. *der*): Flexive verschmelzen mit Lexem
- weitere: ind (z.B. *ein, eine*); dem (z.B. *diese, jene, dieselben, solche*); qnt (z.B. *alle, jeder, viele, beide*); neg (z.B. *kein, keine*); pss (z.B. *mein, ihr*); int (z.B. *welche*)

Klassifizierung: Präpositionen

- weisen Nomen Kasus zu
- stehen meist links (z.B. *in*, *auf*, *für*)
- manche rechts (z.B. *zufolge*)
- wenige: beides möglich (z.B. *wegen*)
- manche umschließen Nomen (z.B. *um . . . willen*)

Klassifizierung: Adverbien

- nicht flektierbar
- manche steigerbar
- adverbial gebrauchte Adjektive bleiben in ihrer Klassifizierung Adj.

Klassifizierung: Konjunktionen

- verbinden gleichartige syntaktische Einheiten (Sätze, Phrasen, Wörter, Wortteile); (z.B. *und, oder, aber, entweder . . . oder*)
- geschlossene Liste

Klassifizierung: Interjektionen

- syntaktisch unverbundene, satzwertige Äußerungen
- drücken Empfindung, Bewertung oder Willen des Sprechers aus (z.B. *aha, igitt, richtig*)
- übermitteln Aufforderung oder Signal der Kontaktaufnahme (z.B. *Hallo, Prost, Hey*)

Klassifizierung: Subjunktionen

- (traditionell zu Konjunktionen)
- Bindewörter
- verbinden propositionale Einheiten verschiedener Stufen (Nebensätze, Infinitivkonstruktionen) (z.B. *dass*, *weil*, *obwohl*, *um*, *statt*)
- stehen linksperipher
- geschlossene Liste

Klassifizierung: Partikeln

- syntaktisch ungebunden (fast beliebige Wortstellung)
- keine eigene Phrase (z.B. *wohl*, *nur*, *kaum*, *nicht*)
- $\rightarrow$ Zuordnung zu Klassifikationsbaum?

Klassifizierung: Satzwörter

- $\rightarrow$ Einwortsätze
- Interjektionen (z.B. *Aha!*)
- weitere: *ja*, *nein*, *Danke*

Schwierigkeiten bei der Zuordnung (1)

- Wortartwechsel
  - *Leid* (vgl. z.B.: *Das tut mir leid*)
  - *Klasse* (vgl. z.B.: *ein klasse Buch*)
  - *ja* (vgl. z.B.: *Das war ein klares Ja*)

Schwierigkeiten bei der Zuordnung (2)

- Zugehörigkeit zu mehreren Wortarten wg. Ambiguität
  - z.B. *aber*:
  - *Er las, aber er war sehr unkonzentriert* (Konj.)
    vs.
  - *Das kann man aber so nicht sagen* (Partikel)

Schwierigkeiten bei der Zuordnung (3)

- Zahlwörter?
  - *eins* (Duden: Kardinalzahl etc.): deklinierbar
  - *zwei* (ebenso): z.B. *der Bund zweier Kaiser*
  - *hundert* (Duden: Kardinalzahl) aber auch *Hundert*! („das Hundert vollmachen“)
  - *tausend* (ebenso)
  - *Million*: eher wie Nomen

Schwierigkeiten bei der Zuordnung (4)

- z.B. singuläres Lexem: *zu* (Partikel, aber ähnlich Konjunktion)
  - verlangt Infinitiv (*Sie half uns das Gepäck zu tragen*), oder
  - mit Partizip I: *die zu erledigenden Aufgaben*

Schwierigkeiten bei der Zuordnung (5)

- z.B. Sonderfall *viel*
  - Duden: Indefinitpron. u. unbest. Zahlwort
  - Heringer: quantitativer Determinierer
  - wie bei *Vieles Erfreuliche stand in dem Brief, Er trank viel Bier*
  - wie Adj.: *viele Tiere, die vielen Tiere, das viele Laub*)

Part-of-Speech Tagging (POS Tagging)

- Wörter eines Textes mit dazugehörigen Wortarten (engl. Part-of-Speech) kennzeichnen
- eine Art der Annotierung des Textes/Korpus
- Wortart gibt viele Informationen über das Wort und seine benachbarten Wörter im Text
  - z.B. Possessivpronomen (z.B. *mein*, *dein*, *sein*, *unser*) $\rightarrow$ rechts davon: häufig Nomen
    vs.
  - Personalpronomen (z.B. *ich*, *du*, *er*, *wir*) $\rightarrow$ rechts davon: Verb
- Tagging manuell oder durch Algorithmen (regelbasierte oder statistische Methoden (z.B. Hidden-Markov-Modelle))

Programme hierzu im Netz (1)

- CIS, LMU München: MarMoT: http://cistern.cis.lmu.de/marmot/ (download)
- CIS, LMU München: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (download)
- Uni Kopenhagen: Brill Tagger: http://cst.dk/online/pos_tagger/uk/index.html (POS-Tagger für das Englische; Deutsch nur Tokeniser und Lemmatiser) (Demo; auch download)
- ETAP-3: http://proling.iitp.ru/ru/etap3 (Demo; engl., russ.)

Programme hierzu im Netz (2)

- Stanford-Parser:
  http://nlp.stanford.edu/software/tagger.shtml (download; engl.)
- Uni Saarbrücken: TnT – Statistical Part-of-Speech Tagging:
  http://www.coli.uni-saarland.de/~thorsten/tnt/ (download)
- Uni Genf: ISSCO TAgger TOOl:
  http://www.issco.unige.ch/en/staff/robert/tatoo/tatoo.html (download)

Weitere Anwendungen

- Maschinelle Übersetzung (Textanalyse, Textsynthese)
- Informationsextraktion
- Information Retrieval
- Text Mining
- Suchmaschinen
- Textgenerierung

Stemming (Stemmatisierung)

- Alternatives Verfahren zum Lemmatisieren
- Flexionsmorpheme von Wortform werden beseitigt
  → Wortstamm (wird der Wortform zugeordnet)
- z.B. engl. Wortformen *process*, *processing*, *processed*
  → Stamm *process*
- Problem: sinnvolle Unterscheidungen können verloren gehen:
- z.B. *stocks* (Aktien etc.) und *stockings* (‘Strümpfe’ etc.)
  → Stamm *stock* (‘Aktie’ etc.)

Programm hierzu im Netz

- Porter-Stemmer (Demo und download):
  http://snowball.tartarus.org
- dt. Beispiele:
  http://snowball.tartarus.org/algorithms/german/stemmer.html

Take-away

- Wortarten / Lexikalische Kategorien
- Schwierigkeiten bei der Zuordnung zu Wortarten
- Anwendungen