Übersicht

Teil I.

Syntax - Linguistische und formale Grundlagen

2. Einführung

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 Satzgliedstellung, Intonation u.a.) **alle wohlgeformten Sätze einer Sprache abgeleitet** werden können.

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

- **Etymologie: σύνταξις [syntaksis] = 'Zusammensetzung'**
  → *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**
  → Regeln der Kombination von Wörtern zu größeren Einheiten
- **als Sprachstrukturanalyse ist Syntax Teil der Grammatik einer Sprache**

2.1.2. Grammatik

- **Etymologie: (τέχνη) γραμματική [(technē) grammatikē] = 'Schreibkunst'**
- **allgemein- und fachsprachlicher Begriff**
- **bezieht sich auf die sprachliche Struktur**
  → Lautstruktur: Phonologie
  → Wortstruktur: Morphologie
  → Satzstruktur: Syntax

- **Syntax: rein formale Analyse des Strukturaufbaus sprachlicher Einheiten oberhalb der Wortebene**
  $\rightarrow$ Erfüllung von Wohlgeformtheitsbedingungen (Grammatikalität)
  $\rightarrow$ unabhängig von Semantik und Pragmatik, vgl. Chomsky 1957, 'Syntactic Structures':

  'colorless green ideas sleep furiously'

Relevanz der Morphologie für Syntax:

- **Wortartenklassifikation**

  $\rightarrow$ Zusammensetzung syntaktischer Einheiten (Kategorien) aus lexikalischen Einheiten (lexikalische Kategorien/Wortarten)

- **Morphosyntax**

  $\rightarrow$ Analyse von Wortformen, insofern sie für die syntaktische Strukturanalyse relevant sind (Kasus und Agreement)

  $\rightarrow$ formale Repräsentation als Merkmalstrukturen

Grammatikbegriff

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

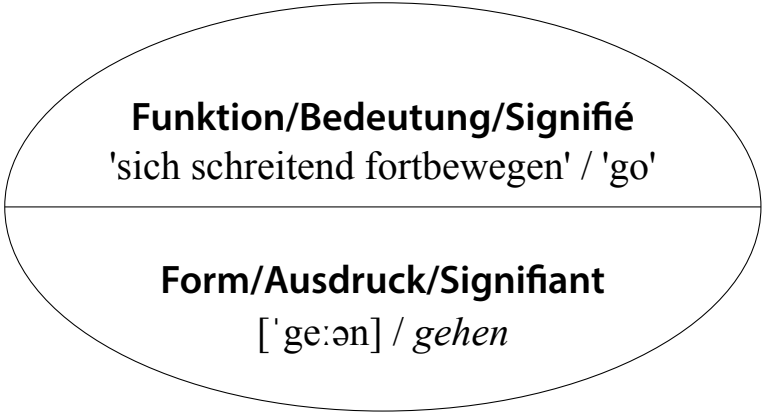
- **d) Grammatik als Regelbuch**
  $\rightarrow$ Lehrwerk, das die Regularitäten einer natürlichen Sprache enthält
- **e) Formale Grammatik**
  $\rightarrow$ Grammatik einer formalen Sprache, die zur Modellierung der Grammatik einer natürlichen Sprache verwendet werden kann

2.1.3. Semiotische Grundlagen

- **Sprachliche Größen als Zeichen (Saussure)**

  → *Form-Funktionspaar*

  → *gilt für Wörter (Lexikon) als auch daraus gebildeten Sätzen*



= Sprachliches Zeichen

- **Wörter als sprachliche Zeichen**
  → *z.B.: Nomen: Objektvorstellung; Verb: Ereignisvorstellung*

- **Sätze als komplexe sprachliche Zeichen**
  → *Satzform zusammengesetzt aus Wortformen (Elementarzeichen)*
  → *These der **Kompositionalität der Bedeutung**: Satzbedeutung folgt aus Bedeutung der Wortformen in Abhängigkeit von der syntaktischen Struktur (nicht arbiträre Beziehung zwischen Form und Bedeutung: kompetenter Sprecher versteht neu gebildete Sätze)*
  → *Syntax: formale und funktionale Regeln der Zusammensetzung (Konstituenten- und Dependenzstruktur)*

2.1.4. Syntaktische Form

Mittel zum Ausdruck syntaktischer Funktion

- **Wortstellung** (strukturell)
  → Lineare Anordnung (SOV vs. SVO usw.)

  Katze jagt Hund : Hund jagt Katze

- **Kasus** (morphosyntaktisch)
  → Rektion: Markierung abhängigen Elements (*dependent-marking*)

- **Agreement** (morphosyntaktisch)
  → Kongruenz von Merkmalen zwischen abhängigen Elementen
  → z.B. Spiegelung von Merkmalen der Subjekt-NP (Numerus, Person) am Verb (*head-marking*)

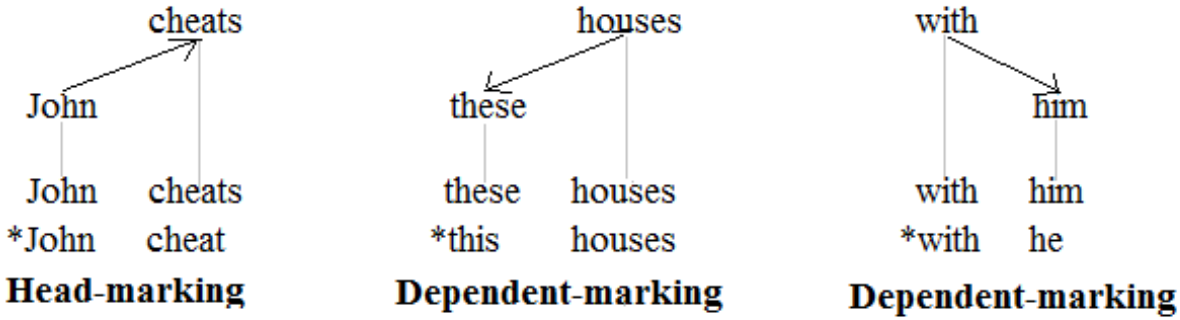


Abbildung 1: *Agreement vs. Kasus / head- vs. dependent-marking (von Tjo3ya, modifiziert von Ceccee - https://commons.wikimedia.org/wiki/File:Head-marking.png, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=36534797)*

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

2.2.1. Satzstruktur

- **Satz als zentraler Untersuchungsgegenstand der Syntax**
  → sprachliche Form einer Äußerung (Sprechakt)
  → Beobachtung: lineare Abfolge von Wörtern (Wortfolge)
  → Syntax: Beschreibung und Analyse der hierarchischen Struktur von Sätzen
  → d.h. ihres Aufbaus aus Wörtern und Phrasen (syntaktische Einheiten), den funktionalen Abhängigkeiten zwischen diesen syntaktischen Einheiten, sowie der Struktur komplexer Sätze

- **Struktur**
  $\rightarrow$ Menge von Relationen, die zwischen Elementen einer Grundmenge bestehen (Relation: Menge geordneter Paare)

- **Syntaktische Struktur**
  $\rightarrow$ 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-Dependens)

2.2.2. Syntagmatische Relation

- **Konstituenten-Struktur**
  $\rightarrow$ aus welchen **syntaktischen Einheiten** besteht ein Satz?

- **Syntagma: Gruppe sprachlicher Elemente in Äußerung**
  $\rightarrow$ durch strukturalistische discovery procedures (syntaktische Tests): Feststellung von syntaktischen 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

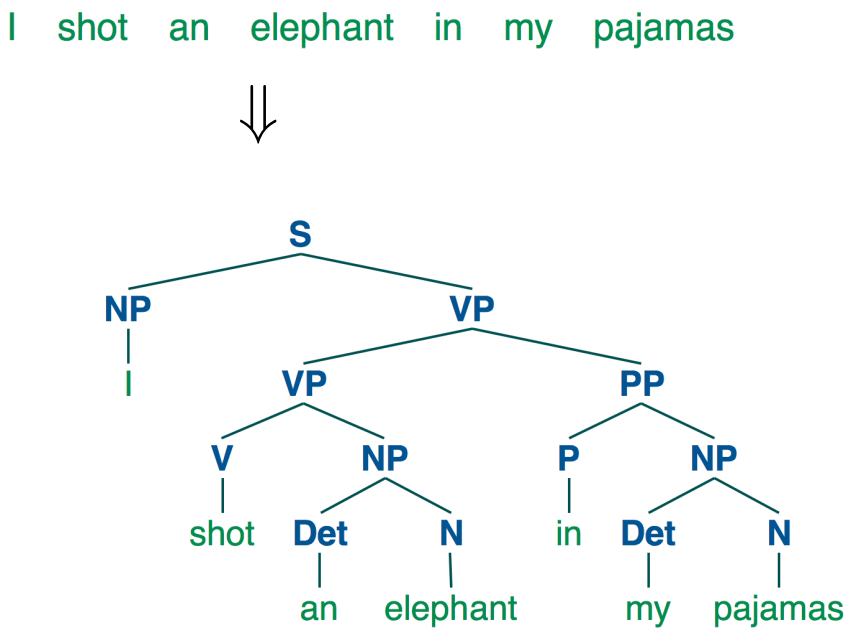


Abbildung 2: *Von der Wortfolge zur syntaktischen Struktur (Konstituenzmodell)*

2.2.3. Grammatische Relationen

- **Dependenz-Struktur**
  $\rightarrow$ in welcher **syntaktische Beziehung** stehen Wörter, welche **Funktion** haben sie im Satz?

- **Funktionale Satzanalyse**
  $\rightarrow$ notwendige und nicht-notwendige Einheiten im Satz
  $\rightarrow$ Abhängigkeitsverhältnisse zwischen Wörtern
  $\rightarrow$ Prädikat + Argumente (notwendige Ergänzungen) + Angaben

- **Syntax natürlicher Sprachen im Dependenzmodell:**
  $\rightarrow$ Regeln der Kombination von Wörtern nach Abhängigkeitsrelationen

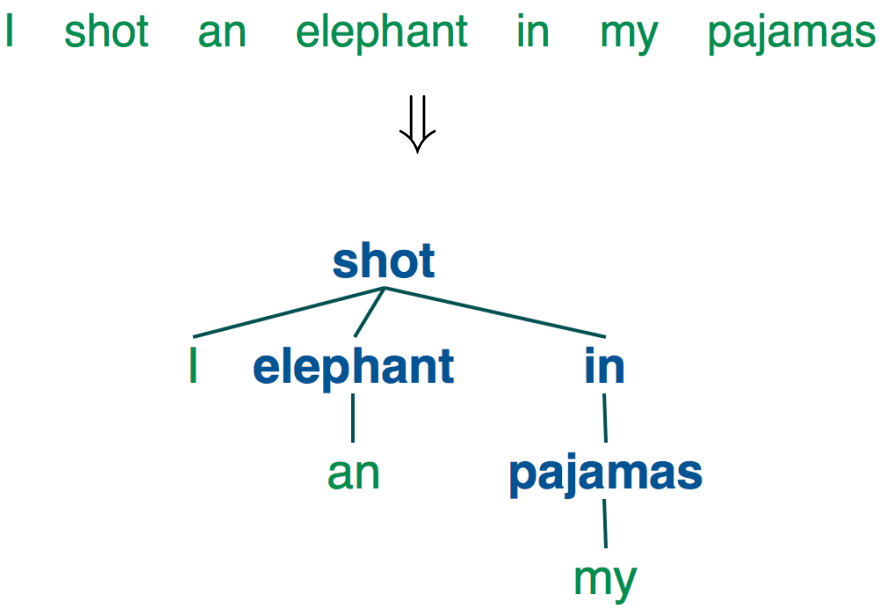


Abbildung 3: *Von der Wortfolge zur syntaktische Struktur (Dependenzmodell)*

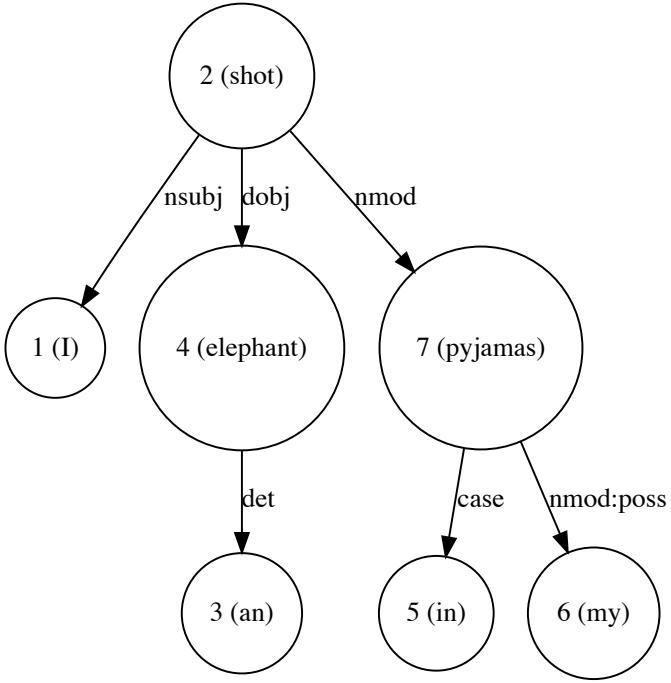


Abbildung 4: *Syntaktische Struktur mit gelabelten Relationen (Dependenzmodell)*

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)"

  → *über Logik von Port-Royal (1662) beeinflusst strukturalistische Distributionsanalyse (Saussure, Bloomfield)*
  → Chomsky (1957, 'Syntactic Structures'): mathematische Modellierung mit kontextfreien Grammatiken

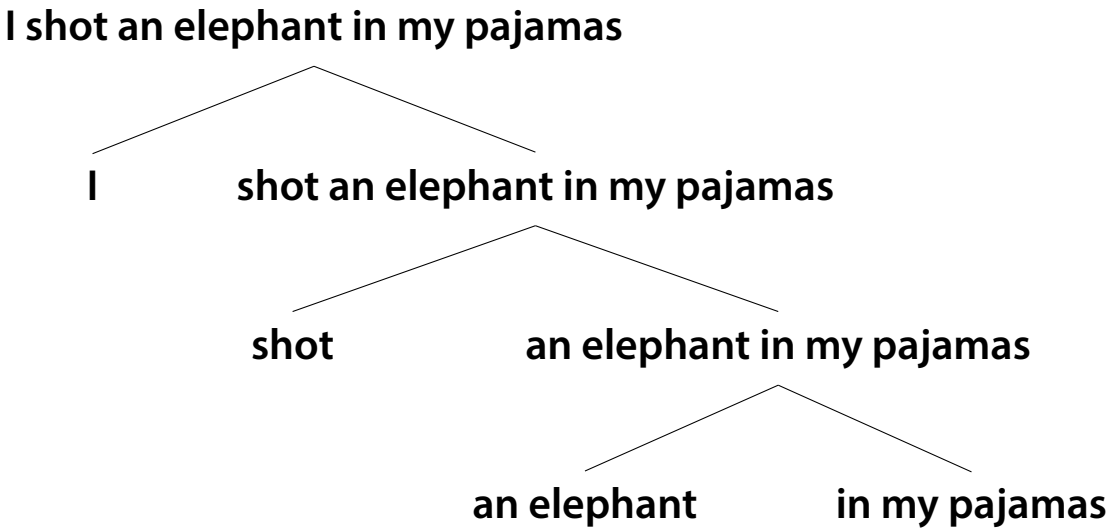


Abbildung 5: *binäre Zergliederung in unmittelbare Konstituenten (immediate constituents; IC)*

- **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')

2.3. Automatische Syntaxanalyse

- **Was muss ein Grammatikmodell als adäquate Beschreibung und Analyse der syntaktischen Struktur einer natürlichen Sprache leisten?**

  $\rightarrow$ **Strukturerkennung**: welche Sätze sind wohlgeformt? (Erkennung genau der grammatisch korrekten Sätze)

  $\rightarrow$ **Strukturwiedergabe**: wie ist die grammatische Struktur eines Satzes aufgebaut? (linguistisch sinnvolle Strukturanalyse)

Möglichkeiten der Syntaxanalyse:

- **Beschreibung des Sprachsystems**
  $\rightarrow$ traditionelle Buch-Grammatik; nicht-computational
- **Aufzählung aller grammatischen Sätze**
  $\rightarrow$ Problem 1: natürliche Sprachen sind unendlich
  $\rightarrow$ Problem 2: Struktur nicht repräsentiert

- **Beschreibung durch formale Grammatik**
  → als Regelsystem ist die Syntax natürlicher Sprachen mathematisch modellierbar durch formale Grammatiken
  → formale Beschreibung, die mit endlichen Mitteln die Analyse der Struktur einer unendlichen Menge an Sätzen ermöglicht
  → erzeugte formale Sprache als Modell der natürlichen Sprache
  → Sprache als die Menge aller wohlgeformten Sätzen
  → Erkennung und Wiedergabe der analysierten Struktur

2.3.1. Formale Grammatiken als Syntaxmodelle

formale Grammatik

- **Formales Regelsystem zur eindeutigen Beschreibung und Erzeugung einer formalen (!) Sprache**
  $\rightarrow$ mathematisches Modell
  $\rightarrow$ kann zur **Modellierung** (Beschreibung und Analyse) der syntaktischen Struktur **natürlicher Sprachen** verwendet werden kann
  $\rightarrow$ **Generierung aller wohlgeformten Sätze** (= generative Grammatik im weiteren Sinne)

Formale Sprache:

- **Menge aller aus Grammatik ableitbaren Wörter**
  $\rightarrow$ in natürlichsprachlicher Syntaxanalyse: sind diese formalsprachlichen Wörter natürlichsprachliche Sätze
- **Syntax formaler Sprache:**
  $\rightarrow$ Regeln der Kombination von Grundsymbolen (aus dem Alphabet) zu den Wörtern der Sprache (=formale Grammatik)

Formale Grammatik:

- **besteht aus:**
  - **Startsymbol**
  - **Nichtterminalsymbole**
    $\rightarrow$ Metasymbole
  - **Terminalsymbole**
    $\rightarrow$ Alphabet / Lexikon
  - **Produktionsregeln**
    $\rightarrow$ durch Einschränkungen der Regeln ergeben sich Sprachen verschiedener Komplexität (Chomsky-Hierarchie)

Kontextfreie Grammatik:

- Phrasenstrukturgrammatik im engeren Sinne
- **CFG-Einschränkung:**
  $\rightarrow$ links nur ein Nichtterminalsymbol: $S \rightarrow NP\ VP$
  $\rightarrow$ Ersetzung unabhängig von Kontext (Kontextfreiheit)
- **syntaktische Regeln:** $NP \rightarrow Det\ N\ (PP)$
  $\rightarrow$ Ersetzungsregeln (linke mit rechter Seite)
  $\rightarrow$ links: syntaktische Kategorien (Phrasen/Satzknoten)
  $\rightarrow$ rechts: obligatorische und optionale Nichtterminale (syntaktische + lexikalische Kategorien)
- **Rekursion:** $NP \rightarrow Det\ N\ (PP)$, $PP \rightarrow P\ NP$

- **lexikalische Regeln:**

  $N \rightarrow$ *'Hund'*

  $N \rightarrow$ *'Katze'*

  $Det \rightarrow$ *'der'* | *'die'*

  $\rightarrow$ Zuordnung lexikalische Kategorien/Wortarten (Präterminale) zu Lexemen (Terminale)

- **Wortarten (=lexikalische Kategorien):** $\rightarrow$ *Präterminale*
- **Lexeme:** $\rightarrow$ *Terminale (Alphabet)*

Auflistung 1: *Kontextfreie Grammatik*

```
1  S → NP VP
2  PP → P NP
3  NP → Det N | Det N PP | 'I'
4  VP → V NP | VP PP
5  Det → 'an' | 'my'
6  N → 'elephant' | 'pajamas'
7  V → 'shot'
8  P → 'in'
```

Klassifizierung syntaktischer Modelle:

- **modellierte Relation**
  $\rightarrow$ Konstituenzgrammatik : Dependenzgrammatik
- **Kategorien**
  $\rightarrow$ atomare Kategorien : Merkmalbündel
- **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:**
  $\rightarrow$ unendliche Menge an Sätzen mit endlichen Mitteln beschreibbar
  $\rightarrow$ rechnergestützt verarbeitbar durch Parsingalgorithmen
  $\rightarrow$ Beantwortung Fragen zur Komplexität natürlicher Sprache (ist jede natürliche Sprache kontextfrei?)
  $\rightarrow$ psycholinguistische Anwendung: Parser als Modell menschlicher Sprachverarbeitung

Nachteile:

- **Probleme mit struktureller Ambiguität**
  $\rightarrow$ wie Entscheidung für richtige (im Kontext intendierte) syntaktische Analyse?
- **Probleme mit Übergenerierung**
  $\rightarrow$ wie Vermeidung Produktion ungrammatischer Sätze?
- **keine vollständige Beschreibung möglich**
  $\rightarrow$ immer nur Annäherung an natürliche Sprache

2.3.2. Parsing als automatische Syntaxanalyse

Parsing:

- **formale Grammatik:**
  $\rightarrow$ Syntaktisches Strukturmodell, das aber nicht mehr ist als eine Sammlung von Strings (Regeln)
  $\rightarrow$ Verfahren notwendig, um zu entscheiden, ob eine Eingabe gemäß einer gegebenen formalen Grammatik wohlgeformt ist
- **Parsing-Algorithmen:**
  $\rightarrow$ 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:**
  $\rightarrow$ gleichzeitig Wiedergabe der in der Suche aufgebauten grammatischen Struktur der Eingabe (Syntaxbaum)

Parsing-Algorithmen:

- **top-down vs. bottom-up**
  $\rightarrow$ von Startsybol zu Blättern oder umgekehrt
- **Parsingalgorithmen mit dynamischer Programmierung**
  $\rightarrow$ 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*

```
parser = nltk.ChartParser(grammar,trace=1)
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 → 'I' *
|[----→     .     .     .     .     .     .| [0:1] S  → NP * VP
|.     [-----]     .     .     .     .     .| [1:2] V  → 'shot' *
|.     [----→     .     .     .     .     .| [1:2] VP → V * NP
|.     .     [-----]     .     .     .     .| [2:3] Det → 'an' *
|.     .     [----→     .     .     .     .| [2:3] NP → Det * N
|.     .     [----→     .     .     .     .| [2:3] NP → Det * N PP
|.     .     .     [-----]     .     .     .| [3:4] N  → 'elephant' *
|.     .     [-----------]     .     .     .| [2:4] NP → Det N *
|.     .     [----------→     .     .     .| [2:4] NP → Det N * PP
```

```
|.      .      [----------→      .      .      .| [2:4] S  → NP * VP
|.      [------------------]      .      .      .| [1:4] VP → V NP *
|.      [------------------→      .      .      .| [1:4] VP → VP * PP
|[------------------------]      .      .      .| [0:4] S  → NP VP *
|.      .      .      .      [-----]      .      .| [4:5] P  → 'in' *
|.      .      .      .      [----→      .      .| [4:5] PP → P * NP
|.      .      .      .      .      [-----]      .| [5:6] Det → 'my' *
|.      .      .      .      .      [----→      .| [5:6] NP → Det * N
|.      .      .      .      .      [----→      .| [5:6] NP → Det * N PP
|.      .      .      .      .      .      [-----]| [6:7] N  → 'pajamas' *
|.      .      .      .      .      [-----------]| [5:7] NP → Det N *
|.      .      .      .      .      [----------→| [5:7] NP → Det N * PP
|.      .      .      .      .      [----------→| [5:7] S  → NP * VP
|.      .      .      .      [------------------]| [4:7] PP → P NP *
|.      .      [-----------------------------]| [2:7] NP → Det N PP *
|.      [-----------------------------------]| [1:7] VP → VP PP *
|.      [-----------------------------------→| [1:7] VP → VP * PP
|[=======================================]| [0:7] S  → NP VP *
|.      .      [----------------------------→| [2:7] S  → NP * VP
|.      [-----------------------------------]| [1:7] VP → V NP *
|.      [-----------------------------------→| [1:7] VP → VP * PP
|[=======================================]| [0:7] S  → NP VP *
```

2.4. Syntaxtheorien

Konstituentengrammatiken:

- **Generative Grammatik (Chomksy):**
  → **Transformational Grammar** (Überführung Tiefen- in Oberflächenstruktur)
  → **X-Bar-Theory** (Ebene zwischen phrasalen und lexikalischen Kategorien)
  → **Government & Binding** (Ausweitung X-Bar-Schema auf Sätze)
  → **Minimalist Program** (Reduktion Strukturannahmen)

- **Unifikations- und Constraint-basierte, lexikalisierte Formalismen:**
  $\rightarrow$ **HPSG** (Head-Driven Phrase Structure Grammar)
  $\rightarrow$ **LFG** (Lexical Functional Grammar)

- **Weitere konstituentenbasierte Formalismen:**
  → **DCG** (Distinct Clause Grammar): Grammatik als Menge von prädikatenlogischen Sätzen; assoziiert mit PROLOG
  → **CG** (Categorical Grammar): Strukturinformation in Kategorien statt in Regeln (S/N statt V: Verb als Funktionswort, das mit einer NP (N) einen Satz (S) bildet)
  → **TAG** (Tree Adjoining Grammar): Teilbäume statt Regeln (Categorical Grammar); mild kontextsensitiv

Dependenzgrammatiken:

- **Traditionelle Dependenzgrammatik:**
  → **Tesnières Valenzgrammatik**
- **Computationale Modelle:**
  → **Bedeutung-Text-Modell** (Meľčuk)
  → **Word Grammar**
  → **Link Grammar**
  → **Constraint Grammar**

- **Stanford-Dependency-Parser:**
  $\rightarrow$ Kombination probabilistischer kontextfreier Grammatik (PCFG) mit Dependenzanalyse
  $\rightarrow$ Transformation Konstituentenbaum in Dependenzbaum (über Phrasenköpfe)
  $\rightarrow$ grammatische Relationen über handgeschriebene Muster (z.B.: Kopf der unmittelbar von S dominierten NP = Subjekt vom Kopf der unmittelbar dominierten VP)

Probabilistische, daten-orientierte Modelle:

- **PCFGs** *(Probabilistische kontextfreie Grammatiken)*:
  $\rightarrow$ Wahrscheinlichkeitsdaten zum Auftreten von Regeln zur Disambiguierung in CFG-Modellen
- ***grammar induction***
  $\rightarrow$ Aufbau Grammatikmodell aus Korpusdaten
  $\rightarrow$ Regeln und Wahrscheinlichkeiten aus syntaktisch annotiertem Korpus (Treebank) lernen

Funktionalistische Syntaxtheorien:

- **Functional Grammar** *(Simon Dik): Modell mit prädikatenlogischem Formalismus*
- **Role & Reference Grammar** *(Van Valin)*

  → *These der Nichtautonomie der Syntax (gegen Generative Grammatik)*

  → Sprache als kommunikationsbasiert (usage-based)

  → Motivation sprachlicher Strukturen durch Pragmatik und Semantik

  → Analyse von Interdependenz semantischer Rollen (Agens, Patiens usw.) und Topik-Fokus-Struktur mit Syntax

Kognitive Linguistik:

- **Construction Grammar** *(Fillmore; Goldberg): unifikationsbasierter Formalismus; Hierarchie von Konstruktionen als syntaktische Einheiten (Kontinuum Lexikon - Syntax)*
- **Cognitive Grammar** *(Langacker): Grammatik als Constraints der Kombination von symbolischen Einheiten*
- **Metapherntheorie** *(Lakoff): konzeptionelle Konstruktionen*

→ *These der Nichtautonomie der Syntax (gegen Generative Grammatik)*

→ 'Linguistic Wars' (60er/70er), Lakoff u.a. gegen Chomsky: Semantik statt Syntax als Basis für Sprachtheorie (Kognitive Linguistik)

→ Sprache als kognitionsbasiert

→ Motivation von sprachlichen Strukturen durch kognitive Prozesse

→ Verständnis von Grammatik als symbolisch: Eigenbedeutung syntaktischer Konstruktionen

Beschreibende Modelle:

- **Deutsches Stellungsfeldermodell**

  $\rightarrow$ sprachspezifisch

  $\rightarrow$ Beschreibung der linearen Struktur deutscher Sätze (Wortstellung) über Stellungfelder (Vorfeld, Mittelfeld und Nachfeld)

2.5. Abbildungen syntaktischer Strukturen

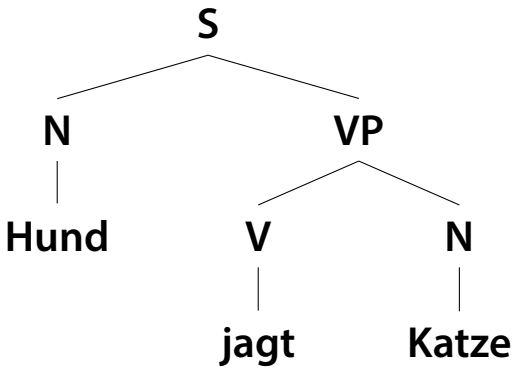
Syntaxbaum, auch: Parsebaum, Ableitungsbaum:

$\rightarrow$ gerichteter Graph

$\rightarrow$ mathematische Repräsentation hierarchischer Struktur

- **Gerichteter Graph:**
  $\rightarrow$ Knoten + gerichtete Kanten
- **Knoten:**
  $\rightarrow$ Elemente der Struktur
- **Kanten:**
  $\rightarrow$ geordnete Paare von Knoten; ggf. gelabelt
  $\rightarrow$ Repräsentation der Relation zwischen zwei Knoten

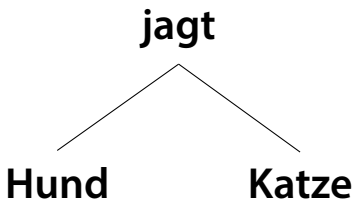
- **Baumdiagramm (Konstituentenstruktur):**



- **Klammerausdruck (Konstituentenstruktur):**

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

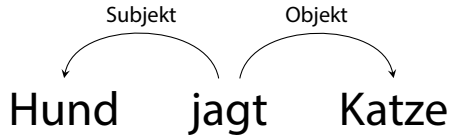
- **Baumdiagramm (Dependenzstruktur):**



- **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:**

  $\rightarrow$ als 'ungeordneter' Baum: abstrahiert von linearer Wortstellung

  $\rightarrow$ als 'geordneter' Baum: Wörter als Blätter

Auflistung 3: *Graphen-Notation (dot-Format)*

```
digraph G{
edge [dir=forward]
node [shape=circle]

1 [label="1 (Hund)"]
2 [label="2 (jagt)"]
2 → 1 [label="nsubj"]
3 [label="3 (Katze)"]
2 → 3 [label="dobj"]
}
```

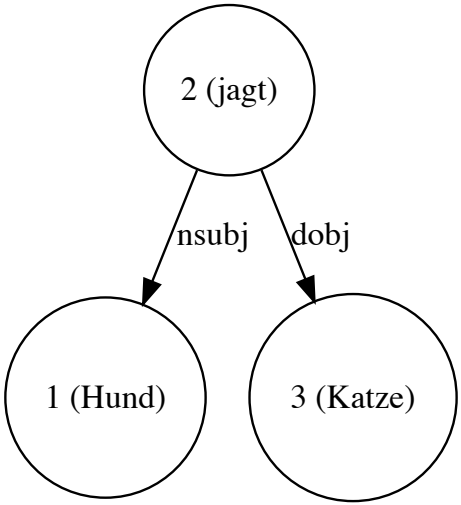


Abbildung 6: *Visualsisierung der Dependenz-Graphstruktur (mit Graphviz)*

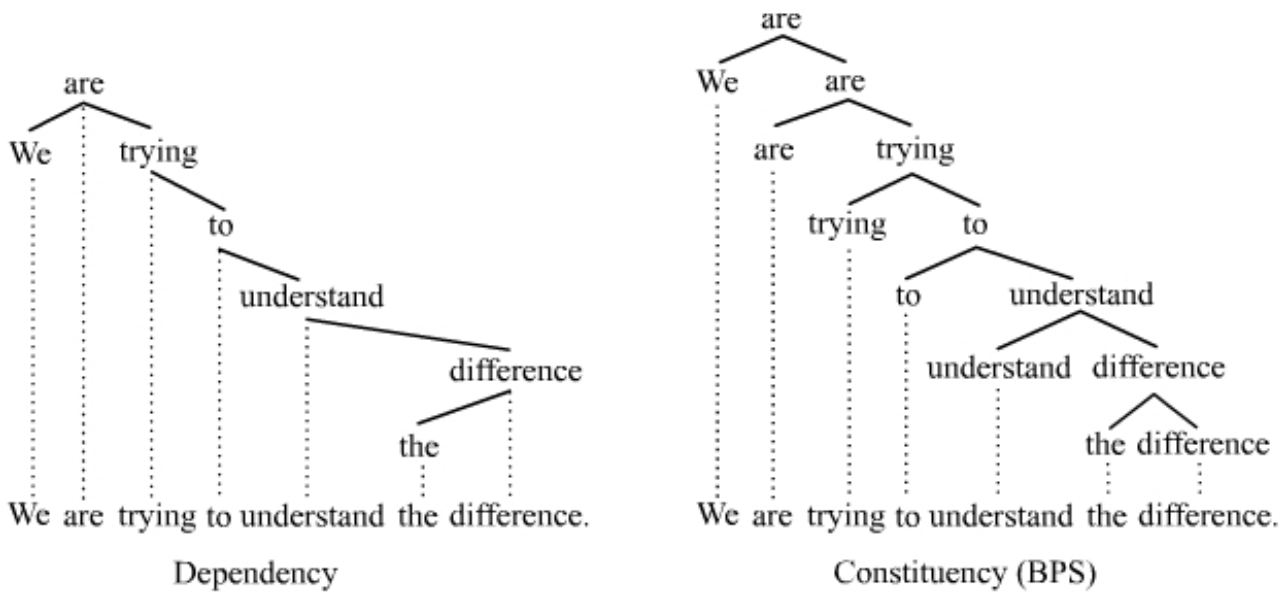


Abbildung 7: *Geordneter Dependenzbaum - Konstituentenbaum (von Tjo3ya - eigenes Werk, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17517283)*

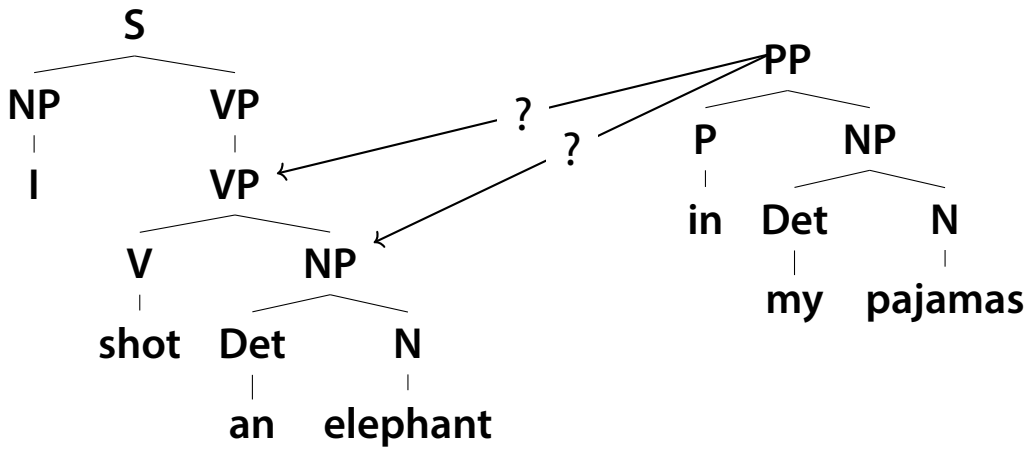
2.6. Syntaktische Ambiguität

- **strukturelle Ambiguität**
  $\rightarrow$ mehr als eine Strukturanalyse möglich

- **Lösung**
  $\rightarrow$ 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)

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

Beispiele: Präpositionalphrasen, Adverbialphrasen (s. Übung)



Koordinierungsambiguität:

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

garden-path-Sätze:

- *The old man the boat.*
- *The horse raced past the barn fell.*

2.7. Computerlinguistische Anwendungen

2.7.1. Anwendungsgebiete

- Identifizierung von Einheiten und Relationen in **Informationsextraktionsanwendungen**
- Voraussetzung **semantischer Analyse** (basierend auf Kompositionalitätsprinzip)
- Disambiguierung in **maschineller Übersetzung und Question Answering** Systemen
- Einsatz in **Korrektursystemen** (Rechtschreibung, Interpunktion)

2.7.2. Voraussetzungen und Folgeanwendungen

- Parsing in NLP-Pipeline:
- `http://www.nltk.org/book/ch07.html#fig-ie-architecture`
- `http://www.nltk.org/book/ch01.html#fig-sds`

Voraussetzungsschritte für automatische Syntaxanalyse:

- Sentence Segmentation

  $\rightarrow$ `Liste von Strings`

- Tokenisierung

  $\rightarrow$ `Liste von Stringlisten`

- Part-of-Speech-Tagging

  $\rightarrow$ `Liste von Tupellisten (token,pos-tag)`

- (Stemming)

- (morphologisches Parsing [Kasus, Agreement])

Mögliche Folgeanwendungen:

- Entity Extraction

  → `Liste von Bäumen`

- Relation Extraction

  → `Liste von Tripeln: (entity, relation, entity)`

- Semantic Role Labelling

  → *shallow semantic parsing, z.B. Framenet:*

  → *Lexikon von semantic frames (Prädikat+Argumente): Informationen z.B. zu Subkateg. und diathetische Alternativen)*

- Semantic Parsing

  → *natural language understanding*