Linguistische Informatik

Gerhard Heyer
Universität Leipzig
heyer@informatik.uni-leipzig.de

UNIVERSITÄT LEIPZIG
Institut für Informatik

ASV
Automatische Sprachverarbeitung

Syntax

Problem:

Gegeben eine Menge von Wortformen (oder Wörtern), welche Reihenfolgen sind grammatikalische oder sinnvolle Sätze?

Ein Sack von Wörtern

die Leipzig scheint warm in Sonne
in Leipzig warm Sonne die scheint
warm die in Sonne scheint Leipzig

Die Sonne scheint in Leipzig warm

Syntax

Gegenstand:

Wörter bzw. Wortformen und deren Kombination zu Sätzen

***Definition* (Chomsky):**

“Syntax is the study of principles and processes by which sentences are constructed in particular languages”

- ***Grammatik der Einzelsprachen***
  (z.B. Deutsch, Englisch, …)
- ***Universalgrammatik***
  (Bedingungen menschlichen Sprachverstehens)

Automatische Sprachverarbeitung

Ziele einer Syntax für natürliche Sprachen

1) Syntaktische Struktur einer Einzelsprache

- **welche Elemente ?**
- **wie kombiniert ?**

2) Grammatik einer Einzelsprache

- **syntaktische Merkmale der Einzelsprache**
- **Formalisierung der Grammatik**
- **Bedingungen der Grammatikalität**

3) Prinzipien der Universalgrammatik

- **Merkmale menschlicher Sprachfähigkeit**
- **Grundlagen und Grenzen ihrer Formalisierung**

Requirements on the syntactic description

1. ***Linear sequence of words (time linearity)***
   - *only one word after the other can be uttered and perceived*
2. ***Sentences contain syntactic ambiguities***
   - *„Flying planes can be dangerous“ (Chomsky)*
3. ***Hierarchic structuring of phrases***
   - *Some expressions are „closer“ to each other than others, e.g. „flying planes“, „can be“ rather than „planes can“*
   - *Empirical tests for determining constituents, e.g.*
     - *substitution*
     - *permutation*

Formale Sprachen

a) Erzeugungsgrammatik
b) Erkennungsgrammatik

Chomsky - Grammatik

< s, T, N, P >

T = Mengen der Terminalen Symbole
N = Mengen der Nichtterminale
s = Startsymbol, $s \in N$
P = Produktionsregeln
(Ersetzungsregeln)

Als Vokabular V bezeichnen wir $V = N \cup T$

Ersetzungsregeln

$u \rightarrow v$

Eine Kette w_2 ist aus einer Kette w_1 direkt ableitbar, wenn es eine Ersetzungsregel

$p \rightarrow q$

gibt und sich w_2 von w_1 dadurch unterscheidet, daß die Teilkette p in w_1 durch die Teilkette q in w_2 ersetzt ist.

Allgemein ist w_n ableitbar aus w_1, wenn es eine Folge direkt ableitbarer Ketten wie folgt gibt:

$w_1 \ldots w_i \ldots w_n$

Typ 0 : unbeschränkte Ersetzungssysteme

keine Beschränkungen an die Form der Regeln

Typ 1 : kontext-sensitive Grammatik

x A y → x v y
A∈ N, x, y∈V*, v∈V⁺
Beschreibung: **Ersetze A durch v im Kontext x, y**

Beispiel
Mit einer kontext-sensitiven Grammatik lässt sich die Menge $\{a^n b^n c^n\}$ ableiten.
Grammatik: **T={ a, b, c }, N={ s, x, y, z }**
Regeln:
s → abc
s → axbcy
x → axbc
x → az
cy → yc
zb → bz
cb → bc
zy → bc

Typ 2 : kontext-freie Grammatik

$A \rightarrow v$

$A \in N, v \in V^*$

Beispiel 1

Mit einer kontext-freien Grammatik lässt sich die Menge $\{a^n b^n\}$ ableiten.

Grammatik: **T={a, b}, N={s, x}**

Regeln:

$s \rightarrow ab$

$s \rightarrow axb$

$x \rightarrow ab$

$x \rightarrow axb$

Beispiel KF 2

Mit einer kontext- freien Grammatik lässt sich die Menge $\{a^n\ b^k\ c^n\}$ ableiten.

Grammatik: **T={a, b, c}, N={s, x, y}**

Regeln:

s → ac
s → axc
x → axc
x → ayc
y → yb
y → b

ABER

Mit einer kontext- freien Grammatik lässt sich nicht die Menge $\{a^n\ b^n\ c^n\}$ ableiten.

Typ 3 : reguläre Grammatik

Regeln der Form

A → a

A → B b (links-linear) bzw.

A → b B (rechts-linear)

mit A, B ∈ N, v ∈ V*

Beispiel

Mit einer regulären Grammatik lässt sich nicht die Menge $\{a^n b^n\}$ ableiten.

Ein Beispiel

Beispiel: ***Der Bayer stellt die Maß auf den Tisch.***

1) Welche Elemente ?

Wortformen (nicht Wörter):
{der,die, den, Bayer, Maß, Tisch, stellt, auf}

im Vergleich dazu die entsprechenden Wörter: {*der, die, Bayer, Maß, Tisch, stellen, auf*}

2) Wie kombiniert ?

Tests für Konstituenten: Was sind richtige Strukturen ?

1. Ersetzungsprobe

Wortfolgen, die sich füreinander ersetzen lassen, ohne dass sich an der Grammatikalität des Ganzen etwas ändert, sind (möglicherweise) Konstituenten.

2. Pronominalisierungstest

Pronomina: Er, sie, dort, damals …

Was sich pronominalisieren läßt (worauf man sich mit einer *Proform* beziehen kann), ist eine Konstituente.

Die Straßenbahn stand **durch den Zusammenstoß mit einem Auto**
sie ***dadurch***
auf der Kreuzung des Augustusplatzes.
dort

"Ede will einen Pudding essen." *Das* will ich auch.

Tests für Konstituenten II

3. Weglassprobe

„ die Straßenbahn stand…“

4. Fragetest

wer, wann, wo, wohin, wie, womit, was, warum, …

5. Koordinierungstest (Verbindung durch “und”)

6. Verschiebeprobe

Auf der Kreuzung des Augustusplatzes stand die Straßenbahn durch den Zusammenstoß mit einem Auto.

Konstituentenstruktur-Syntax

Der syntaktische Aufbau sämtlicher Sätze einer Sprache (und nur dieser) ist dann vollständig beschrieben, wenn festgelegt ist,

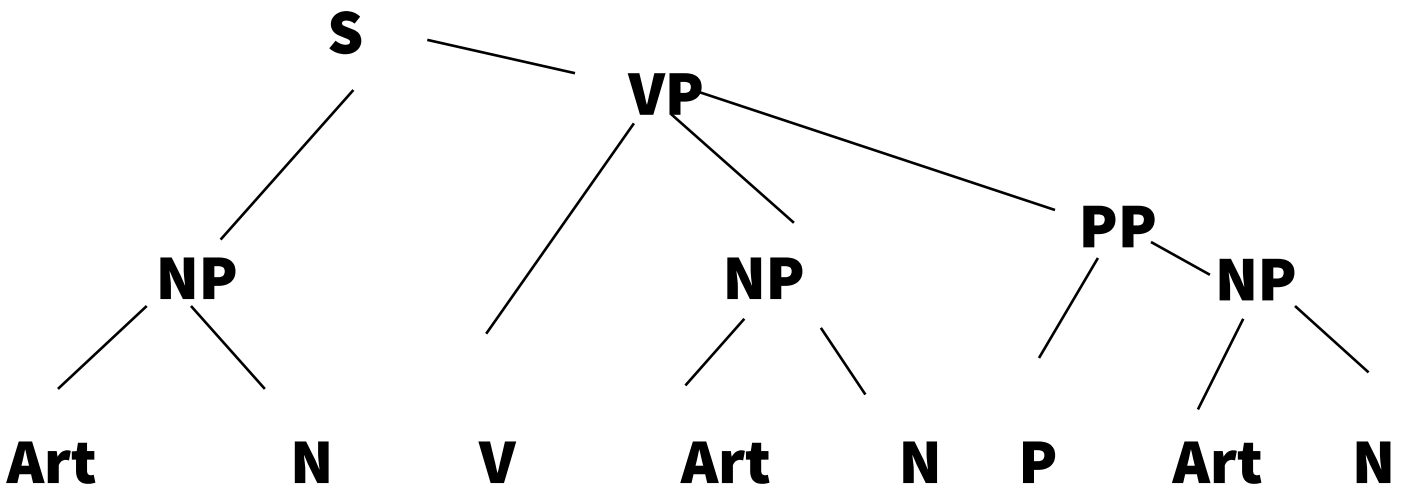
(a) zu welchen Kategorien die Wörter der Sprache gehören und

(b) welche Folgen von Kategorien von jeweils welcher Kategorie *unmittelbar dominiert* werden können.

Der Beispielsatz richtig zerlegt

Regeln:

S → NP VP
NP → Art N
VP → V NP PP
PP → P NP



der Bayer stellt die Maß auf den Tisch

***Abarbeitung (Parsing)*: Top-Down oder Bottom-up**

Modell für die Sprachgenerierung

1. **Erzeugen der syntaktisch richtigen Struktur**
2. **Ersetzen der Kategorien durch (passende) Terminale**

S → NP VP
Art N VP
der N VP
der Bayer VP
.
.
.
der Bayer stellt die ...

Wichtig: Reihenfolge ist starr!

Automatische Sprachverarbeitung

Angemessenheit von syntaktischen Beschreibungen

- ***Beobachtungsadäquatheit***

  (Erzeugung der richtigen terminalen Ketten)
- ***Beschreibungsadäquat***

  (richtige Terminalketten plus intuitiv richtige Struktur)
- ***Erklärungsadäquat***

  (Beschreibungsadäquat plus richtige Grammatik, z.B. Rekursivität)

Dependenzgrammatik

Was ist die "richtige" Grammatik für natürliche Sprachen?

Alternative zu Konstituenten: Dependenzen

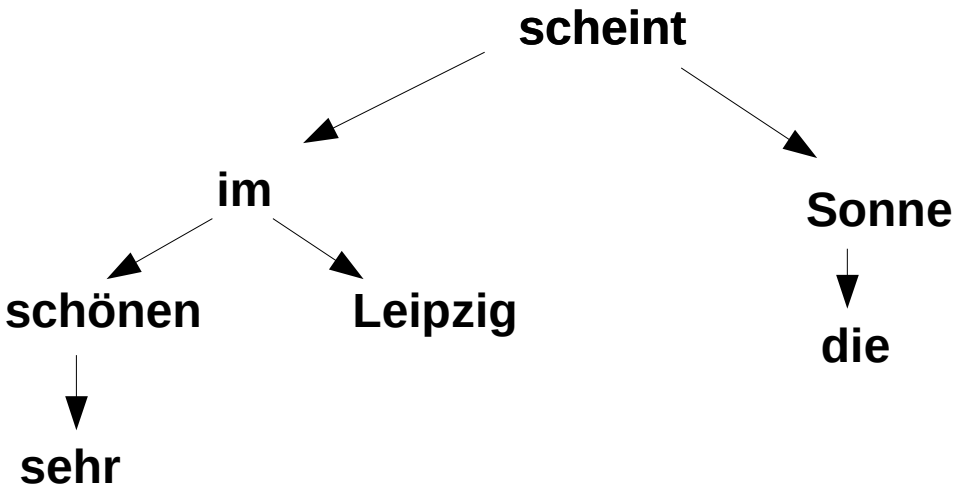
Grundidee ist, einen Satz nicht in Konstituenten zu zerlegen, sondern die zwischen den Ausdrücken bestehenden Abhängigkeiten darzustellen *(Dependenzen)*

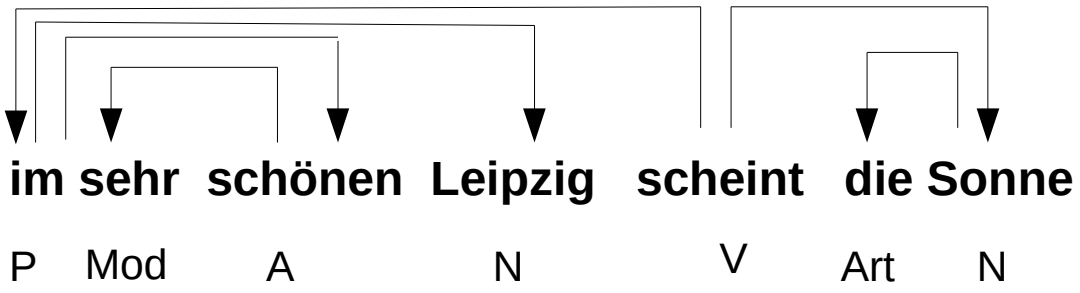
Beispiel: „Die Sonne scheint im sehr schönen Leipzig“
Wenn wir „schönen“ löschen, müssen wir auch „sehr“ löschen, „sehr“ ist also abhängig von „schönen“

Dies ist auch gleich ein Test für Dependenzen

Automatische Sprachverarbeitung

Dependenzgrammatik





ASV Automatische Sprachverarbeitung

Dependenzgrammatik

Wesentlich ist eine Unterscheidung zwischen *bedeutungstragenden* und *funktionalen* Wörtern

Der Satz wird von dem Verb dominiert

Dependenzregeln für den Beispielsatz:

1. Fin → N * V

2. N → (D) (A) *

3. V → * (P)

4. P → * N

5. A → Mod *

(* bezeichnet die Position des übergordneten Knotens)

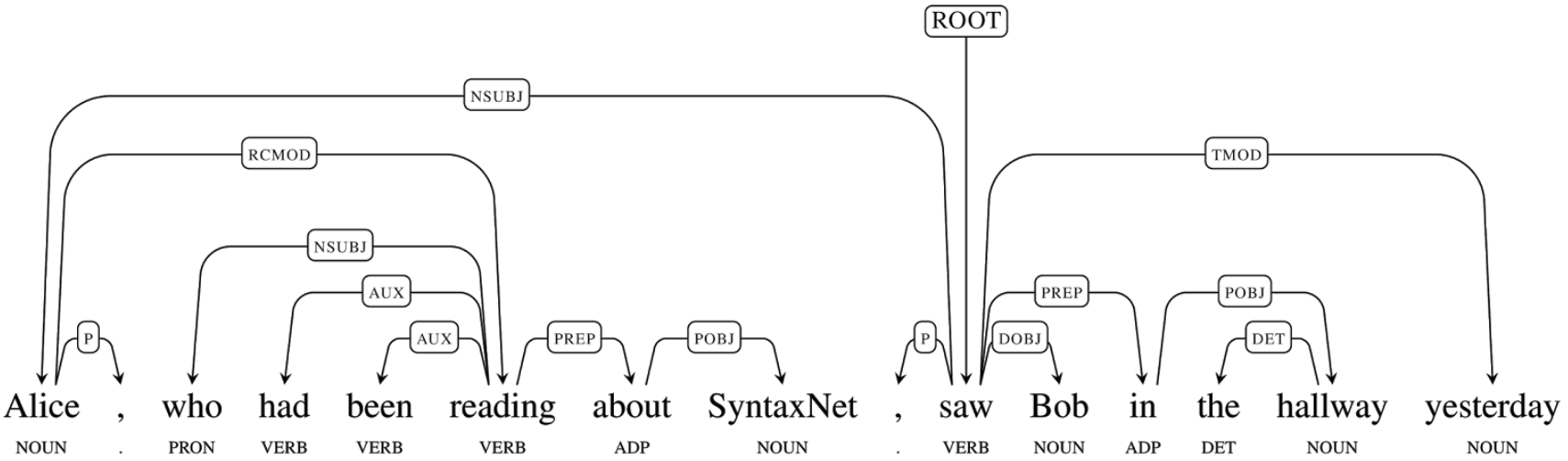
Dependenzgrammatik

Dependenzen werden oft verwendet als Grundlage für Information- und Relationenextraktion

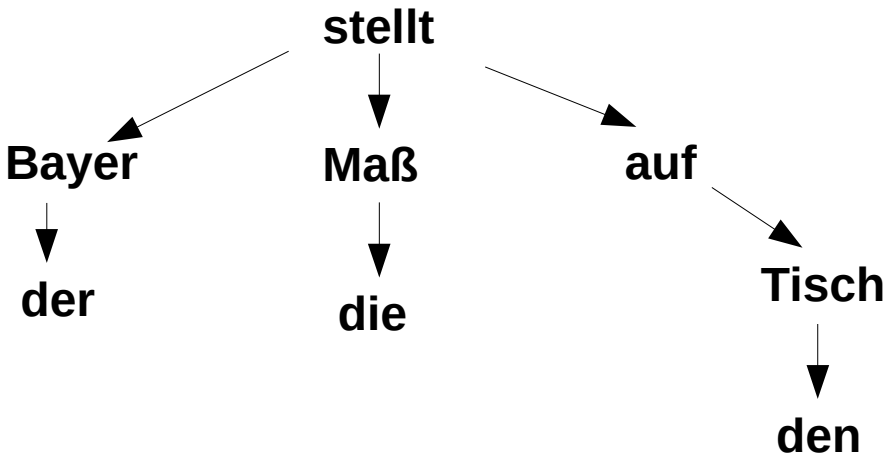
Parsen meist „middle-out“

vgl. SyntaxNet

https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html



Dependenzgrammatik für anderen Beispielsatz



Die Grammatikregeln müssen entsprechend ergänzt werden:

3. V → * (N) (P)

Automatische Sprachverarbeitung

Beispielanalysen

In der Forschungsinfrastruktur CLARIN-D finden sich zahlreiche Demo-Beispiele für Konstitutenten- und Dependenzanalysen

https://www.clarin-d.net/de/auswerten/web-basierte-analysewerkzeuge

Komplexitätsfragen

Was ist die "richtige" Komplexität für natürliche Sprachen?

a) unbeschränkt? OK aber zu wenig Struktur bzw. Beschränkung, daher zu wenig Erklärungswert

b) regulär ? zu schwach wegen rekursiven Strukturen z.B.

A={the cat, the dog, the rat, the elefant, ...}
B={chased, bit, admired, ate, befriended, ...}

Die Sprache
x^n y^{n-1} *died* mit x ∈ A und y ∈ B
ist wohlgeformtes Englisch, aber nicht regulär.

Komplexitätsfragen 2

1) Richtige Struktur: **Reguläre Sprachen und Rekursivität**

Rekursive Präpositionalphrasen (NP -> NP PP, PP -> P NP) und Nominalphrasen (N -> A N).

Chomskys Beipiel (1957)

1) Der Mann s_1 liest das Buch s_2
2) Das Buch ist 1957 geschrieben
3) Der Mann s_3 empfiehlt das Buch s_2

Jede der beiden Satzvariablen s_1 und s_3 können die Sätze 1 und 3, s_2 den Satz 2 als Wert nehmen.

Rekursive Verknüpfung der Sätze:

Der Mann,
der das Buch,
das 1957 geschrieben worden ist,
liest,
empfiehlt das Buch.

Das ist eine Struktur der Form $a^n b^n$, also nicht regulär beschreibbar.

Phrasenstrukturgrammatik kontextfrei oder kontextsensitiv ?

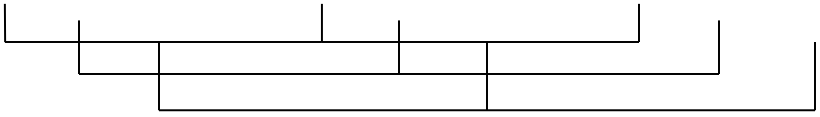
Peter a1
bzw. Maria a2
bzw. Paul a3

wandern b1
bzw. schwimmen b2
bzw. fahren Fahrrad b3

nach Naumburg c1
bzw. an der Ostsee c2
bzw. im Rosenthal c3

liefert eine *nicht kontextfreie* Konstruktion vom Typ
a^n b^n c^n
mit folgender Interdependenz:

a1 a2 a3 ... b1 b2 b3 ... c1 c2 c3



Ergänzende Literatur

Grewendorf, Hamm, Sternefeld, Sprachliches Wissen, Suhrkamp Taschenbuch Wissenschaft swt 695

Barton, Berwick, Ristad, Computational Complexity and natural language, MIT Press: Cambridge (Ma) 1987