

Übersicht

3 Syntaktische Kategorien

3.1 Kategorisierung syntaktischer Einheiten

3.2 Lexikalische Kategorien

3.2.1 Wortarten-Analyse

3.2.2 Wortarten des Deutschen

3.3 Methoden zur Identifizierung von Konstituenten

3.3.1 Substitutionstest

3.3.2 Permutationstest

3.3.3 Eliminierungstest

3.3.4 Koordinationstest

3.4 Phrasenkategorien

3.4.1 Eigenschaften von Phrasen

3.4.2 Phrasenkategorien des Deutschen

3.5 Tagsets und Formate

3.5.1 Part-of-Speech-Tagsets

3.5.2 Syntaktische Tagsets

3.5.3 Repräsentationsformate Part-of-Speech-Tags

3.5.4 Repräsentationsformate syntaktischer Tags

3 Syntaktische Kategorien

3.1 Kategorisierung syntaktischer Einheiten

Analysemethoden syntaktischer Einheiten

- **Segmentierung**

- Zerlegung Satz in **Konstituenten**: Wörter, Phrasen, Teilsätze
- Analyse **syntagmatischer Beziehung zwischen Einheiten**
- feststellbar über **Konstituententests**

- **Klassifizierung / Kategorisierung**

- Bildung von **Mengen mit gleichen Eigenschaften** (Klassen)
- morphologische, syntaktische und semantische **Kriterien**
- syntaktisch: **Austauschbarkeit im gleichen Kontext**
- Analyse **paradigmatischer Beziehung zwischen Einheiten**

Syntagmatische vs. paradigmatische Dimension

(= lineare Kombinierbarkeit vs. vertikale Austauschbarkeit)

der $\left\{ \begin{array}{l} \underline{\text{gro\ss e}} \\ \underline{\text{kleine}} \end{array} \right\}$ *Hund jagt* $\left\{ \begin{array}{l} \underline{\text{die Katze.}} \\ \underline{\text{Ferdinand.}} \end{array} \right\}$

Ein gro\ss er $\left\{ \begin{array}{l} \underline{\text{Hund}} \\ \underline{*Ferdinand} \end{array} \right\}$ *jagt die Katze.*

$\left\{ \begin{array}{l} \underline{\text{Sie}} \\ \text{die sich im Gehen nach dem Hund umschauende } \underline{\text{Frau}} \end{array} \right\}$ *st\ss rzt.*

Er $\left\{ \begin{array}{l} \underline{\text{sieht}} \text{ einen Hund} \\ \underline{\text{geht}} \end{array} \right\} \text{ auf dem Weg.}$

Er $\left\{ \begin{array}{l} \underline{\text{sieht}} \\ \underline{* \text{geht}} \end{array} \right\} \text{ einen Hund.}$

Motivation für Klassifizierung in Syntaxanalyse:

- Beschreibung der hierarchischen syntaktischen Struktur über **wort- und phrasenklassenbasierte Schemata**
 - ökonomisch: viele Satz schemata durch wenige Regeln generierbar
 - beschreibungsadäquat: Phrasen empirisch feststellbar
- **syntaktische Regeln**
 - Regeln der Kombination von **Klassen sich syntaktisch gleichverhaltender Wörtern** (lexikalische Kategorien) zu Phrasen und Sätzen
- **lexikalische Regeln**
 - Zuordnung lexikalischer Einheiten zu ihren lexikalischen Kategorien

- **traditionelle Grammatik (siehe unten) = reine Wortarten-Syntax**

→ ohne Phrasenebene: sehr viele Satzschema: $S \rightarrow \left\{ \begin{array}{l} \text{Satzschema 1} \\ \dots \end{array} \right\}$

Auflistung 1: Syntaktische und Lexikalische Regeln

```
1 ##### Syntaktische Regeln #####
2         S → NP VP
3         PP → P NP
4         NP → Det N | N
5         VP → V NP | VP PP
6
7 ##### Lexikalische Regeln #####
8         Det → 'an' | 'my'
9         N → 'elephant' | 'pajamas' | 'I'
10        V → 'shot'
11        P → 'in'
```

Auflistung 2: Generierung Satz schemata

```
1  # http://www.nltk.org/howto/generate.html
2  grammar = nltk.CFG.fromstring("""
3      S → NP VP
4      PP → P NP
5      NP → Det N | Det N PP | N
6      VP → V NP | VP PP
7      Det → 'Det'
8      N → 'N'
9      V → 'V'
10     P → 'P'
11     """)
12
13  from nltk.parse.generate import generate
14  for sentence in generate(grammar, depth=6):
15      print(' '.join(sentence))
```

12

```
32 # Det N P N V N P Det N
33 # Det N P N V N P N
34 # N V Det N
35 # N V N
36 # N V Det N P Det N
37 # N V Det N P N
38 # N V N P Det N
39 # N V N P N
40
41 len(list(generate(grammar, depth=6)))
42 #24
43 len(list(generate(grammar, depth=7)))
44 #64
45 len(list(generate(grammar, depth=8)))
46 #408
47
```

```
48 for sentence in generate(grammar, n=6):
49     print(' '.join(sentence))
50
51 #Det N V Det N
52 #Det N V Det N P Det N
53 #Det N V Det N P Det N P Det N
54 #Det N V Det N P Det N P Det N P Det N
55 #Det N V Det N P Det N P Det N P Det N P Det N
56 #Det N V Det N P Det N P Det N P Det N P Det N
    P Det N
```

Syntaktischer Grundfunktionen - prototypische Wortart

- **Prädikat**

- Bezeichnung von Sachverhalten (Handlungen, Ereignisse, Zustände)

- prototypische Wortart: **Verb**

- **Argument (auch: Komplement, Ergänzung)**

- Referenz auf im Sachverhalt beteiligte Sache (Person, Ort, Ding)

- prototypische Wortart: **Nomen**

- **Modifikator**

- optionale Bedeutungshinzufügung (Eigenschaften, Umstände)

- prototypischer nominaler Modifikator: **Adjektiv** (= Attribut)

- prototypischer verbaler Modifikator: **Adverb** (= Adjunkt/Angabe)

Beispiel: *Peter* (Arg.) *kauft* (Präd.) *bald* (Adjunkt) *ein schnelles* (Attr.) *Auto* (Arg.)

Syntaktische vs. semantische Kategorisierung

- traditionell Grammatik: **semantische Wortklassifizierung**
→ z. B.: Nomen, von lat. *nomen*: Namen einer Sache/Person/Ort usw.
Adjektiv: Eigenschaftswort
- **keine direkte Entsprechung Semantik : syntaktische Funktion**
→ z. B.: prototypisches Nomen kann syntaktisch Teil des Prädikats sein, also eine andere syntaktische Funktion erfüllen (Prädikativum):
Er ist Lehrer.
→ z. B.: Wörter mit nicht-nominaler Semantik können die prototypische nominale Strukturposition einnehmen (als Argument fungieren): *Blau ist eine Farbe.*

- Wortarten sind **sprachabhängig**
 - es gibt Sprachen, die keine Eigenschaftswortklasse haben (Dyirbal, Lakota; s. VanValin 2000, 12)
 - die typische syntaktische Funktion, die in indogermanischen Sprachen Adjektive übernehmen (Attributfunktion), wird hier von Nomen (Dyirbal) bzw. Verben (Lakota) übernommen
- moderne Linguistik: **Definition Wortklassen über morphosyntaktische Eigenschaften**
- Bestimmung Klassenmitglieder über **syntaktisches Verhalten**:
 - Generative Grammatik: **Besetzung gleicher Strukturpositionen**
 - Strukturalismus: **Auftreten in gleichen Kontexten** (distributionsäquivalent)
 - Distribution = Menge der Kontexte

Auflistung 3: *Distributionsanalyse*

```
1 #siehe http://www.nltk.org/book/ch05.html
2 import nltk
3 text = nltk.Text(word.lower() for word in
    nltk.corpus.brown.words())
4
5 text.similar('woman')
6 #man day time year car moment world family
    house boy child country job state girl place
    war way case question
7
8 text.similar('bought')
9 #made done put said found had seen given left
    heard been brought got set was called felt
    in that told
```

3.2 Lexikalische Kategorien

3.2.1 Wortarten-Analyse

- Klassifikation von Wörter nach **morphologischen, syntaktischen oder semantischen Kriterien**
- Wort = **atomare syntaktische Einheit**
→ terminale Konstituenten im Syntaxbaum
- Wortklasse = **Wortart** = **Part-of-Speech** = **lexikalische Kategorie**
→ präterminale Konstituenten im Syntaxbaum

- **semantisches Kriterium:** Differenzierung Wörter über ihre **Bedeutung**
- **morphologisches Kriterium:** Differenzierung Wörter über die Art ihrer **Flexion / Derivation**

→ **Flexionsparadigmen:** $\left\{ \begin{array}{c} \text{Tür} \\ \text{Welt} \end{array} \right\} -en$ vs. $\left\{ \begin{array}{c} \text{geh} \\ \text{steh} \end{array} \right\} -e/st/t$ (*Welt-st)

→ **Derivationsmorphologie:** $\left\{ \begin{array}{c} \text{new} \\ \text{beautiful} \end{array} \right\} -ly$

(Adjektive bilden in Kombination mit -ly Adverbien)

- **syntaktisches Kriterium:** Differenzierung Wörter über **Distribution**
 - Auftreten in gleichen Kontexten (distributionsäquivalent)
 - z. B.: Adjektiv zwischen DET und NOUN oder nach Form von *sein*

- **Differenzierungen:**
 - **Auto- vs. Synsemantika**
 - **Inhaltswörter:** selbständige lexikalische Bedeutung; satzgliedfähig (Funktion als Phrasenkopf)
 - **Funktionswörter:** grammatische Bedeutung (abhängig von Bezugswort); nicht satzgliedfähig
 - **offene vs. geschlossene Klassen**
 - endliche/abgeschlossene vs. potentiell unendliche Menge
 - Bildung neuer Wörter u.a. durch Derivationsregeln
- **historisch: Acht-Wortarten-Lehre (Dionysios Thrax, 2. Jhd. v. Chr)**
 - Nomen, Verb, Partizip, Artikel, Pronomen, Präposition, Adverb und Konjunktion

3.2.2 Wortarten des Deutschen

Lexikalische Hauptkategorien (Inhaltswörter):

Nomen (NOUN, NN, N):

- offene Klasse; bezeichnet Lebewesen, Sachen (Dinge), Begriffe (Abstrakta), Individuen, Eigenschaften
- deklinierbares Wort
- minimaler Bestandteil eines Arguments (verbalen Komplements)
- Subklassen: Substantive, Eigennamen (*proper nouns* NNP), nominalisierte Adjektive
- Beispiele: *Mensch; Ferdinand; (das) Gute*

Verb (VERB, VB, V):

- offene Klasse; bezeichnet Zustände, Vorgänge, Tätigkeiten, Handlungen
- konjugierbares Wort
- minimaler Bestandteil des Satzprädikats
- Subklassen (nach Anzahl der Argumente): intransitiv (1), transitiv (2), ditransitiv (3)
- Beispiele: *gehen, sehen, geben*

Adjektiv (ADJ , JJ):

- offene Klasse; bezeichnet Eigenschaften und Merkmale
- deklinierbar (im attributiven Gebrauch) und komparierbar
- attributiver Gebrauch: *der kleine Junge* (modifiziert Nomen)
- prädikativer Gebrauch: *der Junge ist klein*
- adverbialer Gebrauch: *Er singt laut*

Adverb (ADV , RB):

- offene Klasse; bezeichnet nähere Umstände
- nicht flektierbares Wort
- modifiziert Verben, Sätze, Adjektive und Adverbien
- Beispiele: *hier, bald, gern, wohl*

Nominale Begleiter und Proformen (Funktionswörter):

Pronomen (PRON, PR):

- geschlossene Klasse; Verweis / Referenz / nähere Bestimmung
- deklinierbares Wort, das eine Nominalphrase vertritt (Proform)
- gleiche syntaktische Distribution wie Nomen, andere Semantik
→ anaphorischer oder deiktischer Bezug (Kotext vs. Kontext)
- Funktionswort, aber satzgliedfähig (im selbstständigen Gebrauch)
- Personal- (PRP), Indefinit-, Demonstrativ- und Fragepronomen
- Beispiele: *er / du / einer / dieser / wer (geht)*

Determinativ (DET, DT):

- geschlossene Klasse; Verweis / Referenz / nähere Bestimmung
- nominaler Modifikator (nur ein Determinativ pro NP):
 - Artikel (Definitheitsmarker):
 - * bestimmt = vorerwähnt/bekannt
 - * unbestimmt = neu/unbekannt
 - Quantifizierer
 - attributiv gebrauchte Pronomen: Possessiv- (PRP\$), Reflexiv-Demonstrativ- und Fragepronomen
- Beispiele: $\left\{ \begin{array}{l} \text{der} \\ \text{ein} \end{array} \right\} \text{Hund}, \left\{ \begin{array}{l} \text{alle} \\ \text{keine} \end{array} \right\} \text{Hunde}, \left\{ \begin{array}{l} \text{dieser} \\ \text{euer} \end{array} \right\} \text{Hund}$

Weitere lexikalische Kategorien (Funktionswörter):

Adposition (ADP):

- geschlossene Klasse; bezeichnet Verhältnisse, Beziehungen
- Präposition (P, IN, APPR) oder Postposition (APP0)
- Beispiele: *wegen (Unwetter), auf (dem Dach); (der Uhrzeit) halber*

Konjunktion (CONJ):

- geschlossene Klasse; bezeichnet Verknüpfungen im logischen, zeitlichen, begründenden, modalen u. ä. Sinn
- verbindet gleichartige Konstituenten
- koordinierende (CCONJ) und subordinierende (SCONJ) Konjunktionen
- Beispiele: *und, aber, weil*

Partikel (, die) (PRT, RP):

- geschlossene Klasse; bezeichnet die Sprechereinstellung, -bewertung
- Negationspartikel: *nicht*
- Intensitätspartikel: *zu, sehr, wenig*
- Modalpartikel / Abtönungspartikel (Sprechereinstellung): *schon, ja, einfach, doch, bloß*
- Diskurspartikel (Gesprächssteuerung): *also, ähm*
- Ausdruckspartikel (Interjektion, satzwertig): *oh, juhu!*

Auflistung 4: POS-Tagging mit NLTK

```
1 #siehe http://www.nltk.org/book/ch05.html
2 import nltk
3
4 text = word_tokenize("They refuse to permit us
   to obtain the refuse permit")
5 nltk.pos_tag(text)
6 # [('They', 'PRP'), ('refuse', 'VBP'), ('to',
   'TO'), ('permit', 'VB'), ('us', 'PRP'),
   ('to', 'TO'), ('obtain', 'VB'), ('the',
   'DT'), ('refuse', 'NN'), ('permit', 'NN')]
```

3.3 Methoden zur Identifizierung von Konstituenten

- **Konstituenten = syntagmatische Einheiten** (Wörter, Phrasen, Teilsätze)
- Feststellung durch **Konstituententests**
- Hauptkriterien: **Verschiebbarkeit und Ersetzbarkeit** unter Erhalt der **Grammatikalität**
- Konstituenten-Hierarchie: **unmittelbare vs. mittelbare Konstituenten**

3.3.1 Substitutionstest

- auch: **Ersatzprobe**
 - Eine unter Erhalt der Grammatikalität des Satzes durch eine andere Wortfolge ersetzbare Wortfolge ist Konstituente
- **Feststellung von Einheiten** (Austauschbarkeit im gleichen Kontext; allgemeiner: auch für Wortarten, Flexionsparadigmen)
- **Pronominalisierung**
 - Eine durch Pronomen ersetzbare Wortfolge ist Konstituente
- für **Feststellung Verb mit Erweiterungen (komplexe VP) als Konstituente**
 - Ersatz komplexer VP durch (intransitives) Verb

Anwendung Substitutionstest

Beispiel Dürscheid 2010, Kap. 3.6:

Der Junge verkauft die Äpfel des Bauern.

Identifizierung der Wortfolgen *der Junge* und *die Äpfel des Bauern* als Konstituenten durch **Pronominalisierung**:

Er verkauft sie.

Identifizierung der Wortfolge *verkauft die Äpfel des Bauern* als **komplexe VP-Konstituente** durch Ersatz mit intransitivem Verb:

Der Junge arbeitet.

3.3.2 Permutationstest

- auch: **Verschiebeprobe**
 - Im Satz ohne Zerstörung der Grammatikalität verschiebbare Wortfolge ist Konstituente
- im Deutschen: Verschiebung vor finites Verb
 - **Topikalisierung**: rhetorisch-pragmatische Funktion
- verwendet zum **Testen von Wortstellungsmöglichkeiten**
- **Auflösung von Ambiguität**:

Der Junge beobachtete das Mädchen mit dem Fernglas. (ambig)

Das Mädchen mit dem Fernglas beobachtete der Junge. (NP-att.)

Mit dem Fernglas beobachtete der Junge das Mädchen. (VP-att.)

Anwendung Permutationstest

Beispiel Dürscheid 2010, Kap. 3.6:

Der Junge verkauft die Äpfel des Bauern.

Identifizierung der Wortfolgen *der Junge* und *die Äpfel des Bauern* als Konstituenten durch **Permutation**:

Die Äpfel des Bauern verkauft der Junge.

3.3.3 Eliminierungstest

- auch: **Weglassprobe**
 - Eine ohne Zerstörung der Grammatikalität eines Satzes weglassbare Wortfolge ist Konstituente
- Feststellung syntaktisch notwendiger bzw. optionaler Einheiten (**Dependenzbeziehungen**)

Anwendung Eliminierungstest

Beispiel Dürscheid 2010, Kap. 3.6:

Der Junge verkauft die Äpfel des Bauern.

Identifizierung der Wortfolge *des Bauern* als **attributive Konstituente** durch Eliminierung:

Der Junge verkauft die Äpfel.

3.3.4 Koordinationstest

- **Koordination:** Verbindung mit *und* / *aber*
→ Eine mit einer anderen Wortfolge unter Erhalt der Grammatikalität des Satzes koordinierbare Wortfolge ist Konstituente
- geeignet für die **Ermittlung von Phrasenteilen** (Attributen usw.)
- Analyse der Struktur von **komplexen Konstituenten**
- Konstituenten müssen vom **gleichen Typ** sein: *Er schrieb einen Brief und eine Karte* und *Er schrieb an dich und an mich*, aber nicht **Er schrieb einen Brief und an mich*.

Anwendung Koordinationstest

Beispiel Dürscheid 2010, Kap. 3.6:

Der Junge verkauft die Äpfel des Bauern.

Identifizierung der Wortfolge *des Bauern* als **Konstituente** durch Koordination:

Der Junge verkauft die Äpfel des Bauern und der Bäuerin

Festgestellte Konstituentenstruktur:

[Der Junge] [[verkauft] [[die Äpfel] [des Bauern]]]

3.4 Phrasenkategorien

3.4.1 Eigenschaften von Phrasen

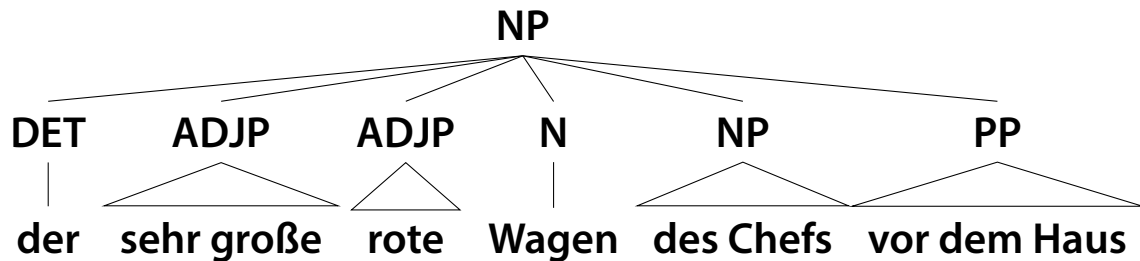
- im gleichen Kontext austauschbare Konstituenten bilden **Konstituentenklasse** → *Phrasenkategorien*
- Phrase = Konstituente, in der ein (Inhalts-)Wort als **Phrasenkopf um Wörter oder Phrasen erweitert** ist
- alle Wörter und Phrasen in der Phrase sind zum Kopf **dependent**
- Kopf vererbt **morphosyntaktische Merkmale** an Phrase (Kasus usw.)
- Kopf steuert **syntaktisches Verhalten** der Konstituente im Satz
- Kopf bestimmt die **Phrasenkategorie** (Wortart X → Phrasenkat. XP)
→ nicht-terminale Knoten im Syntaxbaum

3.4.2 Phrasenkategorien des Deutschen

Nominalphrase NP:

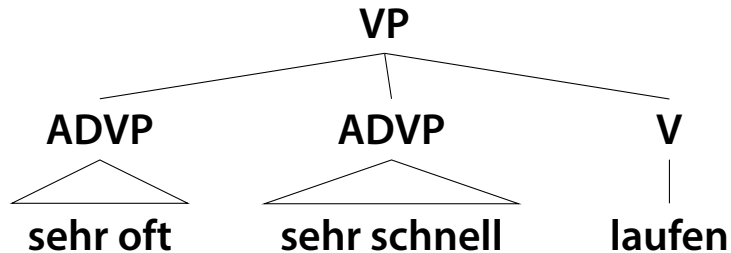
- Nomen als Phrasenkopf
- Beispiel: *der alte Mann*
- Pronomen: vertreten Nominalphrasen
- Phrasenschema NP:

(DET | NP) (ADJP)* N (PP / NP / Relativsatz)*



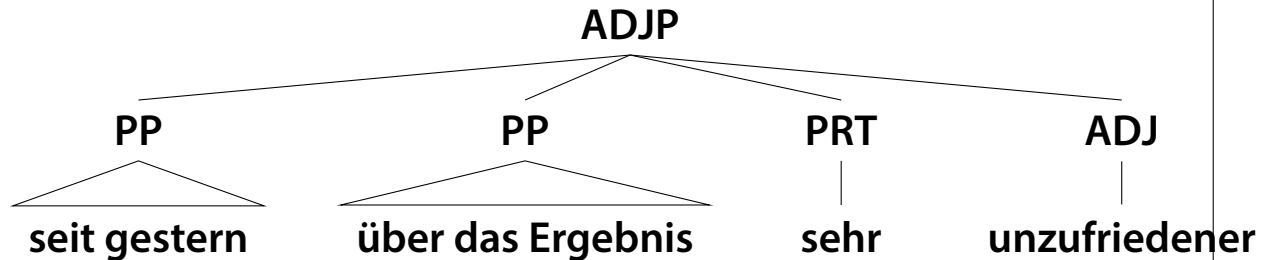
Verbalphrase VP:

- Verb als Phrasenkopf
- Beispiel: *langsam gehen*
- Phrasenschema VP: **(ADVP)* V**



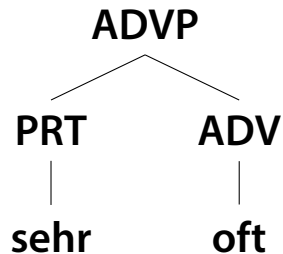
Adjektivphrase ADJP, AP:

- Adjektiv als Phrasenkopf
- Beispiel: *sehr groß*
- Phrasenschema ADJP: (PP)* (PRT) ADJ



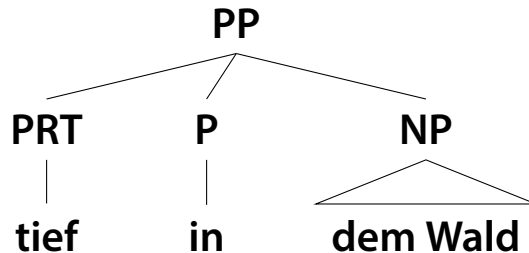
Adverbphrase ADVP , AVP:

- Adverb als Phrasenkopf
- Beispiel: *ziemlich selten*
- Phrasenschema ADVP: (PRT) ADV



Adpositionalphrase / Präpositionalphrase PP:

- Präposition oder Postposition als Phrasenkopf
- Beispiel: *mit dem Fahrrad*
- Phrasenschema PP: (PRT) P NP



3.5 Tagsets und Formate

- **Tagset** = Sammlung von Kategorienlabels
- traditionelle Analysen: wenige lexikalische Kategorien
- in Korpuslinguistik/Computerlinguistik: umfangreichere Tagsets
→ umfassen z. T. auch morphologische Kriterien
- **Wichtige Tagsets:**
 - **Brown Corpus:** 87 POS-Tags
 - **Penn Treebank:** 45 POS-Tags (vereinfachtes Brown Corpus Tagset)
 - **Universal POS-Tagset (UD):** 17 POS-Tags
- (POS-Tags oben: 1. Universal POS Tagset; 2. Penn Treebank / Brown Corpus)

Auflistung 5: *NLTK POS-Tag-Hilfefunktion*

```
1 nltk.help.upenn_tagset('NN.*')
2 # NN: noun, common, singular or mass
3 #   common-carrier cabbage knuckle-duster
4   Casino afghan shed thermostat
5 #   investment slide humour falloff slick
6   wind hyena override subhumanity
7 #   machinist ...
8 # NNP: noun, proper, singular
9 #   Motown Venneboerger Czystochwa Ranzer
10  Conchita Trumplane Christos
11 #   Oceanside Escobar Kreisler Sawyer Cougar
12  Yvette Ervin ODI Darryl CTCA
13 #   Shannon A.K.C. Meltex Liverpool ...
14 # NNPS: noun, proper, plural
```

```
12 nltk.help.brown_tagset('NN.*')
13 # NN: noun, singular, common
14 #     failure burden court fire appointment
      awarding compensation Mayor
15 #     interim committee fact effect airport
      management surveillance jail
16 #     doctor intern extern night weekend duty
      legislation Tax Office ...
17 # NN$: noun, singular, common, genitive
18 #     season's world's player's night's
      chapter's golf's football's
19 #     baseball's club's U.'s coach's bride's
      bridegroom's board's county's
20 #     firm's company's superintendent's mob's
      Navy's ...
```

3.5.1 Part-of-Speech-Tagesets

- Tagset **Browncorpus**:

https://en.wikipedia.org/wiki/Brown_Corpus#Part-of-speech_tags_used

- Tagset der **Penn Treebank** (getaggetes + geparstes Zeitungskorpus):

→ verwendet im englischen Stanford-Parser-Modell

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

- **Universal-Dependencies POS-Tagset**:

<http://universaldependencies.org/u/pos/index.html>

- **Universal Tagset** (vereinfachtes UD-POS-Tagset, s. NLTK-5):

<http://www.nltk.org/book/ch05.html#tab-universal-tagset>

- **STTS** (Stuttgart-Tübingen Tagset)

→ verwendet im deutschen Stanford-Parser-Modell

<http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html>

https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/STTS_Tagset_Tiger

3.5.2 Syntaktische Tagesets

- **Penn-Treebank** (Stanford: english.pcfg)

<http://web.mit.edu/6.863/www/PennTreebankTags.html>

- **STTS gemäß Tiger-Annotationsschema** (Stanford: german.pcfg)

Tiger Corpus ist ein syntaktisch annotiertes deutsches Korpus

[https://www.linguistik.hu-berlin.de/de/institut/professuren/](https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/Tiger_Knotenlabels)

[korpuslinguistik/mitarbeiter-innen/hagen/Tiger_Knotenlabels](https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/Tiger_Knotenlabels)

https://www.linguistics.ruhr-uni-bochum.de/~dipper/pub/tiger_annot.pdf

- (Phrasentags oben: 1. Penn Treebank, 2. STTS/Tiger)

3.5.3 Repräsentationsformate

Part-of-Speech-Tags

- **Auszeichnungssprache (Markup):**
 - XML-TEI-Format (s. nächste Auflistung)
 - Auszeichnung mit Trennerzeichen, z.B. Brown Corpus:
The/at jury/nn further/rbr said/vbd
- **tabular** (Vertikaltext, Tokenlisten): s. Auflistung IOB-Format
- **Tupel-Listen** (z. B. NLTK):
[('the', 'ART'), ('jury', 'N'), ('further', 'ADV'), ('said', 'V')]

Auflistung 6: BNC Sample im XML-TEI-Format (British National Corpus)

```
1 <stext type="OTHERSP">
2   <s n="1">
3     <w c5="ITJ" hw="ah" pos="INTERJ">Ah
4       </w>
5     <w c5="AV0" hw="there" pos="ADV">there
6       </w>
7     <w c5="PNP" hw="we" pos="PRON">we </w>
8     <w c5="VBB" hw="be" pos="VERB">are</w>
9     <c c5="PUN">,</c>
10    <unclear/>
11    <c c5="PUN">.</c>
12  </s>
13 </stext>
```

Auflistung 7: IOB-Format (Ausschnitt CONLL2000-Korpus)

```
1 Balcor NNP B-NP
2 , , O
3 which WDT B-NP
4 has VBZ B-VP
5 interests NNS B-NP
6 in IN B-PP
7 real JJ B-NP
8 estate NN I-NP
9 , , O
10 said VBD B-VP
11 the DT B-NP
12 position NN I-NP
13 is VBZ B-VP
14 newly RB I-VP
15 created VBN I-VP
```

```
16 | . . 0
17 |
18 | Mr. NNP B-NP
19 | Meador NNP I-NP
20 | had VBD B-VP
21 | been VBN I-VP
22 | executive JJ B-NP
23 | vice NN I-NP
24 | president NN I-NP
25 | of IN B-PP
26 | Balcor NNP B-NP
27 | . . 0
```

3.5.4 Repräsentationsformate syntaktischer Tags

- **hierarchisch: Treebanks (Menge von Syntaxbäumen):**

- als Klammerausdrücke (s. unten Penn Treebank)
- XML-annotiert (s. o. BNC-Sample: Satzebene)
- relational (über Phrasen-/Satz-IDs)

<http://www.nltk.org/howto/corpus.html#parsed-corpora>

<http://universaldependencies.org/docs/format.html#syntactic-annotation>

- **IOB-Format (= Inside–Outside–Beginning):**

Tagging Tokens als Beginn bzw. weiterer Teil einer Phrase

Auflistung 8: Penn Treebank (Verwendung in NLTK)

```
1 from nltk.corpus import treebank
2 print(treebank.words('wsj_0003.mrg'))
3 #[('A', 'DT'), ('form', 'NN'), ('of', 'IN'),
   ...]
4 print(treebank.parsed_sents('wsj_0003.mrg')[0])
5 # (S
6 #   (S-TPC-1
7 #     (NP-SBJ
8 #       (NP (NP (DT A) (NN form)) (PP (IN of)
9 #         (NP (NN asbestos)))))
10 #       (RRC ...)...)...)
11 #   ...
12 #   (VP (VBD reported) (SBAR (-NONE- 0) (S
13 #     (-NONE- *T*-1))))
14 #   (. .))
```