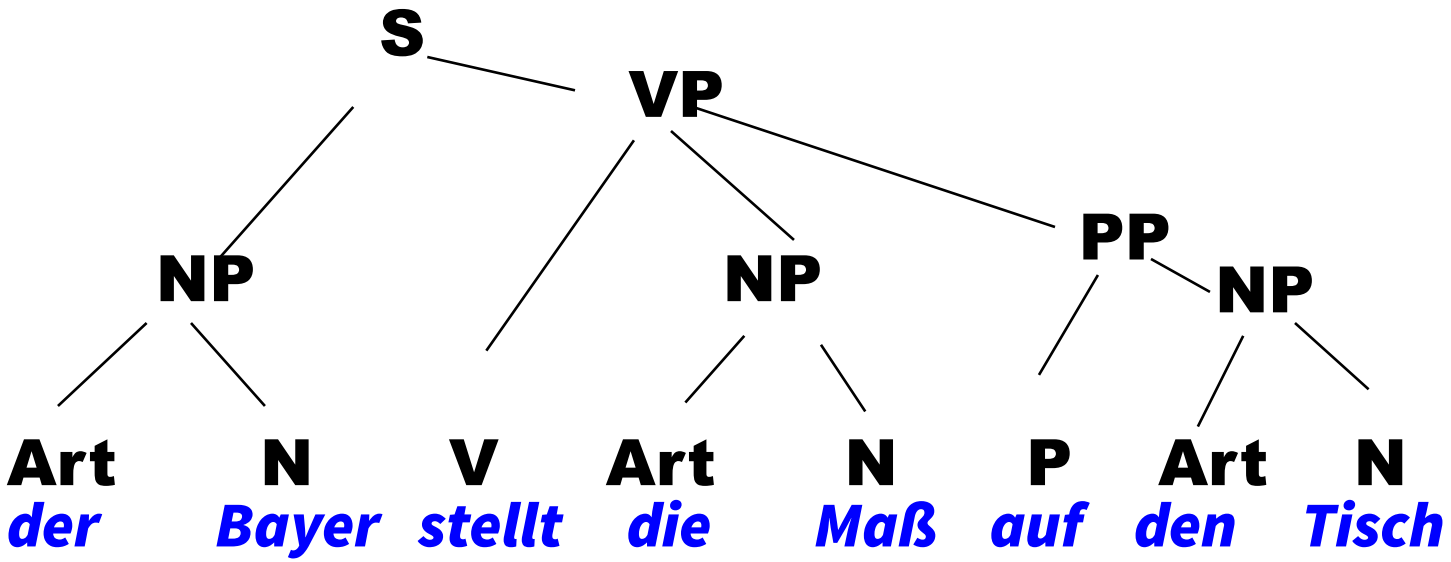
Automatische Sprachverarbeitung

Der Beispielsatz richtig zerlegt

Regeln:

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

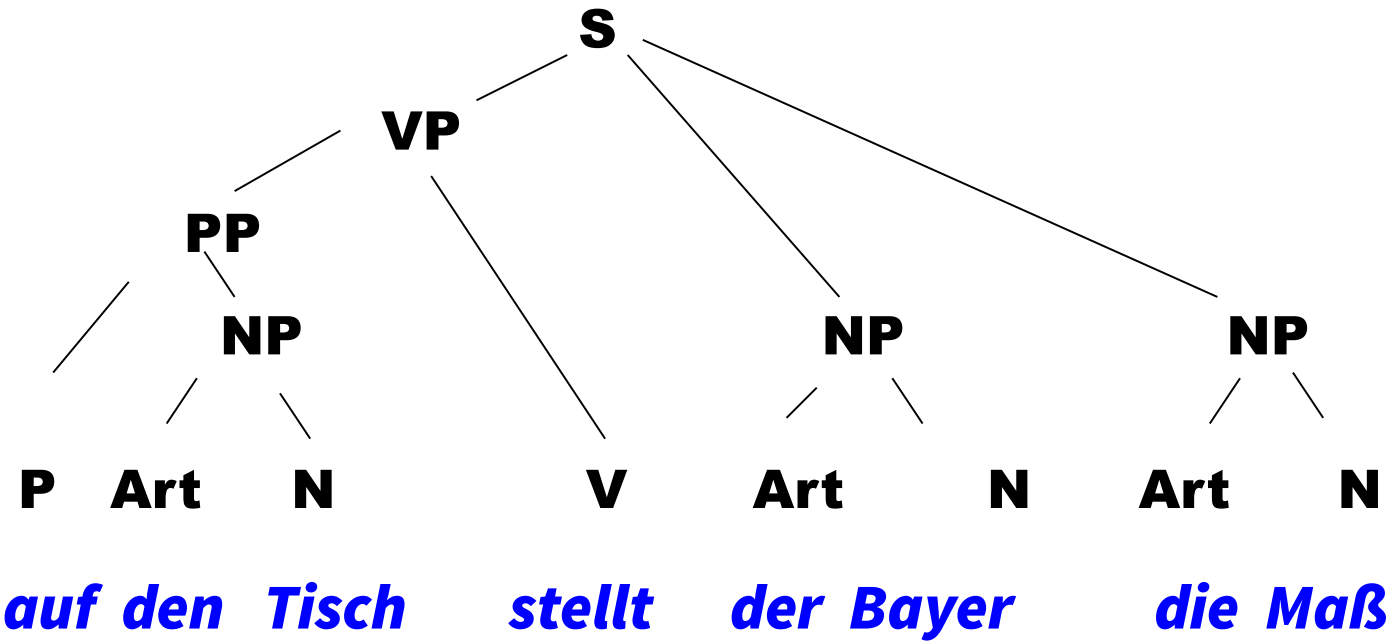


Abarbeitung (Parsing)*: *Top-Down* oder *Bottom-up

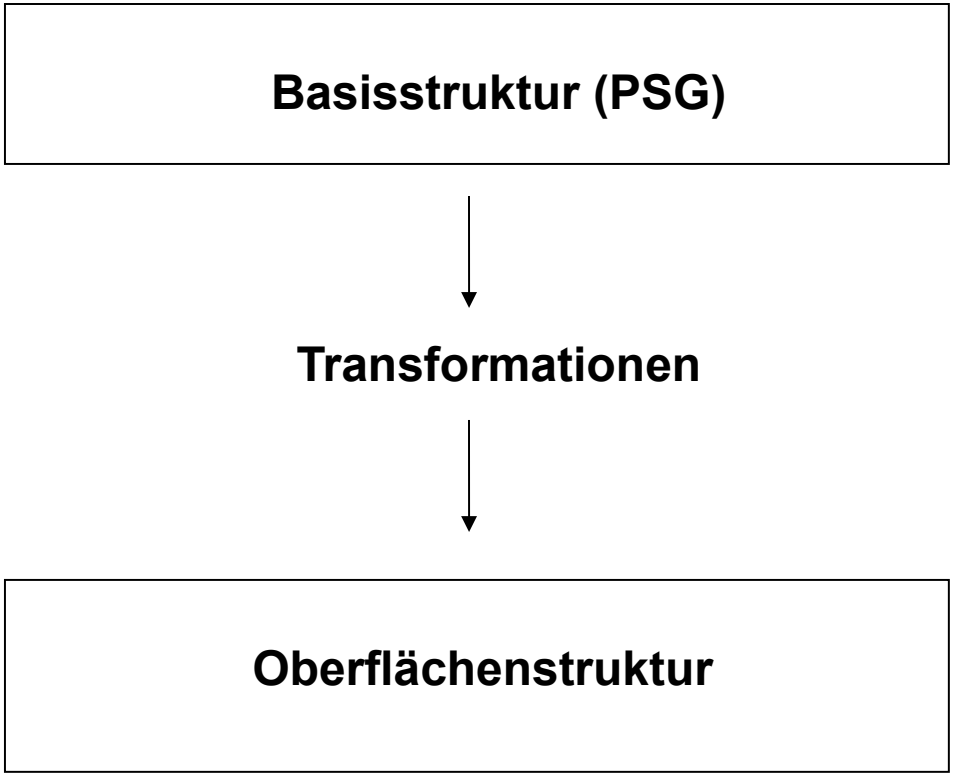
Problem freie Wortstellung

Regeln:

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



Traditioneller Lösungsansatz: Transformationen



Automatische Sprachverarbeitung

Notation für Transformationen

Strukturbeschreibung (structural description)

w1	w2	w3	w4	w5
[Die	Sonne]NP	[scheint	[in	Leipzig]PP]VP]
1	2	3	4	5

Strukturveränderung (structural change)

Fokus_Ort:	4	5	3	1	2
Frage:	3	4	5	1	2

Ggf. können neue und zusätzliche Elemente eingefügt werden

Zipfsches Gesetz: sprachabhängige Konstante c

Die textabhängige Konstante *k* ist abhängig von der Korpusgröße. Durch die Normalisierung dieses Parameters auf die Anzahl der in einem Korpus insgesamt vorkommenden tokens erhalten wir die sprachabhängige Konstante *c*, die für alle Korpora einer Sprache gelten sollte.

Die sprachabhängige Konstante *c* berücksichtigt anstelle der absoluten Häufigkeit eines Wortes seine relative Häufigkeit und wird bestimmt über:

$$r * f/N = k/N \sim c$$

Aus den oben genannten Daten des Projekts Deutscher Wortschatz mit N = 222 538 789 tokens und k = 18 000 000 ergibt sich so beispielsweise für das Deutsche eine Konstante von

$$c = 18.000.000/222.539.789 \sim 0.08$$

Sprachspezifische Konstante

Die sprachabhängige Konstante c sowie die Steigung der Ausgleichs-geraden a ist unterschiedlich für verschiedene Sprachen.

Sprache	c	a
Deutsch	0.0898	-1.040
Englisch UK	0.0935	-1.034
Finnisch	0.1058	-0.812
Arabisch	0.1140	-0.989
Chinesisch	0.1238	-0.979

web_public_2019_10K

Zipfsches Gesetz: Vorhersagen (auf Basis N und c)

N Gesamtanzahl aller Wortformen des Textes (tokens)

t Umfang des Vokabulars (types)

f/N relative Häfigkeit der Wortformen, die n mal auftreten

r_f größter Rang derjenigen Wortformen, die genau n mal auftreten

l_f Anzahl der Wortformen, die genau n mal auftreten

Es gilt:

$$r_f \times f/N = c \qquad \text{also:} \quad r_f = c \times N/f$$

Für das Vokabular gilt:

t ist so groß wie der größte Rang der häufigkeits-sortierten Liste. Falls Wörter mit Häufigkeit 1 vorkommen folgt damit:

$$t = r_1 = c \times N/1 = c \times N$$

Anteil von Wortformen, die genau n-mal auftreten

Für I_f gilt:

$$I_f = r_f - rf_{+1} = c \times N/f - c \times N/(f+1) = cN/(f*(f+1)) = t/(f*(f+1))$$

Für I_1 gilt insbesondere:

$$I_1 = t/2$$

→ **Die hälfte des Vokabulars eines Textes tritt wahrscheinlich nur 1 mal auf. (Achtung: Hälfte der types, nicht der token!)**

allgemein: Anteil der Wortformen, die genau f mal auftreten, am Vokabular eines Textes

$$I_f/t = (t/(f*(f+1))) / t = 1/(f*(f+1))$$

Bedingte Wahrscheinlichkeiten

Für bedingte Wahrscheinlichkeit der Kombination von zwei Wortformen **gilt:**

$$(5) \qquad P(w_j \mid w_i) == \frac{|w_i, w_j|}{|w_i|}$$

Bsp.: P(Schuhmann | Herr) = |Herr, Schuhmann| / |Herr| =
P(liest | Schuhmann) = |Schuhmann, liest| / |Schuhmann| =
P(ein | liest) = |liest, ein| / |liest| =
P(liest | Müller) = |Müller, liest| / |Müller| =

Bedingte Wahrscheinlichkeiten

Für das Aufeinanderfolgen dreier Wortformen ergibt sich:

$$(6) \qquad P(w_i, w_j, w_k) = P((w_i, w_j), w_k) = P(w_i, w_j) \cdot P(w_k \mid w_i, w_j)$$
$$= P(w_i) \cdot P(w_j \mid w_i) \cdot P(w_k \mid w_i, w_j)$$

**P(Herr, Schuhmann, liest) = P(Herr)*P(Schuhmann | Herr) *
P(liest | Herr, Schuhmann)**

Berechnungsgrundlage *n-gram Modell*

Annahme, dass nur die vorangehenden n-1 Wortformen von Einfluss auf die Wahrscheinlichkeit der nächsten Wortform sind, wobei ***n = 3 (daher tri-gram)***

(8) $P(w_n \mid w_1, \ldots, w_{n-1}) = P(w_n \mid w_{n-2}, w_{n-1})$

$$P(w_{1,n}) = P(w_1) * P(w_2 \mid w_1) * P(w_3 \mid w_{1,2}) * \ldots * P(w_n \mid w_{n-2}, w_{n-1})$$

$$= P(w_1) * P(w_2 \mid w_1) * \prod_{i=3}^{n} P(w_i \mid w_{i-2,i-1})$$

$$= \prod_{i=1}^{n} P(w_i \mid w_{i-2}, w_{i-1})$$

Für die Berechnung der Wahrscheinlichkeit eines ganzen Satzes gilt:

$$(10) \qquad P(w_1, w_2, w_3, \ldots, w_n)$$

$$= P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_1, w_2) \cdot \ldots \cdot P(w_n \mid w_1, w_2, w_3, \ldots, w_{n-1})$$

$$= \frac{|w_1|}{\sum_{k=1}^{m} |w_k|} \cdot \frac{|w_1, w_2|}{|w_1|} \cdot \frac{|w_1, w_2, w_3|}{|w_1, w_2|} \cdot \frac{|w_2, w_3, w_4|}{|w_2, w_3|} \cdot \ldots \cdot \frac{|w_{n-2}, w_{n-1}, w_n|}{|w_{n-2}, w_{n-1}|}$$

Beispiel (1)

P(Herr, Schuhmann, isst, ein, Brot)

= P(Herr) * P(Schuhmann | Herr) * P(isst | Herr,Schuhmann) * P(ein | Schuhmann,isst) * P(Brot | isst,ein)

= |Herr|/|“alle Wortformen“| * |Herr,Schuhmann|/|Herr| * |Herr,Schuhmann,isst|/|Herr,Schuhmann| * |Schuhmann,isst,ein|/|Schuhmann,isst| * |isst,ein,Brot|/|isst,ein|

= |Herr,Schuhmann,isst|*|Schuhmann,isst,ein|*|isst,ein,Brot| / |“alle Wortformen“|*|Schuhmann,isst|*|isst,ein|

=