# Lexikalische Kategorien (Wortarten)

Dr. Benjamin Roth

CIS LMU München

# Gliederung

# Wozu Wortarten (Part-of-Speech, 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
- Ambiguität: Wortart einer Form wird oft erst durch den Kontext bestimmbar:
  *Time flies like an arrow.*

# POS-tagging für Anwendungen in der Computerlinguistik

- POS-tagging (Part-of-speech-tagging): Automatische Wortartbestimmung
- Stemming: Grundform eines Wortes kann gefunden werden, wenn Wortart bekannt.
- Maschinelle übersetzung: Richtige Übersetzung hängt von verwendeter Wortart ab.
- Zusammen mit Tokenisierung einer der am häufigsten Verwendeten Vorverarbeitungsschritte in der Computerlinguistik.

# Outline

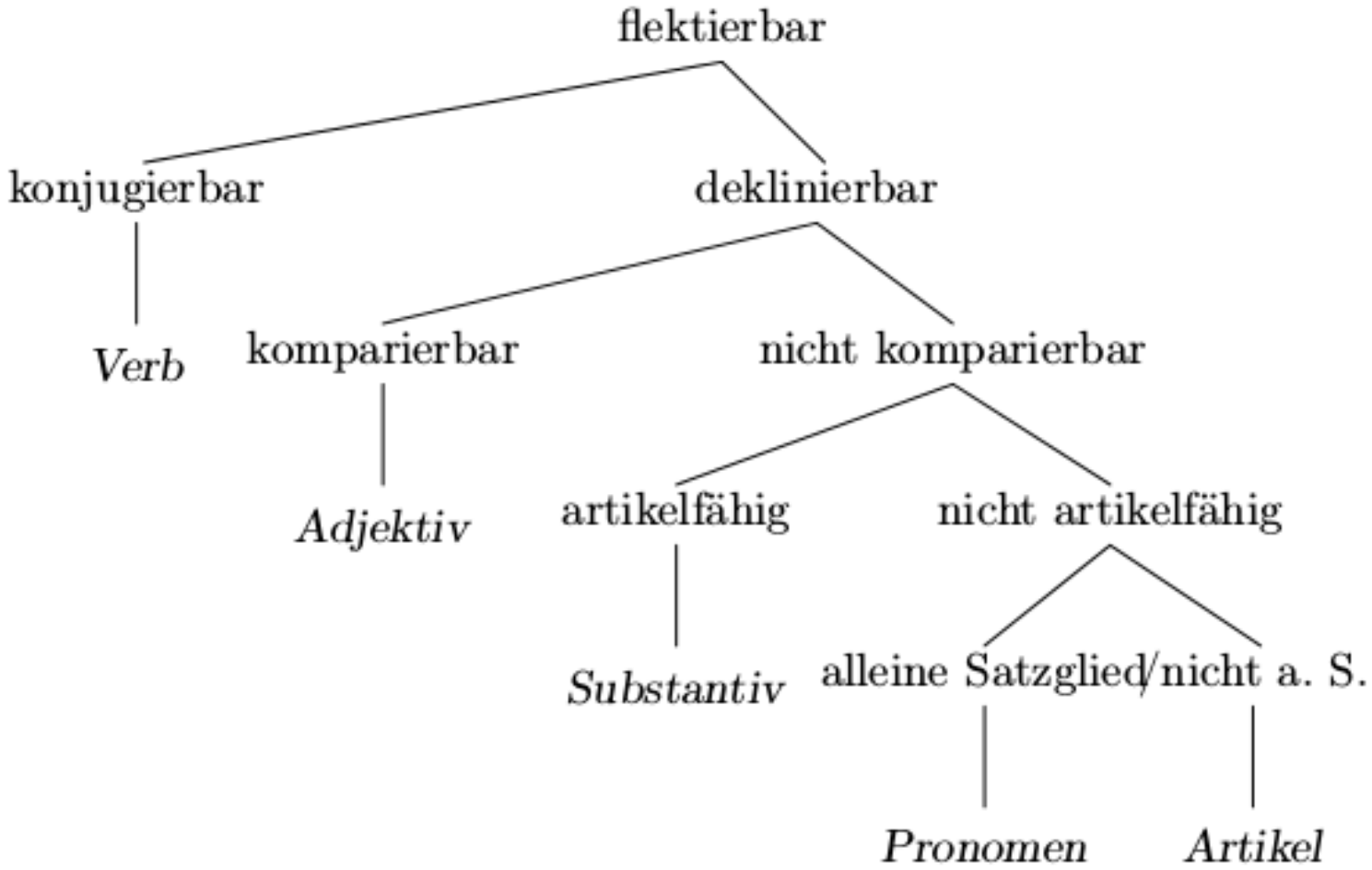
# Kriterien zu Wortartklassifizierung

- Lexeme bilden z.T. offene Listen $\Rightarrow$ nicht aufzählbar
  (vgl. dagegen: grammatische Morpheme bilden geschlossene Listen)
- Linguistische Kriterien sind nötig zur Klassifizierung
- Eine gängige Art der Klassifizierung richtet sich nach **morphologisch-syntaktischen Kriterien**.

# Morphologisch-syntaktische Kriterien

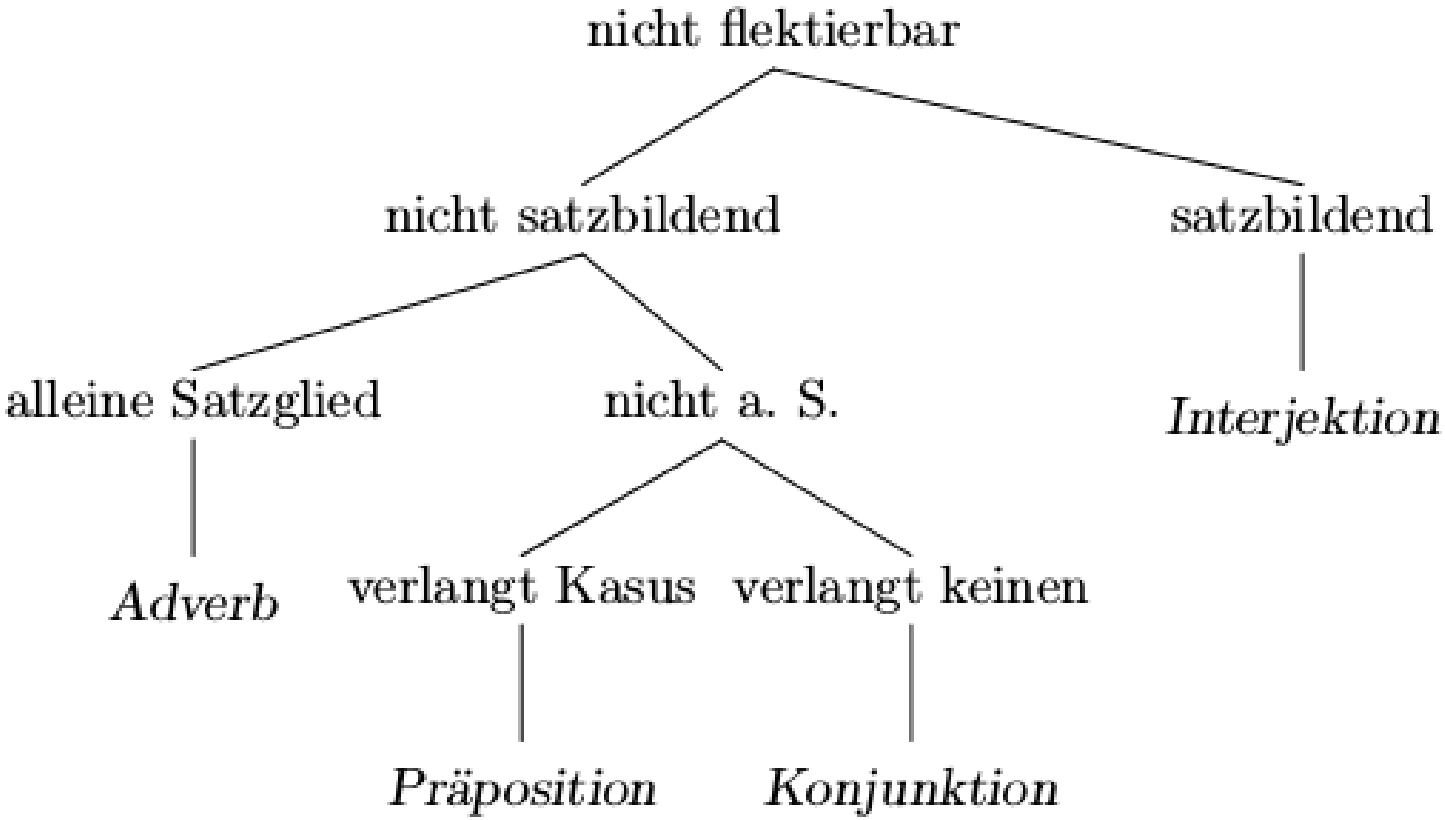
- Morphologisch:
  - **flektierbar**: Substantiv, Adjektiv, Pronomen, Numerale, Verb, Artikel
  - **nicht flektierbar**: Präposition, Kunjunktion, Partikel
  - bei Adverbien ist oft nicht klar ob sie flektierbar sind, man kann ja den Komparativ bilden.
- Syntaktisch:
  - die Fähigkeit **als Satzglied zu fungieren**
  - die Fähigkeit **einen Artikel zu binden**
  - die Fähigkeit **einen bestimmten Kasus zu fordern**
- Die Hauptunterscheidung wird zwischen **flektierbaren** und **nicht-flektierbaren** Lexemen getroffen, die Wortarten werden davon ausgehend weiter eingeteilt.

# Wortarten (flektierbare Lexeme)




Klassifizierung von Heringer, H.-J.: Morphologie. Paderborn 2009.

# Wortarten (nicht-flektierbare Lexeme)




Klassifizierung von Heringer, H.-J.: Morphologie. Paderborn 2009.

# Weitere Kriterien: Semantisch

- Autosemantika: Substantiv, Adjektiv, Adverb, (Voll-)Verb
- Synsemantika: Hilfsverb (sein, haben, werden), Hilfspartikel (zu)
- Pronomen, Präposition, Artikel und Partikel lassen sich schlecht in dieses Schema einordnen

# Weitere Kriterien: Nach Produktivität

- offene Klassen sind Bestandteile des Lexikons und können durch Wortbildungsregeln jederzeit erweitert werden: Verb, Nomen, Adjektiv, Adverb
- geschlossene Klassen sind im Prinzip aufzählbar und somit in die Grammatik integrierbar: Präposition, Artikel, Konjunktion

# Weitere Kriterien: Nach Kasuszuweisung

Kann das Lexem den Kasus eines Satzgliedes bestimmen?

- z.B. Subjekt, Akkusativ-, Dativ-, Genitive-Objekte bei Verben
- Später mehr zu Satzgliedern.

| | NOM | AKK | DAT | GEN |
|---|---|---|---|---|
| *Verb* | X | X | X | X |
| *Präp.* | | X | X | X |
| *Adj.* | | (X) | X | X |
| *Nomen* | | | | X |

# Outline

# Wortarten: Übersicht und Beispiele

| | | |
|---|---|---|
| tree, dog, freedom \| Baum, Ofen | → | Substantiv / Nomen (Hauptwort) |
| run, kick, work \| sprechen, müssen | → | Verb (Zeitwort) |
| big, red, beautiful \| braun, ehrlich | → | Adjektiv (Eigenschaftswort) |
| a(n), some, any, the, this, that \| der, die, das | → | Artikel / Determinator (Geschlechtswort) |
| du, sie | → | Pronomen (Fürwort) |
| sieben, anderthalb | → | Numerale (Zahlwort) |
| today, there, well, strangely \| heute, sehr | → | Adverb (Umstandwort) |
| in, on, below, against \| für, auf | → | Präposition (Verhältniswort) |
| that, because, although \| wenn, weil | → | Konjunktion (Bindewort) |
| ouch, oops, oh, psst \| ... ! | → | Interjektion (Empfindungswort / Ausrufewort) |

# Wortart Verb

- Konjugierbar: morphologische Kennzeichnung nach Person, Tempus, Numerus, ...
- Typischerweise Kongruenz in Person, Numerus und/oder Genus mit einem oder mehreren Argumenten (z.B. mit Subjekt)
- Einteilung nach Stelligkeit: Valenzklassen
  - Verben ohne Ergänzung
    Es **schneit**.
  - Intransitive Verben (nur Subjekt)
    Martin **schnarcht**.
  - Transitive Verben (Subjekt und Akkusativobjekt)
    Der Professor **lobt** seine Studenten.
  - Ditransitive Verben (Subjekt, Akkusativ- und Dativobjekt)
    Hans **verkauft** sein Auto einem Freund.
  - Verben mit Genitiv- oder Dativobjekt (ohne Akkusativobjekt):
    Wir **gedenken** der Toten.
    Die Spieler **danken** dem Trainer.
  - Verben mit Präpositionalobjekt:
    Ich **ziehe** nach China.

# Besondere Verbklassen

- **Kopulaverben** (sein, werden) spezifizieren lediglich das Tempus, während der semantische Gehalt vom Nomen oder Adjektiv beigetragen wird.
  *Die Vorwürfe* **sind** *schwerwiegend*.
- Bei sog. **Lightverb Constructions** (keine eigene Wortart; dt. *Stützverben*) kommt die Hauptbedeutung durch ein Satzglied, mit dem das Verb eine lexikalisierte Verbindung eingegangen ist:
  *Ich* **ziehe** *alle Optionen* **in Erwägung**.
  *Er* **erhebt** *schwere* **Vorwürfe**.
- **Modalverben** (können, müssen, sollen, ...) spezifizieren die Möglichkeit oder Notwendigkeit einer Aussage.
  *Ich* **kann** *morgen nicht zum Training kommen.*

# Wortart: Nomen und Pronomen

- Deklinierbar: morphologische Kennzeichnung von Kasus, Genus und Numerus
- Nomen: Festes Genus, offene Klasse
- Pronomen:
  - geschlossene Klasse
  - verweisen auf etwas, haben als Zeichen alleine keine Referenz.
- Pronomina-Unterklassen
  1. Personalpronomina: ich, du er, sie, es, mich, dir
  2. Reflexivpronomina: sich
  3. Possessivpronomina: mein, dein, sein
  4. Demonstrativpronomina: diesen
  5. Relativpronomen: der, welcher
  6. Interrogativpronomen: welcher, wer, was
  7. Indefinitpronomen: jemand, etwas, alle, kein

# Wortart Adjektive

- **Attributive Verwendung**: *das große Haus.*
- **Prädikative Verwendung**: *Das Haus ist groß.*
- Rein **attributive** Adjektive: *der ehemalige Präsident*
  vs. **der Präsident ist ehemalig*
- Rein **prädikative** Adjektive: *die Regierung ist schuld*
  vs. **die schulde Regierung*
- **Deklinierbar** (nur wenn attributive Verwendung möglich!)
- Meist **komparierbar**
- bestimmte Adjektive verlangen **Ergänzungen**:
  *seinem Bruder ähnlich ...*
  *sich seiner Schuld bewusst ...*
  *bei uns beliebt ...*
  *in Köln wohnhaft ...*
  *seiner Überzeugung sicher ...*
  *der Idee dienlich...*
  *...sein*

# Wortart Adverb

- Modifizieren ein Verb oder Adjektiv.
  Sie ist **schon** da.
  Ich werde **bald** gehen.
  Das hat mir **sehr** geholfen.
  Ein **äußerst** hilfreiches Buch.
- Nicht flektierbar.
- Manche steigerbar.
- Konvention: Adverbial gebrauchte Adjektive bleiben in ihrer Wortart-Kategorie Adjektiv.
  *Er fährt* **schnell**.
  $\Rightarrow$ Wortart vs. syntaktische Funktion

# Wortart Artikel (Determinierer)

- Geschlossene Liste
- Syntaktische Funktion: komplettieren eine Nominalphrase (s.a. X-bar-Theorie)
- Definite Artikel: **der** Hut, **die** Katze, **das** Haus
  $\Rightarrow$ verweisen normalerweise auf Entitäten, die bereits bekannt sind, schon in den Diskurs eingeführt wurden, oder deren Existenz logisch aus anderen Informationen folgt (s.a. Diskursanalye, Pragmatik).
- Indefinite Artikel: **ein** Hut, **eine** Katze, **ein** Haus
  $\Rightarrow$ führen z.B. neue Referenten in den Diskurs ein, auf die später referenziert werden kann.
- Beispiel:
  *Hans hat* **ein** *Haus gekauft.* **Der** *Kredit für* **das** *Haus war günstig.*
- Weitere Artikel: Demonstrativart. (z.B. diese, jene, dieselben, solche); Quantifikatoren (z.B. alle, jeder, viele, beide); Negatoren (z.B. kein, keine); Possessivart. (z.B. mein, ihr); Interogativart. (z.B. welche)
- Artikel (wie auch Adjektive) sind typischerweise kongruent zu einem Nomen in Numerus, Genus und Kasus.

# Wortart Präpositionen (Adpositionen)

- weisen Nomen Kasus zu
- **Präpositionen** stehen links (z.B. in, auf, für)
  *nach München, wegen der Kinder*
- Seltenere Adpositionen:
  - manche rechts (z.B. zufolge)
    *seiner Frau zuliebe, den Freunden entgegen*
  - wenige: links und rechts möglich (z.B. wegen)
  - manche umschließen Nomen (z.B. um . . . willen)
    *um der Liebe willen, von Gesetzes wegen*

# Wortart Konjunktionen

- **Konjunktionen** verbinden syntaktische Einheiten der gleichen syntaktischen Kategorie (Sätze, Phrasen, Wörter, Wortteile); (z.B. und, oder, aber, entweder . . . oder)
- geschlossene Liste
- **Koordinierende** (nebenordnende) Konjunktionen:
  *Er schläft,* **aber** *sie arbeitet noch.*
- **Subjunktionen** (satzeinbettende Konjunktionen):
  dass, weil, obwohl, . . .
  **Weil** *er berühmt ist, lassen sie ihn durch.*

# Wortart Interjektionen (Satzwörter)

- Syntaktisch unverbundene, satzwertige Äußerungen.
- Drücken Empfindung, Bewertung oder Willen des Sprechers aus (z.B. aha, igitt, richtig, ja, nein, Danke)
- Übermitteln Aufforderung oder Signal der Kontaktaufnahme (z.B. Hallo, Prost, Hey)

# Wortart Partikel

- Übernehmen lediglich syntaktische oder pragmatische Hilfsfunktionen.
- Bilden keine eigene Phrase.
- Beispiele:
  *Das kann man* **aber** *so nicht sagen.*
  *Das ist* **halt** *so.*
  **am** *schönsten*
  **zu** *schnell*
- Lassen sich oft schwer in ein Schema einordnen.

# Das Partizip zwischen Verb und Adjektiv

- Partizipien verhalten sich wie Verben, denn
  - sie können Akkusativ zuweisen
  - sie "erben" die Argumentstruktur des Verbs, aus dem sie abgeleitet werden.

  *die Tätigkeit befriedigt mich* $\Rightarrow$ *eine mich befriedigende Tätigkeit*
  *der Schüler liest das Buch* $\Rightarrow$ *der das Buch lesende Schüler*
- Partizipien verhalten sich wie Adjektive, denn sie flektieren wie Adjektive (können aber oft nicht prädikativ verwendet werden)
  *die befriedigend***en** *und nützlich***en** *Tätigkeiten*
  *Freude an befriedigend***er** *und nützlich***er** *Tätigkeit*
  **Die Frau ist laufend.*
- Konvention: Dem Partizip wird die Wortart *Verb* zugewiesen (Wortart vs. syntaktische Funktion).

# Outline

# Schwierigkeiten bei der Wortartenzuweisung

Wortartwechsel

- Leid (vgl. z.B.: Das tut mir leid) (Nomen $\Leftrightarrow$ Verbpartikel)
- Klasse (vgl. z.B.: ein klasse Buch) (Nomen $\Leftrightarrow$ Adjektiv)
- ja (vgl. z.B.: Das war ein klares Ja) (Satzwort $\Leftrightarrow$ Nomen)

Zugehörigkeit zu mehreren Wortarten / Ambiguität

- Er las, **aber** er war sehr unkonzentriert (Konj.)
- Das kann man **aber** so nicht sagen (Partikel)

# Schwierigkeiten bei der Wortartenzuweisung

z.B. Zahlwörter

- *eins/ein/eine...*: deklinierbar (ähnlich zu Determinerern oder Pronomen?)
- *zwei* auch deklinierbar: z.B. der Bund zweier Kaiser
- *tausend* (ebenso)
- Million: eher wie Nomen

Sonderfall viel

- Teils wie Determinierer:
  *Vieles Erfreuliche stand in dem Brief*
  *Er trank viel Bier*
- Teils wie Adjektiv:
  *viele Tiere*
  *die vielen Tiere*
  *das viele Laub*

# Outline

# 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
  - Personalpronomen (z.B. ich, du, er, wir) $\Rightarrow$ rechts davon: Verb
- **Aufgabe: Welche Wortarten könnten die Wörter haben? (und warum)**
  *Das Ützlipütz prümft den pliemen Wenzipil krät.*

# Part-of-Speech Tagging (POS Tagging)

- Tagging manuell oder durch Algorithmen (regelbasierte oder statistische Methoden (z.B. Hidden-Markov-Modelle)
- Programme im Netz:
  - CIS, LMU München: MarMoT
    `cistern.cis.lmu.de/marmot/`
  - CIS, LMU München: TreeTagger
    `www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`
  - Stanford:
    `http://nlp.stanford.edu/software/tagger.shtml`

# Stemming (Stemmatisierung)

- Alternatives Verfahren zum Lemmatisieren
- Flexionsmorpheme von Wortform werden beseitigt
  $\Rightarrow$ Wortstamm (wird der Wortform zugeordnet)
- z.B. engl. Wortformen *process*, *processing*, *processed*
  $\Rightarrow$ Stamm *process*
- Problem: sinnvolle Unterscheidungen können verloren gehen:
- z.B. *stocks* (Aktien etc.) und *stockings* ('Strümpfe' etc.)
  $\Rightarrow$ Stamm *stock*
- Programm in Netz z.B. Porter Stemmer:
  Porter-Stemmer (Demo und download):
  `http://snowball.tartarus.org`