Übersicht

2 Einführung

- 2.1 Syntax natürlicher Sprachen
 - 2.1.1 Syntaxbegriff und Grammatikbegriff
 - 2.1.2 Syntaktische Ausdrucksmittel

2.2 Syntaktische Struktur

- 2.2.1 Syntagmatische Relation
- 2.2.2 Grammatische Relationen

2.3 Automatische Syntaxanalyse

- 2.3.1 Formale Grammatiken als Syntaxmodelle
- 2.3.2 Syntaktische Ambiguität
- 2.3.3 Parsing als automatische Syntaxanalyse
- 2.3.4 Computerlinguistische Anwendungen
- 2.3.5 Abbildungen syntaktischer Strukturen

2 Einführung

tet werden können."

2.1 Syntax natürlicher Sprachen

Definition Syntax (nach Bußmann, Lexikon der Sprachwissenschaft): "Teilbereich der Grammatik natürlicher Sprachen (auch: Satzlehre)." "System von Regeln, die beschreiben wie aus einem Inventar von Grundelementen (Morphemen, Wörtern, Satzgliedern) durch spezifische syntaktische Mittel (Morphologische Markierung, Wort- und Satzgliedstel-

lung, Intonation u.a.) alle wohlgeformten Sätze einer Sprache abgelei-

Definition Syntax (nach mediensprache.net/de/lexikon):

"Teilgebiet der Linguistik, das sich mit der Kombination von Wörtern zu komplexen Einheiten (Analyse des Aufbaus von Satzstrukturen und der Zusammenfügung von Wörtern zu größeren Einheiten) beschäftigt, ohne sich für den internen strukturellen Aufbau der Wörter zu interessieren."

"Der Begriff kann auch benutzt werden, um den strukturellen Aufbau eines Satzes zu bezeichnen ('Syntax eines Satzes' und so weiter). "

2.1.1 Syntaxbegriff und Grammatikbegriff

Syntax:

- Etymologie: σύνταξις [syntaksis] = 'Zusammensetzung'
 - \rightarrow aus σύν= 'zusammen', τάξις = 'Ordnung, Reihenfolge'
- allgemein (Semiotik): Syntax als Struktur einer Zeichenfolge
 - → Regeln der Kombination elementarer Zeichen zu komplexen Zeichen
- Syntax natürlicher Sprachen: Struktur von Wortfolgen
 - ightarrow Regeln der Kombination von Wörtern zu größeren Einheiten
- Teil der Grammatik (= Sprachstruktur-(Analyse))

Grammatik:

- Etymologie: (τέχνη) γραμματική [(technē) grammatikē]
 - = 'Schreibkunst' / 'Buchstaben-Fertigkeit'
- Bezug zu strukturellem Aufbau Sprache:
 - → Lautstruktur: **Phonologie**
 - → Wortstruktur: **Morphologie**
 - → Satzstruktur: *Syntax* (Strukturaufbau aus Wörtern)

Grammatikbegriff

- · a) Grammatik als Sprachstruktur
 - → phonologische, morphologische und syntaktische **Regula- ritäten** einer natürlichen Sprache
- b) Grammatik als Theorie der Sprachstruktur
 - → Sprachwissenschaftliche **Beschreibung der Regularitäten** einer natürlichen Sprache (Modell)
- c) Grammatik als Wissen um Sprachstruktur
 - → Wissen des Sprechers um diese Regularitäten

- d) Grammatik als Regelbuch
 - → **Lehrwerk**, das die **Regularitäten** einer natürlichen Sprache enthält
- e) Formale Grammatik
 - → mathematisches Regelsystem einer formalen Sprache, das zur Modellierung der Grammatik einer natürlichen Sprache verwendet werden kann

Abgrenzung Syntax zu anderen Disziplinen:

- Abgrenzung zur Morphologie:
 - Syntax: Analyse des *Strukturaufbaus* sprachlicher Einheiten oberhalb der Wortebene
- Abgrenzung zu Semantik und Pragmatik:
 - Syntax: unabhängig von semantischer Interpretation, vgl. Chomsky 1957, 'Syntactic Structures':
 - 'colorless green ideas sleep furiously'
 - * erfüllt Wohlgeformtheitsbed., ist also grammatisch
 - * aber: hat keine sinnvolle semantische Interpretation

Relevanz der Morphologie für Syntax:

- Wortartenklassifikation:
 - → **Zusammensetzung** syntaktischer Einheiten aus *Klassen* von Wörtern (lexikalische Kategorien)
- Flexionsmorphologie:
 - → Analyse von *Wortformen*, insofern sie für die syntaktische Strukturanalyse relevant sind (*Morphosyntax*)
 - → *Kasus* und *Agreement* als morphologische Ausdrucksmittel syntaktischer Funktionen
 - → formale Repräsentation als **Merkmalstrukturen**

2.1.2 Syntaktische Ausdrucksmittel

- Wortstellung (strukturell):
 - → Markierung syntaktischer Funktion durch *lineare Anordnung*

→ **Beispiel:** Subjekt-Verb-Objekt-Wortstellung:



- **Kasus** (morphosyntaktisch):
 - → morphologische Markierung der syntaktischen Funktion am abhängigen Element
 - \rightarrow Rektion (dependent-marking)
 - → **Beispiel:** Objekt-Kasus-Markierung:



- Kongruenz/Agreement (morphosyntaktisch):
 - → Kongruenz = Übereinstimmung von Merkmalen zwischen abhängigen Elementen
 - → morphologische Markierung der syntaktischen Funktion des abhängigen Elements am Kopf (head-marking)
 - → **Beispiel:** Subjekt-Kongruenz (in Numerus und Person):



- → Verb kongruiert in nominalen Kategorien mit Subjekt-NP
- (Argument für Regel: $S \rightarrow NP VP$ (Subjekt > Objekt))

2.2 Syntaktische Struktur

Definition Satz (nach Lewandowski, Linguistisches Wörterbuch):

"grammatisch, intonatorisch und inhaltlich nach den Regularitäten der jeweiligen Sprache linear und hierarchisch organisierte Einheit als Mittel zu Ausdruck, Darstellung und Appell, zur Kommunikation von Vorstellungen oder Gedanken über Sachverhalte."

Definition Satz (nach mediensprache.net/de/lexikon):

"kleinste (im Blick auf Inhalt, Struktur und Intonation) selbstständige und vollständige sprachliche Äußerung"

Satzstruktur:

- Satz als zentraler Untersuchungsgegenstand der Syntax
 - → sprachliche Form einer Äußerung (Sprachhandlung)
 - → **Beobachtung: lineare Abfolge** von Wörtern (Wortfolge)
 - → **Syntax:** Beschreibung und Analyse der **hierarchischen Struk**tur von Sätzen:
 - des **Aufbaus** einfacher Sätze (*clause*) aus Wörtern und Phrasen (syntaktische Einheiten)
 - der funktionalen Abhängigkeiten zwischen diesen syntaktischen Einheiten
 - der Struktur komplexer Sätze (sentence)

Struktur

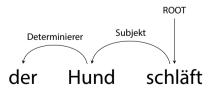
→ *Menge von Relationen*, die zwischen Elementen einer Grundmenge bestehen (Relation: Menge geordneter Paare / Tupel)

Syntaktische Struktur

→ Menge von Relationen, die zwischen Elementen des Lexikons einer natürlichen Sprache (Wörtern) und/oder daraus gebildeten syntaktischen Einheiten bestehen

- Zwei syntaktische Relationstypen:
 - Konstituenz = Teil-Ganzes-Beziehung zwischen Wörtern und aus diesen bestehende syntaktische Einheiten (Syntagmen)
 - Dependenz = Abhängigkeitsbeziehungen zwischen Wörtern (Regens kontrolliert Dependens)
 - * Lexikon: $\{der, die, den, Hund, Katze, jagt, schlaeft\}$
 - * Satzstruktur (Dependenz-Relation):

$$\{(Hund, der), (schlaeft, Hund), (ROOT, schlaeft)\}$$



2.2.1 Syntagmatische Relation

- Konstituenten-Struktur
 - → aus welchen **syntaktischen Einheiten** besteht ein Satz?
- Syntagma: Gruppe sprachlicher Elemente in Äußerung
 - → durch strukturalist. *discovery procedures* (syntaktische Tests): Feststellung von syntakt. Einheiten oberhalb Wortebene und unterhalb Satzebene (**Phrasen / Konstituenten / Satzglieder**)
- Syntax natürlicher Sprachen im Konstituentenmodell:
 - \rightarrow Regeln der (rekursiven) Kombination von Wörtern zu Satzgliedern, einfachen und komplexen Sätzen

shot an elephant in my pajamas

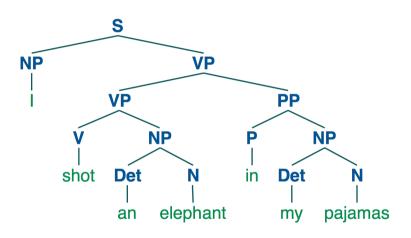


Abbildung 1: Von der Wortfolge zur syntaktischen Struktur (Konstituenzmodell)

2.2.2 Grammatische Relationen

Dependenz-Struktur

→ in welcher **syntaktische Beziehung** stehen Wörter, welche **Funktion** haben sie im Satz?

Funktionale Satzanalyse

- → notwendige und nicht-notwendige Einheiten im Satz
- → Abhängigkeitsverhältnisse zwischen Wörtern
- → Prädikat + Argumente (notwend. Ergänzungen) + Angaben

Syntax natürlicher Sprachen im Dependenzmodell:

→ Regeln der Kombin. von Wörtern nach Abhängigkeitsrelat.

Einführung

shot an elephant in my pajamas

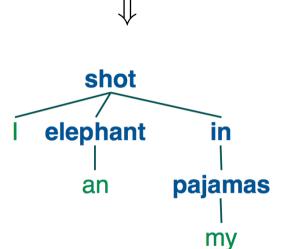


Abbildung 2: Von der Wortfolge zur syntaktische Struktur (Dependenzmodell)

Historischer Hintergrund:

- Paradigma 1: Aristotelische Logik (Begriffslogik)
 - → binäre Struktur von Aussagen: Subjekt-Prädikat (syllogistisches Prädikat = einstelliges Prädikat im Sinne der Prädikatenlogik, s. u.)
 - → Kategorisches Urteil: "Alle Menschen (Subjekt) sind Säugetiere (Prädikat)"
 - \rightarrow über Logik von Port-Royal (1662) beeinflusst strukturalistische Distributions analyse (Saussure, Bloomfield)
 - → Chomsky (1957, 'Syntactic Structures'): mathematische Modellierung mit kontextfreien Grammatiken

I shot an elephant in my pajamas

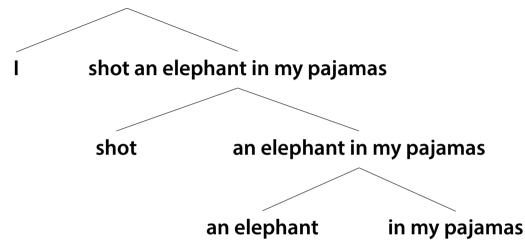


Abbildung 3: binäre Zergliederung in unmittelbare Konstituenten (immediate constituents; IC)

Paradigma 2: Prädikatenlogik (Frege)

- → mehrstellige Prädikate
- → Verb als Satzzentrum: Prädikat + Argumente
- → Vorläufer: Sanskrit-Grammatiker Panini (5./4. Jhd. v. Chr.)
- → Schulgrammatik: implizit dependenzbezogen: Analyse grammatischer Funktionen wie Subjekt, Objekt
- → Valenz-/Dependenzgrammatik: verallgemeinerte dependenzbezogene Syntaxtheorie (Tesnière 1959, 'Éléments de syntaxe structurale')

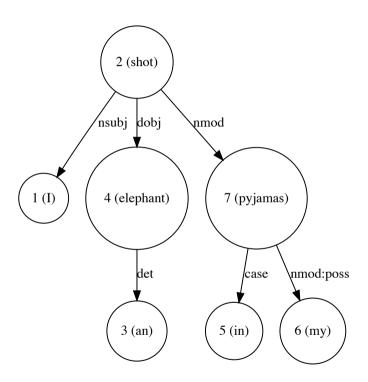


Abbildung 4: Syntaktische Struktur mit gelabelten Relationen (Dependenzmodell)

2.3 Automatische Syntaxanalyse

- Was muss ein Grammatikmodell für die Analyse der syntaktischen Struktur einer natürlichen Sprache leisten?
 - 1. **Strukturerkennung**: ist ein Satz wohlgeformt?
 - Ziel: Erkennung genau der grammatisch korrekten Sätze

Strukturwiedergabe: wie ist die ein Satz aufgebaut?

- Ziel: linguistisch adäquate Strukturanalyse

Möglichkeiten der Syntaxanalyse:

1. Beschreibung des Sprachsystems

→ traditionelle Buch-Grammatik; nicht-computational

2. Aufzählung aller grammatischen Sätze

- → Problem 1: natürliche Sprachen sind unendlich
- → Problem 2: Struktur nicht repräsentiert

3. Beschreibung durch formale Grammatik

- → mathematisches Modell des syntaktischen Regelsystems (computational)
- → ermöglicht die Analyse der Struktur einer unendlichen Menge an Sätzen mit endlichen Mitteln

2.3.1 Formale Grammatiken als Syntaxmodelle

Formale Grammatik:

- Formales Regelsystem zur eindeutigen Beschreibung und Erzeugung einer formalen (!) Sprache
 - → **Generierung aller wohlgeformten Sätze** (= Sprache)
 - \rightarrow generative Grammatik
- kann auch als Modell zur Erkennung und Wiedergabe der syntaktischen Struktur *natürlicher* Sprachen verwendet werden

- Formale Sprache = Menge aller aus Grundsymbolen (z. B. $\{a, b, c\}$) mit den Grammatikregeln ableitbaren formalsprachlichen Wörter (z. B. $\{a, aa, aba, ...\}$)
 - → in Analyse Syntax **natürlicher Sprache**:
 - Grundsymbole sind Wörter des Lexikons: $\{die, der, den, Hund, Katze, jagt\}$
 - die aus der entsprechenden formalen Grammatik als Syntaxmodell ableitbaren formalsprachlichen Wörter sind natürlichsprachliche Sätze:
 - $\{der Hundjagt die Katze, die Katze jagt den Hund, ...\}$

Aufbau einer formalen Grammatik:

- 1. Startsymbol (S)
- 2. Nichtterminalsymbole (z.B. $\{X,Y\}$ oder $\{NP,VP,N,Det,V\}$)
 - → Metasymbole; kommen nur in Zwischenschritten der Ableitung eines Wortes vor
- 3. Terminalsymbole (z.B. $\{a, b, c\}$ oder $\{der, Hund, ...\}$)
 - → Terminalalphabet (Lexikon); Terminalsymbole können nicht weiter ersetzt werden

- 4. Produktionsregeln (z.B. $X \rightarrow aY$ oder $S \rightarrow NPVP$)
 - → Ersetzungregeln; geben an, wie aus Symbolfolgen (beginnend mit S) neue Folgen (Wörter) gebildet werden können
 - → durch Einschränkungen der Regeln ergeben sich Sprachen verschiedener Komplexität (Chomsky-Hierarchie)

Kontextfreie Grammatik:

- Phrasenstrukturgrammatik im engeren Sinne
- CFG-Einschränkung:
 - \rightarrow links nur ein Nichtterminalsymbol: $S \rightarrow NPVP$
 - → Ersetzung unabhängig von Kontext (Kontextfreiheit)
- syntaktische Regeln: NP → Det N (PP)
 - \rightarrow links: syntaktische Kategorien (Phrasen/Satzknoten: S, NP, VP)
 - → rechts: obligatorische und optionale Nichtterminale (syntaktische + lexikalische Kategorien)
 - Rekursion: $NP \rightarrow Det N (PP), PP \rightarrow P NP$ die Katze auf dem Ast in dem Baum auf dem Berg ...

- lexikalische Regeln (Präterminal \rightarrow Terminal):
 - → Zuordnung lexikalische Kategorien/Wortarten zu Lexemen
 - → Präterminale/POS-Label als Untermenge der Nichtterminale

 $N \rightarrow 'Hund'$

 $N \rightarrow 'Katze'$

 $Det \rightarrow 'der' \mid 'die'$

- Wortarten (=lexikalische Kategorien): → Präterminale
- **Lexeme:** → *Terminale (Alphabet)*

Auflistung 1: *Kontextfreie Grammatik*

```
1
     #syntaktische Regeln:
 2
                    S \rightarrow NP VP
 3
                    PP \rightarrow P NP
 4
                    	exttt{NP} 
ightarrow 	exttt{Det} 	exttt{N} 	exttt{ | Det} 	exttt{N} 	exttt{ PP} 	exttt{ | 'I'}
 5
                    VP \rightarrow V NP \mid VP PP
 6
     #lexikalische Regeln:
 7
                    Det \rightarrow 'an' | 'my'
                    N 
ightarrow 'elephant' | 'pajamas'
 8
 9
                    V \rightarrow 'shot'
10
                    P \rightarrow 'in'
```

Klassifizierung syntaktischer Modelle:

- modellierte Relation
 - → Konstituenzgrammatik : Dependenzgrammatik
- Kategorien
 - → atomare Kategorien : komplexe Merkmalstrukturen
- Komplexität der Grammatik (Chomsky-Hierarchie)
 - \rightarrow regulär : kontextfrei : kontext-sensitiv : rekursiv aufzählbar
- Analysetiefe der Grammatik (Rekursion?)
 - \rightarrow flach : verschachtelt

Vorteile Modellierung mit formalen Grammatiken:

- mathematisches Modell:
 - → unendliche Menge an Sätzen **mit endlichen Mitteln beschreib**bar
 - → rechnergestützt verarbeitbar durch Parsingalgorithmen
 - → Beantwortung Fragen zur **Komplexität** natürlicher Sprache (ist jede natürliche Sprache kontextfrei?)
 - → psycholinguistische Anwendung: Parser als **Modell mensch**licher Sprachverarbeitung

Nachteile:

- Probleme mit struktureller Ambiguität
 - \rightarrow wie Entscheidung für richtige (im Kontext intendierte) syntaktische Analyse?
- Probleme mit Übergenerierung
 - → wie Vermeidung Produktion ungrammatischer Sätze?
- keine vollständige Beschreibung möglich
 - → immer nur Ausschnitt natürlicher Sprache modelliert

2.3.2 Syntaktische Ambiguität

strukturelle Ambiguität

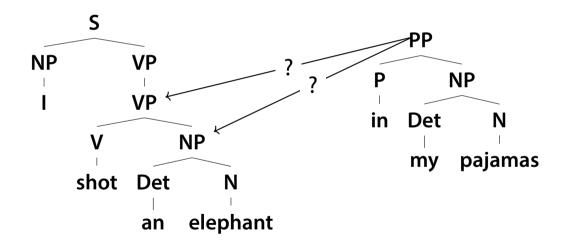
→ mehr als eine Strukturanalyse möglich

Lösung

- → Disambiguierung struktureller Ambiguität durch **PCFGs** (mit statistischen Informationen zu Regelwahrscheinlichkeiten angereicherte CFGs)
- \rightarrow weitere Disambiguierung durch **Lexikalisierung** (erfasst z.B. die Präferenz für PP-Attachment an VP bei setzen/stellen/legen-Verben)

Typ 1: Attachment-Ambiguität: Konstituente kann im Parsebaum an mehr als einer Stelle angebunden werden

Beispiele: Präpositionalphrasen, Adverbialphrasen (s. Übung)



Typ 2: Koordinierungsambiguität:

- [alte [Männer und Frauen]]
- [alte Männer] und [Frauen]]

Typ 3: temporale Ambiguität (garden-path-Sätze):

[The old man] [the boat].

VS

[The old] [man the boat].

[The horse] [raced past the barn] fell.

VS

[The horse [raced past the barn]] [fell].

2.3.3 Parsing als automatische **Syntaxanalyse**

Parsing:

- formale Grammatik:
 - \rightarrow Syntaktisches Strukturmodell, das aber nicht mehr ist als eine Sammlung von Strings (Regeln)
 - → Verfahren notwendig, um zu entscheiden, ob eine Eingabe gemäß einer gegebenen formalen Grammatik wohlgeformt ist

Parsing-Algorithmen:

→ Verfahren zur Verarbeitung von formalen Grammatiken zur Strukturerkennung und -Analyse der Eingabe (Satz als Tokensequenz)

Strukturerkennung:

 \rightarrow Überprüfung der grammatischen Struktur einer Eingabe als Suche einer Ableitung aus den Regeln einer formalen Grammatik (ob Satz in formaler Sprache enthalten ist)

Strukturzuweisung:

→ gleichzeitig Wiedergabe der in der Suche aufgebauten grammatischen Struktur der Eingabe (Syntaxbaum)

Parsing-Algorithmen:

- top-down vs. bottom-up
 - → von Startsybol zu Blättern oder umgekehrt
- Parsingalgorithmen mit dynamischer Programmierung
 - → effizienter Umgang mit Ambiguität/Übergenerierung
- Unifikation
 - \rightarrow Verarbeitung Merkmalstrukturen
- statistische Algorithmen
 - \rightarrow Viterbi, Inside-Outside
- Partielles Parsing / Chunk-Parsing
 - \rightarrow Parsing as Tagging (reguläre Grammatik oder classifier)

Auflistung 2: Tracing CFG Chart-Parsing

```
for tree in parser.parse(sent):
         print(tree)
|. I . shot. an .eleph. in . my .pajam.|
                                                  .| [0:1] 'I'
                                                   .| [1:2] 'shot'
                                                   .| [2:3] 'an'
                                                    .| [3:4] 'elephant'
                                                    .| [4:5] 'in'
                                                    .| [5:6] 'my'
                                            [----]| [6:7] 'pajamas'
                                                    .| [0:1] NP \rightarrow 'I' *
                                                    .| [0:1] S \rightarrow NP * VP
                                                    .| [1:2] V \rightarrow 'shot' *
                                                    .| [1:2] VP \rightarrow V * NP
                                                    .| [2:3] Det \rightarrow 'an' *
                                                    .| [2:3] NP \rightarrow Det * N
                                                    .| [2:3] NP \rightarrow Det * N PP
                                                    .| [3:4] N \rightarrow 'elephant' *
                                                    .| [2:4] NP 
ightarrow Det N *
                                                    .| [2:4] NP 
ightarrow Det N * PP
```

parser = nltk.ChartParser(grammar,trace=1)

```
. | [2:4] S \rightarrow NP * VP
                                 .| [1:4] 	exttt{VP} 
ightarrow 	exttt{V} 	exttt{NP} *
                                 .| [1:4] VP \rightarrow VP * PP
                                 .| [0:4] S 
ightarrow NP VP *
         . . [----] . . . [4:5] P 
ightarrow 'in' *
                                 .| [4:5] PP \rightarrow P * NP
            . . [----]
                                .| [5:6] Det 
ightarrow 'my' *
           . . [----→
                                 .| [5:6] NP 
ightarrow Det * N
                     [---→ .| [5:6] NP \rightarrow Det * N PP
                . . [----]| [6:7] N 
ightarrow 'pajamas' *
                       [-----]| [5:7] NP 
ightarrow Det N *
                . [------] [5:7] NP 	o Det N * PP
        . . . [------] [5:7] S \rightarrow NP * VP
                \lceil ----- \rceil \mid \lceil 4:7 \rceil \text{ PP } \rightarrow \text{P NP } *
          [-----]| [2:7] NP 
ightarrow Det N PP *
     \lceil ----- \rceil \mid \lceil 1:7 \rceil \ \ VP \ 	o \ \ VP \ \ PP \ *
     [-----]| [1:7] VP \rightarrow V NP *
     | [ = = = = = = = = = = = = = = ] | [0:7] S \rightarrow NP VP *
```

2.3.4 Computerlinguistische Anwendungen

Voraussetzungsschritte für automatische Syntaxanalyse:

- Sentence Segmentation
 - ightarrow Liste von Strings
- Tokenisierung
 - \rightarrow Liste von Stringlisten
- Part-of-Speech-Tagging
 - → Liste von Tupellisten (token,pos-tag)
- (Stemming, morphologisches Parsing [Kasus, Agreement])

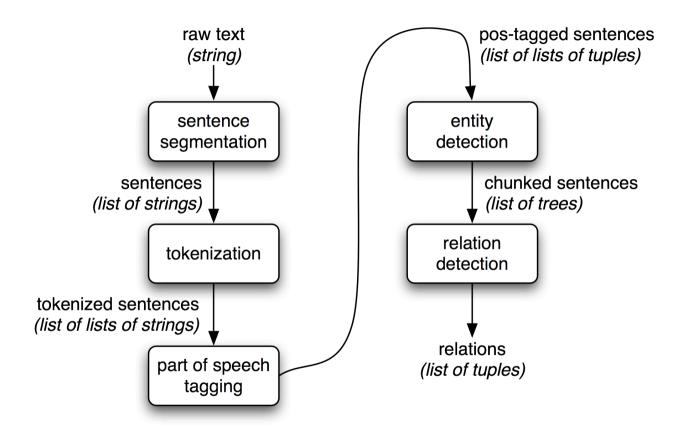


Abbildung 5: Parsing in NLP-Pipeline (Information Extraction), http://www.nltk.org/book/ch07.html # fig-ie-architecture

Mögliche Folgeanwendungen:

- Entity Extraction
 - \rightarrow Liste von Bäumen
- Relation Extraction
 - \rightarrow Liste von Tripeln: (entity, relation, entity)
- Semantic Role Labelling
 - \rightarrow shallow semantic parsing, z.B. Framenet:
 - \rightarrow Lexikon von semantic frames (Prädikat+Argumente): Informationen z.B. zu Subkateg. und diathetische Alternativen)
- Semantic Parsing
 - ightarrow natural language understanding

Einige Anwendungsgebiete:

- Identifizierung von Einheiten und Relationen in Informations**extraktionsanwendungen** (aufgrund syntaktischer Funktion / Position im Syntaxbaum)
- Voraussetzung f
 ür semantische Analyse (basierend auf Kompositionalitätsprinzip)
- Disambiguierung in maschineller Übersetzung und Question **Answering** Systemen
- Einsatz in Korrektursystemen (Rechtschreibung, Interpunktion)

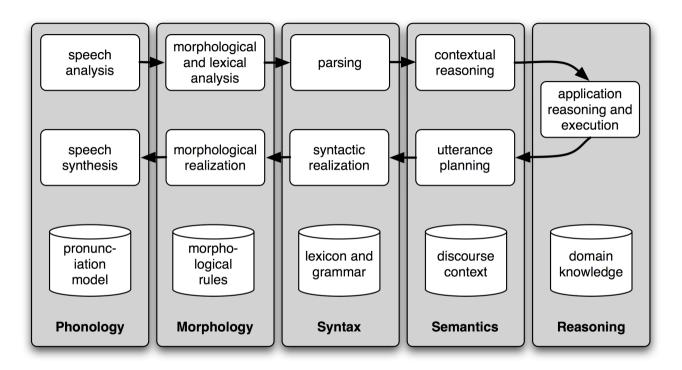


Abbildung 6: Parsing in NLP-Pipeline (Spoken Dialogue System), http://www.nltk.org/book/ch01.html#fig-sds

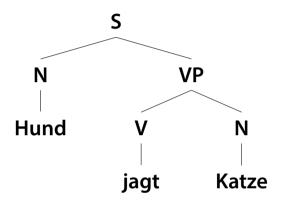
2.3.5 Abbildungen syntaktischer Strukturen

Syntaxbaum (auch: Parsebaum, Ableitungsbaum)

- = gerichteter Graph
- → mathematische Repräsentation hierarchischer Struktur
 - Gerichteter Graph besteht aus:
 - Knoten: → Flemente der Struktur
 - Kanten: → geordnete Paare von Knoten (ggf. gelabelt)
 - \rightarrow Repräsentation der Relation zwischen zwei Knoten

2

• Baumdiagramm (Konstituentenstruktur):



Klammerausdruck (Konstituentenstruktur):

[S [N Hund] [VP [V jagt] [N Katze]]]

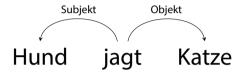
• Baumdiagramm (Dependenzstruktur):



• Klammerausdruck (Dependenzstruktur):

[jagt [Hund] [Katze]]

 Darstellung Dependenzstruktur mit Relationen (gelabelte Kanten):



Notation als Tripel:

(jagt, Subjekt, Hund), (jagt, Objekt, Katze)

- Darstellung als gelabelter gerichtetet Graph:
 - → als 'ungeordneter' Baum: abstrahiert von linearer Wortstellung
 - \rightarrow als 'geordneter' Baum: Wörter als Blätter