# 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
- 2.2.3 Abbildungen syntaktischer Strukturen

## 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 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 und Grammatikbegriff

**Syntax:**

- **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*
- **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**
  $\rightarrow$ 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**
  $\rightarrow$ **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

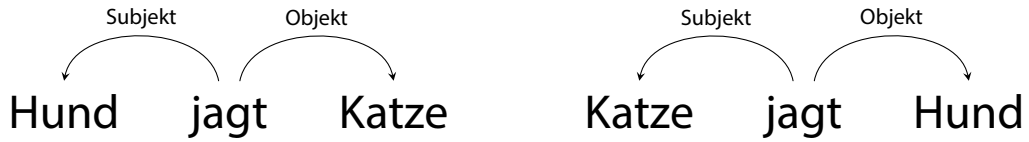
## Relevanz der Morphologie für Syntax:

- **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):

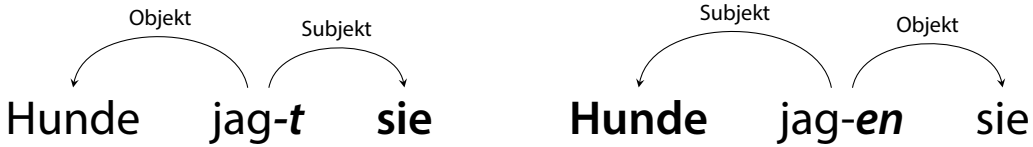
  → *morphologische Markierung* der syntaktischen Funktion am *abhängigen Element*

  → 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 → 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"*

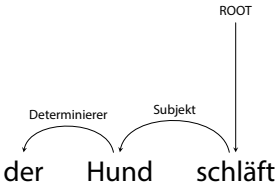
## Satzstruktur:

- **Satz als zentraler Untersuchungsgegenstand der Syntax**
  → sprachliche Form einer ***Äußerung*** (Sprachhandlung)
  → **Beobachtung: lineare Abfolge** von Wörtern (Wortfolge)
  → **Syntax:** Beschreibung und Analyse der **hierarchischen Struktur** 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**
  $\rightarrow$ aus welchen **syntaktischen Einheiten** besteht ein Satz?
- **Syntagma: Gruppe sprachlicher Elemente in Äußerung**
  $\rightarrow$ 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

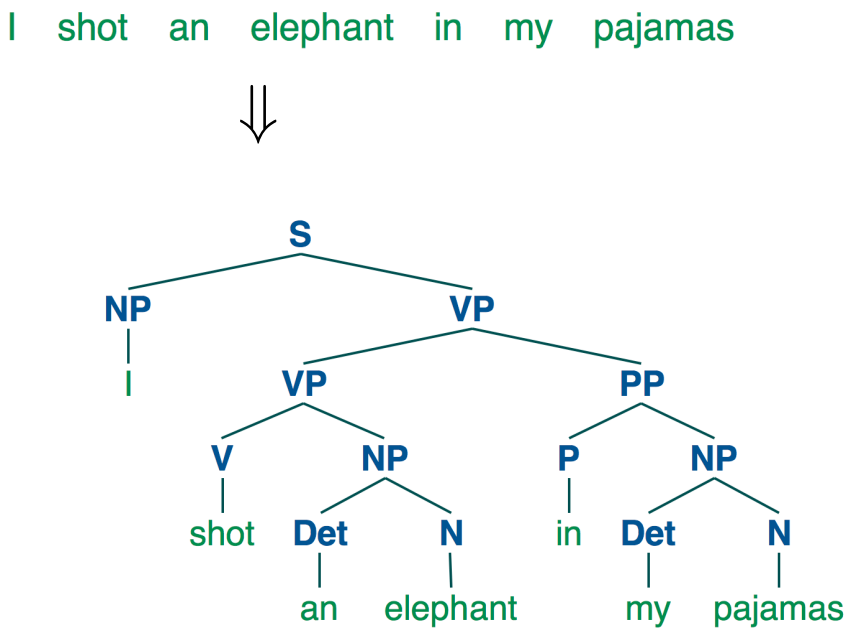



Abbildung 1: *Von der Wortfolge zur syntaktischen Struktur (Konstituenzmodell)*

***Historischer Hintergrund (Konstituentenanalyse):***

- **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

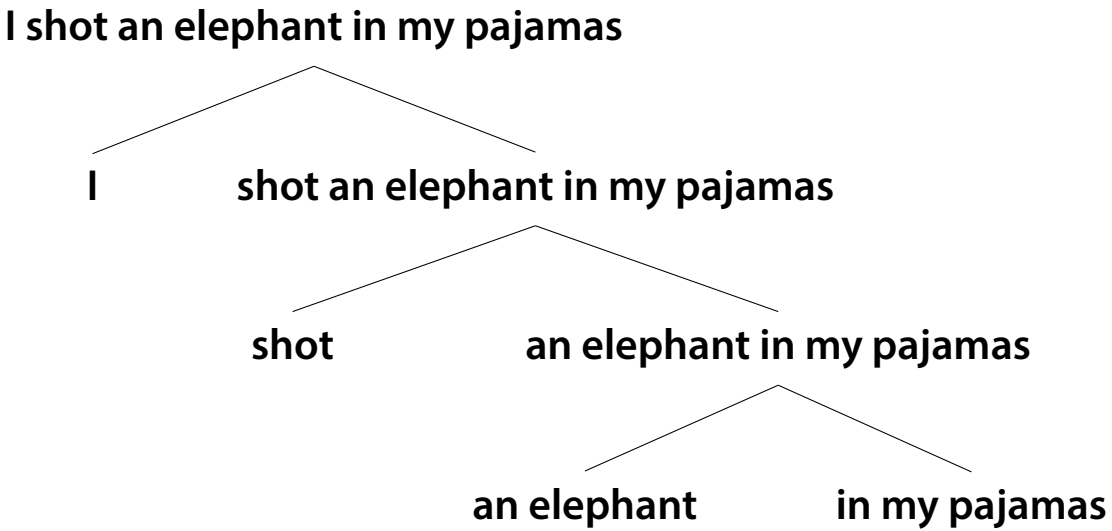



Abbildung 2: *binäre Zergliederung in unmittelbare Konstituenten (immediate constituents; IC)*

# 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.

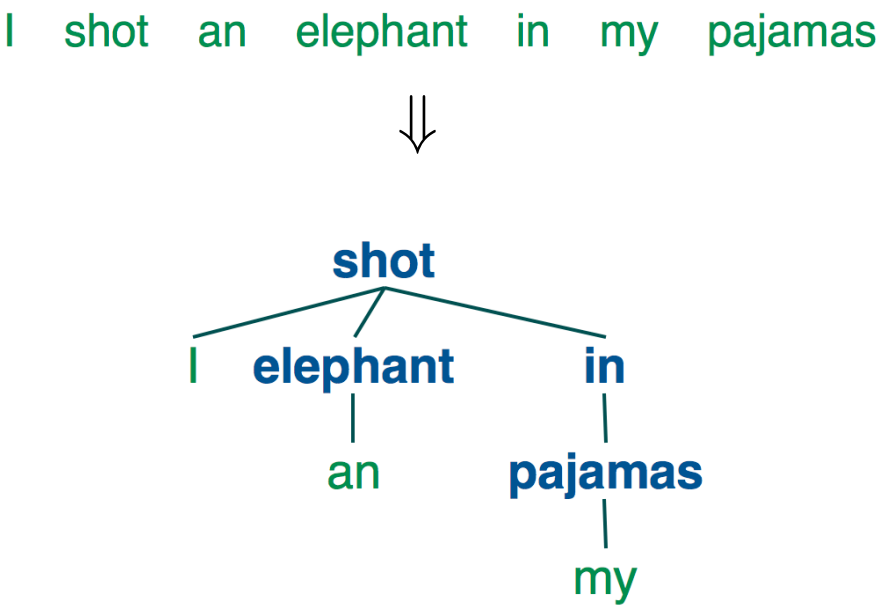



Abbildung 3: *Von der Wortfolge zur syntaktische Struktur (Dependenzmodell)*

## *Historischer Hintergrund (Dependenzanalyse):

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

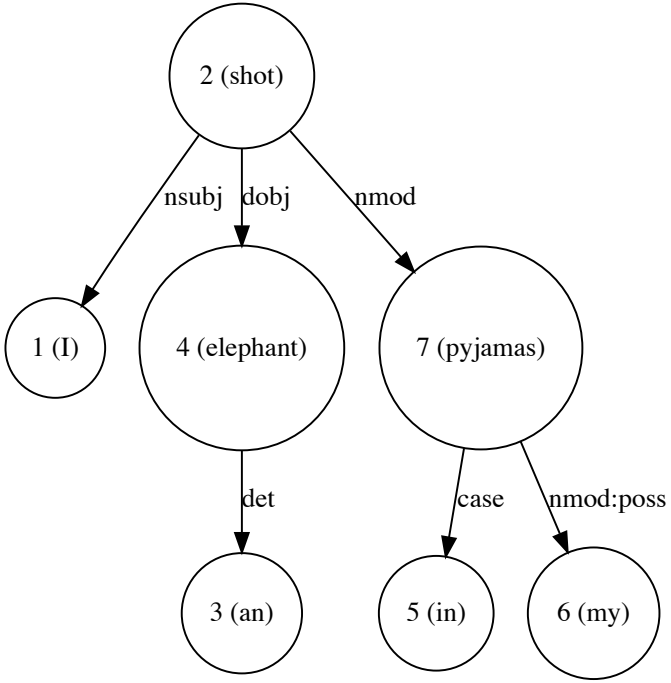


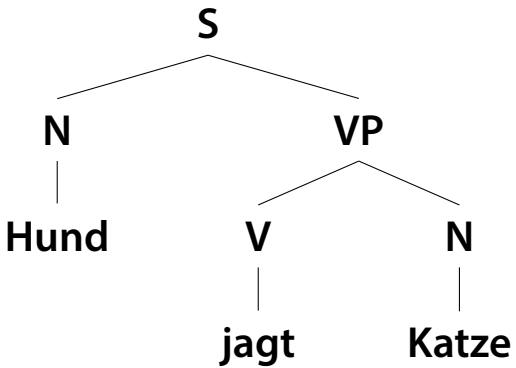

Abbildung 4: *Syntaktische Struktur mit gelabelten Relationen (Dependenzmodell)*

# 2.2.3 Abbildungen syntaktischer Strukturen

**Syntaxbaum (auch: Parsebaum, Ableitungsbaum)**
= ***gerichteter Graph***
→ mathematische Repräsentation hierarchischer Struktur

- **Gerichteter Graph** besteht aus:
  - **Knoten** = Elemente der Struktur
  - **Kanten** = geordnete Paare von Knoten (ggf. gelabelt)
    → Repräsentation der Relation zwischen zwei Knoten

- **Baumdiagramm (Konstituentenstruktur):**

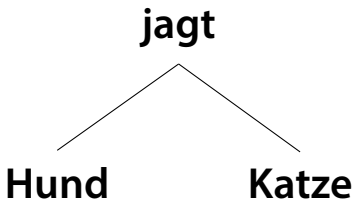


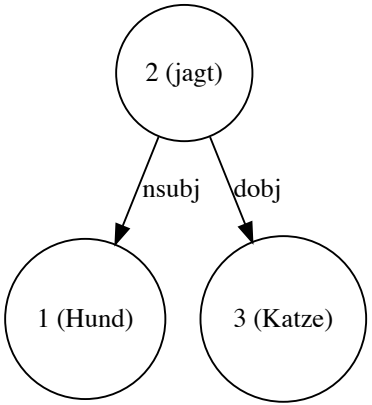

- **Klammerausdruck (Konstituentenstruktur):**

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

- **Baumdiagramm (Dependenzstruktur):**

  $\rightarrow$ optional: Relationen als Label der Kanten

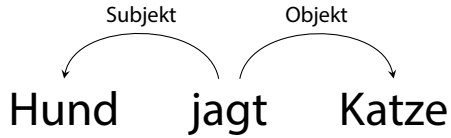



- **Klammerausdruck (Dependenzstruktur):**

  *[jagt [Hund ] [Katze ]]*

- **'Dependenz-Blume'** als alternative Darstellung:
  $\rightarrow$ Erhalt der linearen Anordnung der Wortfolge




- **Notation als Tripel:**

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

# 2.3 Automatische Syntaxanalyse

- 2 Aufgaben eines **Grammatikmodells** für die Analyse der syntaktischen Struktur einer natürlichen Sprache:

  1. ***Strukturerkennung***: *ist ein Satz wohlgeformt?*
     $\rightarrow$ **Erkennung genau der grammatisch korrekten Sätze**
  2. ***Strukturwiedergabe***: *wie ist ein Satz aufgebaut?*
     $\rightarrow$ **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**
  $\rightarrow$ **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. $\{a, aa, aba, ...\}$)

  $\rightarrow$ 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*:
    $\{der - Hund - jagt - die - Katze, die - Katze - jagt - den - Hund, ...\}$

**Aufbau einer formalen Grammatik:**

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

4. **Produktionsregeln** (z.B. $X \to aY$ oder $S \to NP \quad VP$)
   $\to$ Ersetzungregeln; geben an, wie aus Symbolfolgen (beginnend mit $S$) neue Folgen (Wörter) gebildet werden können
   $\to$ 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:** $NP \rightarrow Det\ N\ (PP)$
  $\rightarrow$ **links: syntaktische Kategorien** (Phrasen-/Satzknoten: S, NP)
  $\rightarrow$ 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 → Terminal):**
  → Zuordnung lexikalische Kategorien/Wortarten zu Lexemen
  $N \rightarrow$ *'Hund'*
  $N \rightarrow$ *'Katze'*
  $Det \rightarrow$ *'der'* | *'die'*
- **Wortarten (=lexikalische Kategorien):** → *Präterminale* (Untermenge der Nichtterminale)
- **Lexeme:** → *Terminale*

Auflistung 1: *Kontextfreie Grammatik*

```
#syntaktische Regeln:
        S → NP VP
        PP → P NP
        NP → Det N | Det N PP
        VP → V NP | VP PP
#lexikalische Regeln:
        Det → 'an' | 'my'
        N → 'elephant' | 'pajamas'
        V → 'shot'
        P → 'in'
```

## Klassifizierung syntaktischer Modelle:

- **modellierte Relation**
  $\rightarrow$ Konstituentengrammatik : Dependenzgrammatik
- **Kategorien**
  $\rightarrow$ atomare Kategorien (CFGs) : komplexe Merkmalstrukturen (feature-based CFGs)
- **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 Ausschnitt natürlicher Sprache modelliert

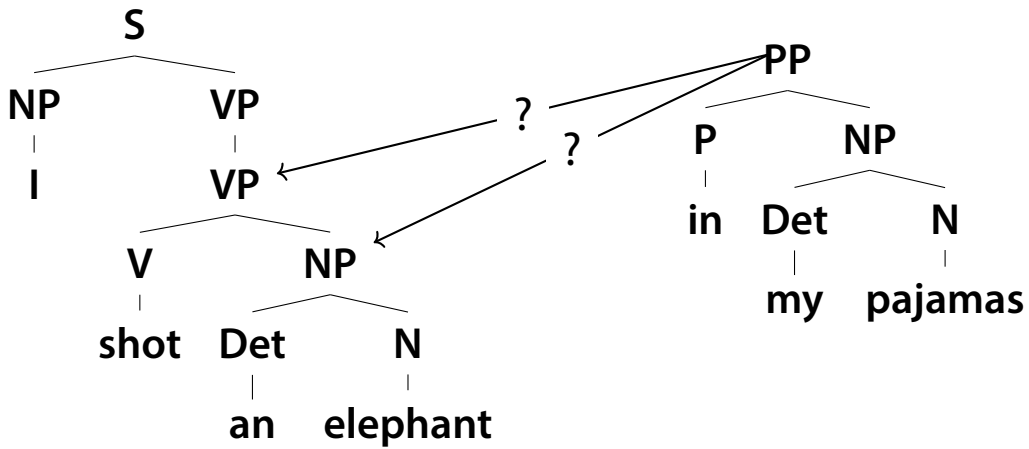
## 2.3.2 Syntaktische Ambiguität

- **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)

**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, dessen Regeln aber nicht mehr sind als eine **Sammlung von Strings**
  $\rightarrow$ **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:**

- **CFG-Parsing**
  $\rightarrow$ top-down vs. bottom-up vs. dynamische Programmierung
- **Unifikationsparsing**
  $\rightarrow$ *Verarbeitung von Merkmalstrukturen*
- **statistische Parsingalgorithmen**
  $\rightarrow$ *Viterbi-Algorithmus* (effizientes Auffinden der wahrscheinlichsten Abbleitung; Wahrscheinlichkeiten aus syntaktisch annotiertem Korpus gelernt = Treebank)

- **Dependency Parsing**
  $\rightarrow$ u.a. *transition-based* (basierend auf aus Dependency Treebank gelerntem Modell)
- **Partielles Parsing / Chunk-Parsing**
  $\rightarrow$ *Parsing as Tagging* (mit regulärer Grammatik oder Training eines Klassifikators)

# 2.3.4 Computerlinguistische Anwendungen

**Einige Anwendungsgebiete:**

- **Entity und Relation Extraction**
  $\rightarrow$ aufgrund syntaktischer Funktion (Dependenzanalyse)
  $\rightarrow$ bzw. aufgrund der Position im Syntaxbaum (CFGs)
- Disambiguierung in **maschineller Übersetzung und Question Answering** Systemen
- Einsatz in **Korrektursystemen** (Rechtschreibung, Interpunktion)

- Voraussetzung für **semantische Analyse**
  → basierend auf **Kompositionalitätsprinzip** (Bedeutung eines Satzes aus Bedeutung der Teile berechenbar)
  → **Montague-Grammatik** mit Lambda-Kalkül als semantische Interpretationsregeln zu syntaktischen Regeln
  → Ergebnis: **prädikatenlogische Repräsentation**
  → (vgl. *Semantik-Vorlesung* 4. Fachsemester)

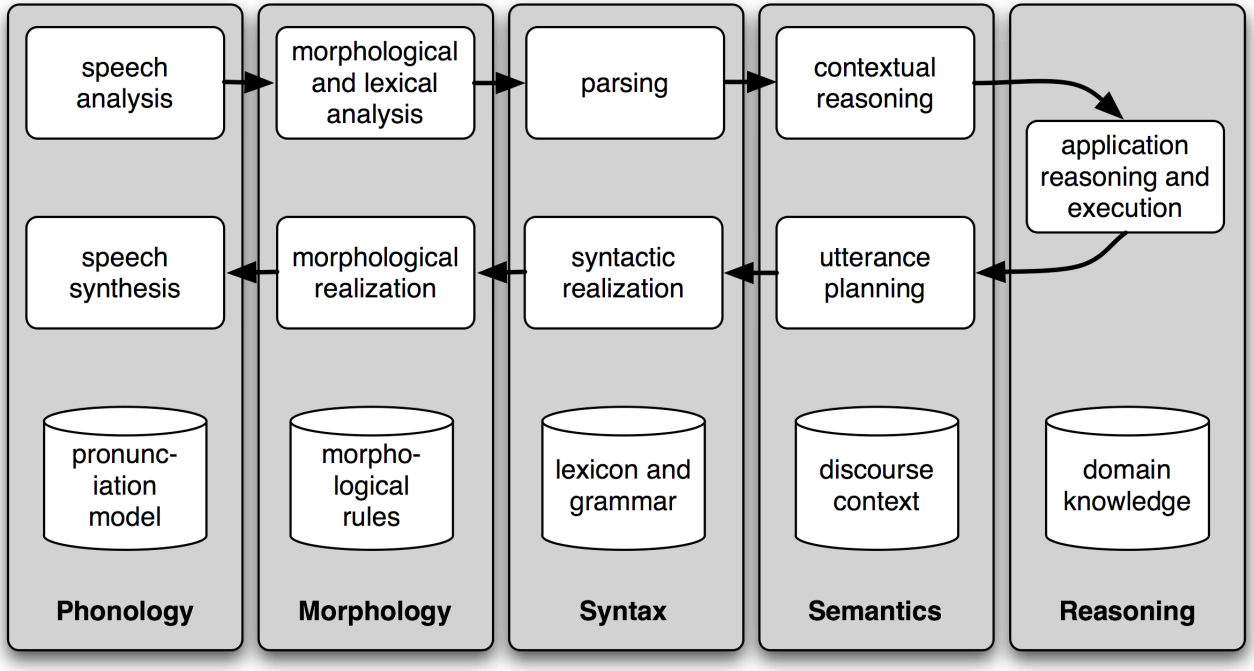



Abbildung 5: *Parsing in NLP-Pipeline (*Spoken Dialogue System*),* `http://www.nltk.org/book/ch01.html#fig-sds`

**Voraussetzungsschritte für automatische Syntaxanalyse:**

- Sentence Segmentation
- Tokenisierung (`split()`)
- Part-of-Speech-Tagging (`token,pos-tag`)
- (Stemming, morphologisches Parsing: Kasus, Agreement)

**Mögliche Folgeanwendungen:**

- **Entity Extraction** (`Tree`-Objekte)
- **Relation Extraction** (`entity, relation, entity`)

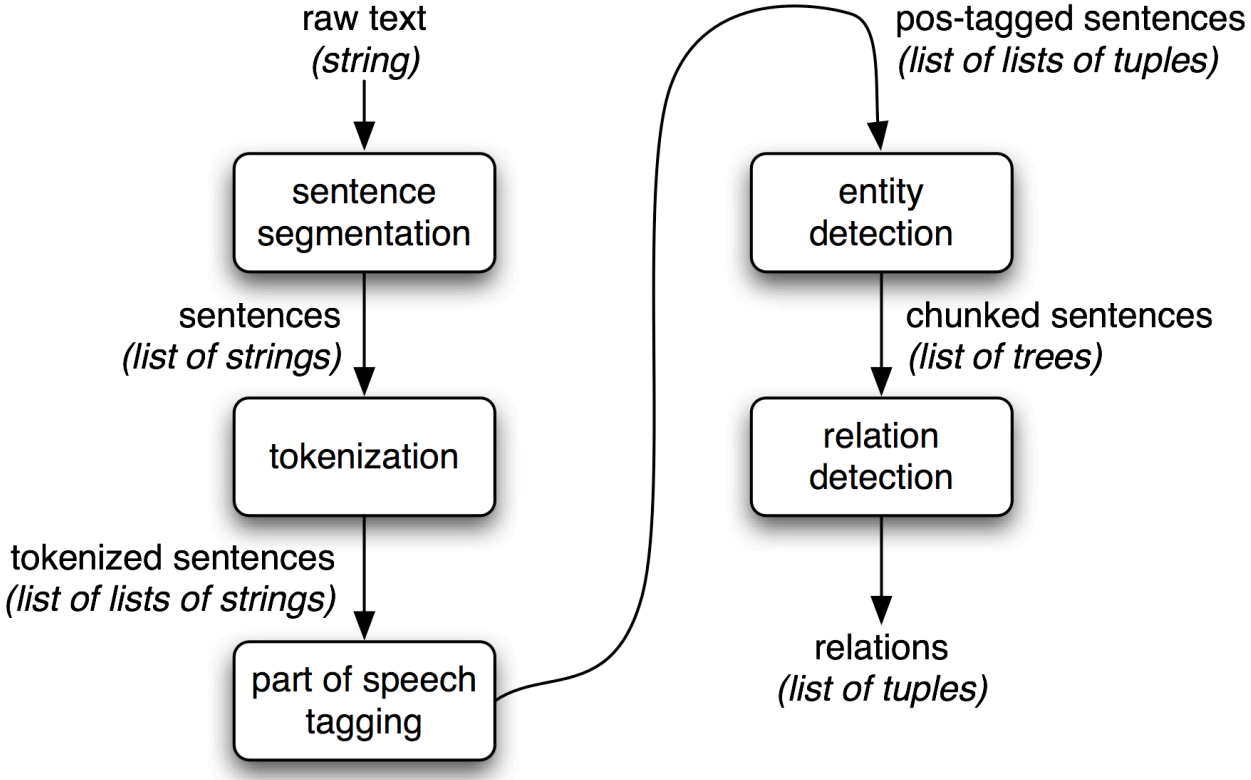



Abbildung 6: *Parsing in NLP-Pipeline* (Information Extraction), `http://www.nltk.org/book/ch07.html#fig-ie-architecture`