

14 Zusammenfassungen

Sitzung 2: Einführung

- **Syntax-, Grammatik- und Satzbegriff**
- **Syntax natürlicher Sprachen**
 - Regeln der Kombination von Wörtern zu Sätzen (Satzlehre)
- **Konstituentenstruktur**
 - Analyse der Hierarchie **syntaktischer Einheiten** (Phrasenstrukturgrammatik im weiteren Sinne)
 - Strukturinformationen in Knoten des Syntaxbaums (Konstituenten = phrasale Einheiten)

- **Dependenzstruktur**

- Analyse der hierarchischen **syntaktischen Abhängigkeitsrelationen** zwischen Wörtern (Wortgrammatik)
- Strukturinformationen in Kanten des Syntaxbaums (grammatische Relationen als funktionale Kategorien)

- **formale Grammatik**

- mathematische Struktur zur Modellierung natürlichsprachlicher Satzstruktur
- kontextfreie Grammatik (CFG) als Phrasenstrukturgrammatik im engeren Sinne (PSG)

- **Parsing**

- algorithmische Verarbeitung von formalen Grammatiken zur automatischen Satzstrukturanalyse
- Erkennung der Wohlgeformtheit (Grammatikalität) einer Eingabe
- Wiedergabe der syntaktischen Struktur (Syntaxbaum)

Sitzung 3: Syntaktische Kategorien

- Syntaktische Einheiten = Konstituenten
 - *Wörter - Phrasen - Sätze*
 - *Wörter = elementare Einheiten*
 - *Phrasen = Gruppen von Wörtern, Erweiterung um Phrasen-*
kopf
 - *Feststellbar durch Konstituententests*

- **Kategorisierungen syntaktischer Einheiten**

→ ***syntaktische Kategorie*** = Menge von syntaktischen Einheiten mit gleichen ***morphosyntaktischen Eigenschaften*** (Abstraktionsklasse)

→ Klassen primär definiert über ***Austauschbarkeit im gleichen Kontext***

→ ***sprachabhängig!***

- **Wortarten = Lexikalische Kategorien**

- *Hauptkategorien: **Nomen, Verb***

- *Modifikatoren: **Adjektiv, Adverb***

- *Nominale Begleiter und Proformen: **Pronomen, Determinativ***

- *Weitere Kategorien: **Adposition, Konjunktion, Partikel***

- **Phrasenkategorien**

- *definiert durch **Wortart des Phrasenkopfs***

- *nur bestimmte Wortarten sind **phrasenbildend***

- *Phrasen können **komplex** sein, d. h. andere Phrasen enthalten*

- ($PP = P + NP$; $NP = NP + PP$)*

- ***Nominal-, Verbal-, Adjektiv-, Adverb-, Adpositional-Phrase***

Sitzung 4: Syntaktische Relationen: Konstituenz

- **Konstituentenstruktur (auch: Phrasenstruktur)**

- Konstituenz = **Teil-Ganzes-Beziehung** zwischen sprachlichen Einheiten (Konstituenten)

- Relation der **unmittelbaren Dominanz** zwischen Einheit und ihren unmittelbaren Konstituenten

- **in phrasalen Einheiten** können neben lexikalischen auch **phasale Einheiten anderer oder gleicher Kategorie** vorkommen

- ⇒ **hierarchischer, rekursiver Strukturaufbau**

- **Merkmalsvererbung** vom **Kopf als Phrasenkern** an Phrase

- Köpfe werden im Syntaxbaum nach oben weitergereicht (**Perkolatio**

- Analyse **diskontinuierlicher Phrasen** über *traces* (Spuren)

- **Kontextfreie Grammatik**

- formale Grammatik mit **kontextfreien Regeln**

- verwendet zur **Modellierung der Konstituentenstruktur natürlicher Sprache**

- Phrasenstrukturgrammatik (**PSG**) im engeren Sinne

- beschreibt Regeln der **Kombination von lexikalischen und phrasalen Kategorien** (nichtterminale Symbole) zu **phrasalen Kategorien und Sätzen** (Startsymbol S)

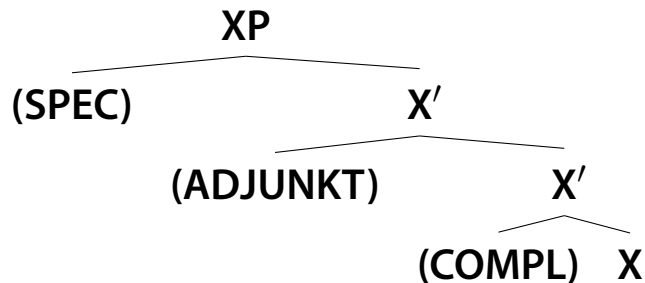
- Eine syntaktische Struktur (**Syntaxbaum**) wird von einer Grammatik erfüllt, wenn eine **Ableitung aus den als Produktionsregeln** aufgefassten Regeln der Grammatik existiert

- **X-Bar-Schema**

→ Beschränkung der Struktur: **binäre Verzweigung**: $A \rightarrow B C$

→ Einführung phrasaler **Analyseebene zwischen Phrase und Kopf (X')**

→ gleichartiges Schema für alle Phrasen:



→ **Komplement**: Schwester von Kopf, Tochter von X'

→ **Adjunkt**: Schwester von X, Tochter von X'

→ **Spezifizierer**: Schwester von X', Tochter von XP

- **CFGs als Konstituentenstrukturmodell**

- Modellierung des **hierarchischen, rekursiven Aufbaus** natürlicher Sprache aus lexikalischen und phrasalen Kategorien
- **X-Bar: Differenzierung Argument-Adjunkt-Spezifizierer**
- Nichtberücksichtigung von Morphosyntax und Subkategorisierung → **Übergenerierung**

- **Erweiterungen von CFGs**

- Einführung **komplexerer atomarer Kategorien**
- **Merkmalsstrukturen** (Unifikationsgrammatiken)
- Auswahl durch **probabilistisches Modell** (PCFG)

Sitzung 5: Syntaktische Relationen: Dependenz

- **Dependenzstruktur**

- Untersuchung von **Abhängigkeiten im Satz** (zwischen dem Vorkommen und der Form von Wörtern)

- **Dependenzrelation**: binäre asymmetrische Relation zwischen Wörtern (Kopf und Dependent)

- 2 Typen von Abhängigkeiten:

- **Rektion** (*bilaterale Abhängigkeit*): → **Komplemente**

- **Modifikation** (*unilaterale Abhängigkeit*): → **Modifikatoren**

- **Valenzgrammatik**: Untersuchung ausgehend vom Verb

- **Komplement** (valenzgrammatisch: **Ergänzung / Aktant**)
 - **obligatorischer Dependent** (gefordert vom Kopf)
 - aber: kann **fakultativ** sein
- **Modifikator**
 - **optionaler Dependent**
 - hängt ab von Kopf, aber wird nicht vom Kopf gefordert
 - *verbal: Adjunkt* (valenzgrammatisch: **Angabe / Zirkumstant**)
 - *nominal: Attribut*

- **Dependenzrelationen als syntaktische Funktionen**
 - **Kategorisierung der Dependenzrelationen nach syntaktischem Verhalten der Dependenden**
 - Feststellung der **syntaktischen Funktion** einer Einheit, die sie in Bezug auf ihren Kopf einnimmt
- **Grammatische Relationen → syntaktische Funktion verbaler Dependenden**
 - **Subjekt:** Kernargument intransitiver Satz, Kongruenz mit Verb
 - **Objekt:** passivierbares Patiens-Argument transitiver Satz
 - **indirektes Objekt:** Recipient-Argument ditransitiver Satz
 - **Adverbial:** nicht-zentrales, peripheres Argument

- **Attributfunktionen → Syntaktische Funktion nominaler Modifikatoren**
→ Adjektiv-/Partizipial-Attribut, Präpositional-/Genitiv-Attribut, Apposition, Determinativ, Attributsatz

- **Dependenzgrammatik**

- formale Repräsentation als **gerichteter Graph**

- **Wortgrammatik**

- Strukturinformation in den Kanten (Relationen)

- Transformation Konstituenten- in Dependenzstruktur möglich

- Hauptvorteil gegenüber PSGs: **Grammatische Funktionen direkt kodiert**

- **Übersicht: Adverbial, Angabe, Ergänzung, Präpositionalobjekt**

Dependenztyp	syntaktische Funktion	Auftreten	Form	Beispiel
Komplement/Ergänzung	Subjekt / Objekt (auch: Präpositionalobjekt)	valenzgefordert	valenzgefordert	<i>denken an jmd.</i>
Komplement/Ergänzung	Adverbial	valenzgefordert	<i>nicht</i> valenzgefordert	<i>ins/zum/... Haus gehen</i>
Adjunkt/Angabe	Adverbial	<i>nicht</i> valenzgefordert	<i>nicht</i> valenzgefordert	<i>Es regnet im Park</i>

– Adverbiale Angabe vs. Präpositionalobjekt:

- * *Erwartet **auf dem Berg*** (adverbiale Angabe: wo?, Dativ)
auf die Sonne (Präpositionalobjekt: worauf?, Akk.).
- * Angabe ist optional: *Er wartet auf die Sonne.*
- * Das präpositionale Komplement von *warten* ist fakultativ: *Er wartet auf dem Berg.*

– Adverbiales Komplement:

- * Verb verlangt keine Formeigenschaft!
- * *Erstellt die Blumen **auf den Tisch*** (adverbiales Komplement).
- * *Erstellt die Blumen **ins Wasser*** (adverbiales Komplement).

Sitzung 6: Morphologische Form syntaktischer Funktionen

- **Sprachliche Ausdrucksmittel syntaktischer Funktionen**
 - strukturell über **Wortstellung**
 - morphologisch über **Flexionsmorphologie**
- **morphologische Kodierung grammatischer Relationen über:**
 - **Kasus:** Markierung der Funktion der Relation zwischen Verb und Dependent durch **Marker am Dependent** (Rektion)
 - **Agreement:** Markierung der Funktion der Relation zwischen Verb und Dependent durch **Merkmalskongruenz**

- **Funktionale Kategorien**

→ Funktionale Syntax: Untersuchung der **systematischen Variation von morphosyntaktischer Kodierung** mit **semantischer und pragmatischer Rolle**

→ **Diathesen**: syntaktische Operation der **Manipulation der Abbildung semantischer Rollen** auf Grammmatische Relationen

→ *Passivierung: Promotion des Patiens-Arguments in Subjektposition*

→ **Topik-Fokus-Struktur**: **kontextabhängige**, pragmatische **Struktur der Äußerung**, die u. a. über syntaktische Operationen wie **Linksversetzung** oder **Cleftsätze** angezeigt werden kann

- **Morphosyntaktische Typologien**

- **Varianz** in der Kodierung syntaktischer Funktionen im **Sprachvergleich**

- Systematische **Differenz in der Abbildung semantischer Rollen** auf Grammatische Relationen: **Akkusativ- vs. Ergativsprachen**

- **Aktiv-Sprachen** wie das Georgische kodieren primär die **semantische Rolle**

- **Topik-prominente Systeme** wie das Japanische kodieren primär die **pragmatische Rolle**

- **Merkmalsstrukturen**

- **formale Repräsentation von grammatischen Kategorien**
- **atomare oder komplexe Werte** (Merkmalsstruktur als Wert, z.B. für Bündelung von Agreementmerkmale)
- **Beschreibung** von lexikalischen Einheiten und Kategorien als komplexe Objekte, die über **Merkmale** definiert sind:

Wortformen: *Hund* $\left[\begin{array}{cc} \text{CAT} & N \\ \text{AGR} & \left[\begin{array}{cc} \text{NUM} & SG \\ \text{GEN} & MASK \end{array} \right] \end{array} \right]$, *der* $\left[\begin{array}{cc} \text{CAT} & DET \\ \text{AGR} & \left[\begin{array}{cc} \text{NUM} & SG \\ \text{GEN} & MASK \\ \text{CASE} & NOM \end{array} \right] \end{array} \right]$

lexikalische Kategorien: $\left[\text{CAT} \quad N \right]$ $\left[\text{CAT} \quad DET \right]$

(unterspezifiziert)

- **Verwendung in Syntaxanalyse**

→ Verwendung in **PSG-Regeln** zusammen mit **Constraintregeln** zum **Ausdruck von Abhängigkeiten** zwischen durch unterspezifizierte Merkmalsstrukturen repräsentierten **Kategorien**

→ nominales Agreement: **Beschränkung** der durch die PSG-Regel repräsentierten **Kombination** von Determinativ und Nomen **auf Übereinstimmung im AGR-Merkmal**:

$$[\text{CAT} \quad \text{NP}] \rightarrow \begin{bmatrix} \text{CAT} & \text{DET} \\ \text{AGR} & \boxed{1} \end{bmatrix} \begin{bmatrix} \text{CAT} & \text{N} \\ \text{AGR} & \boxed{1} \end{bmatrix}$$

- **Unifikation**

→ **Constraintregel**: entspricht Anweisung auf Durchführung von **Unifikation** zur **Feststellung der Vereinbarkeit**

→ nominales Agreement: Feststellung der Vereinbarkeit dieser **AGR-Teil-Merkmalstrukturen**:

$$\begin{bmatrix} \text{NUM} & \text{SG} \\ \text{GEN} & \text{MASK} \\ \text{CASE} & \text{NOM} \end{bmatrix} \sqcup \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{GEN} & \text{MASK} \end{bmatrix} = \begin{bmatrix} \text{NUM} & \text{SG} \\ \text{GEN} & \text{MASK} \\ \text{CASE} & \text{NOM} \end{bmatrix}$$

Sitzung 7: Unifikationsgrammatiken

- **Unifikationsbasierte Erweiterungen von CFGs**
 - Modellierung von **Agreement-, Rektions- und Subkategorisierungs-Constraints**
 - Modellierung von wortstellungsbezogenen Abhängigkeiten wie **Subjekt-Verb-Inversion** und *long distance dependencies*

- **Subkategorisierung**

→ Differenzierung der Klasse der Verben **nach Anzahl und Art ihrer Argumente** (z. B. auch nach abhängigen Sätzen)

→ **Subkategorisierungsprinzip**: Verb kann nur in **Umgebung** auftreten, die **seinem Subkategorisierungsrahmen entspricht**

→ mit **kontextsensitiven Regeln** oder als **Merkmalsconstraint** modellierbar

Sitzung 8: Komplexe Satzkonstruktionen und Wortstellung

- **Wortstellung**
 - **strukturelle** Kodierung syntaktischer Funktion
 - **Positionierung** syntaktischer Einheiten
- **Wortstellungssyntax des Deutschen**
 - Verbstellungstypen: **V1, V2, VE**
 - **Verbstellungs-Split** kodiert Satzfunktion:

- V2 (Verbzweitstellung): **Aussagesatz**
- V1 (Verberststellung): **Aufforderungs-/Wunsch-/Fragesatz**
- VE (Verbendstellung): **Nebensatz**

- **Stellungsfeldermodell**

- **Lineares Modell** der Wortstellung des Deutschen, Analyse der Stellungsmöglichkeiten der Satzglieder
- **Einteilung in Felder**, ausgehend vom flektiertem Verbalkomplex als **Satzklammer**
- **diskontinuierliche Verbalphrase** kennzeichnend für Neu-hochdeutsch
- bei Verbzweitstellung kann **ein** beliebiges Satzglied ins Vorfeld gestellt werden (**Topikalisierung** bzw. **Fokussierung**)
- **Topik-Es** als Platzhalter wenn Vorfeld-Position unbesetzt
- **Wortstellungsregeln** der Anordnung von Satzgliedern im Mittelfeld, insbesondere '**Thema-vor Rhema**' (pragmatische Wortstellung)

- **Komplexe Satzkonstruktionen**

- **Einfache Sätze als Konstituenten von komplexen Sätzen**
- **Koordination = gleichrangige Verbindung**: Sätze bilden als **Ko-Konstituenten** einen komplexen Satz
- **Subordination = Einbettung** eines Satzes als **Satzglied des übergeordneten Satzes** (Matrixsatz)
- in **Dependenzanalyse**: **Verb des eingebetteten Satzes ist Dependent** von Verb des übergeordneten Satzes
- in **Konstituentenanalyse**: je nach Typ **andere Position im Syntaxbaum**: z.B. Objektsatz als Subkonstituente von VP
- **rekursive Einbettung**

- **Typen von eingebetteten Sätzen**
 - **Komplementsatz:** Subjekt- und Objektsatz
 - **Adverbialsatz**
 - **Attributsatz:** Relativsatz, adnominaler Substantivsatz
 - **Prädikativsatz**
- **Infinite Satzkonstruktionen**
 - können wie finite Sätze **als Satzglied auftreten**
 - **nicht-flektiert**, kein Subjekt
 - **Kontrolle durch Subjekt oder Objekt des Matrixsatzes**

- **Verbale Konstruktionen des Deutschen**

- **Hilfs-und Modalverben (Auxiliare)** bilden mit **infiniter Verbform** einen **Verbalkomplex**

- **Auxiliar** als **linker Teil der Satzklammer**

- **Satzklammer**: Aufteilung Satz in Felder → Vorfeld, Mittelfeld, Nachfeld

- **Kopula** als **prädikatives Hilfsverb**, das mit einem Nomen, Adjektiv oder Satz eine **Eigenschaft** über das Subjekt oder Objekt prädiziert

Sitzung 9: Parsing-Algorithmen

- 2 Klassen von **Parsing-Algorithmen: top-down / bottom-up**
 - top-down: **PREDICT + SCAN** (*Regelanwendung + Abgleich*)
 - probiert jede anwendbare Ersetzungsregel aus
 - im Problemfall: ***Backtracking*** *notwendig*
 - bottom-up: **SHIFT + REDUCE** (*Einlesen + Regelnw. rückwärts*)
 - verschiebt Token auf **Stapel** u. führt sie auf Regeln zurück

- **Vergleich** top-down vs. bottom-up:
 - Start der Analyse:
 - **Startsymbol** vs. **1. Wort der Eingabe**
 - Schwäche:
 - **strukturelle** vs. **lexikalische** Ambiguität
 - im Extremfall für beide **exponentielle Laufzeit**

- **Earley Parser: Top-Down-Parsing mit Extras**
 - 3 Operationen: **PREDICTION + SCANNING + COMPLETION**
 - *Voraussage*: wenn . vor Nichtterminal
 - *Überprüfung*: wenn . vor Terminal
 - *Vervollständigung*: wenn . letzte Position
 - **Zwischenergebnisse** werden in Datenstruktur (**Chart**) gespeichert (**Dynamische Programmierung**)
 - auch für ambige Grammatiken **maximal polynomielle Laufzeit**
 - **erweiterbar zu merkmalsbasiertem Parsing**
 - aber: Unifikation ist sehr **rechenaufwändig**

- **Statistisches Parsing:**
 - **nicht alle möglichen Ableitungen** werden ausprobiert, die **wahrscheinlichste** soll bestimmt werden
- per Hand geparste Sätze dienen als **Trainingsdaten**
- Eingabe wird in Merkmale umgewandelt (***Feature Extraction***)
- **Merkmalsvektoren** werden durch **gelernte Gewichte** auf eine **Wahrscheinlichkeitsverteilung** abgebildet
- die **Likelihood** der Trainingsdaten soll **maximiert** werden

Sitzung 10: Unifikation

- **Subsumption:**
 - für Typen definiert durch die \sqsubseteq -Relation
 - bei Merkmalstrukturen muss es **alle Knoten der "allgemeineren" Merkmalstruktur auch in der spezifischeren geben (+ compatible Typen)**

- **Unifikation:**

- sowohl für Typen als auch Merkmalstrukturen **kleinste obere Schranke in der Subsumptionsbeziehung**

- für **Merkmale** zweischrittig:

- 1. **Identifikation äquivalenter Knoten**

- 2. **Unifikation ihrer Typen**

- **Bedingungen:**

- Pfade sind **Ketten von Merkmalen**
- Beschreibungen legen die **Menge von Merkmalstrukturen**, die sie erfüllen, **eindeutig fest**
- **Beschreibungen** werden im NLTK durch ihren **allgemeinsten Erfüller** ausgedrückt