Einführung in die Computerlinguistik und Sprachtechnologie

Vorlesung im WiSe 2018/19
(B-GSW-12)

Prof. Dr. Udo Hahn

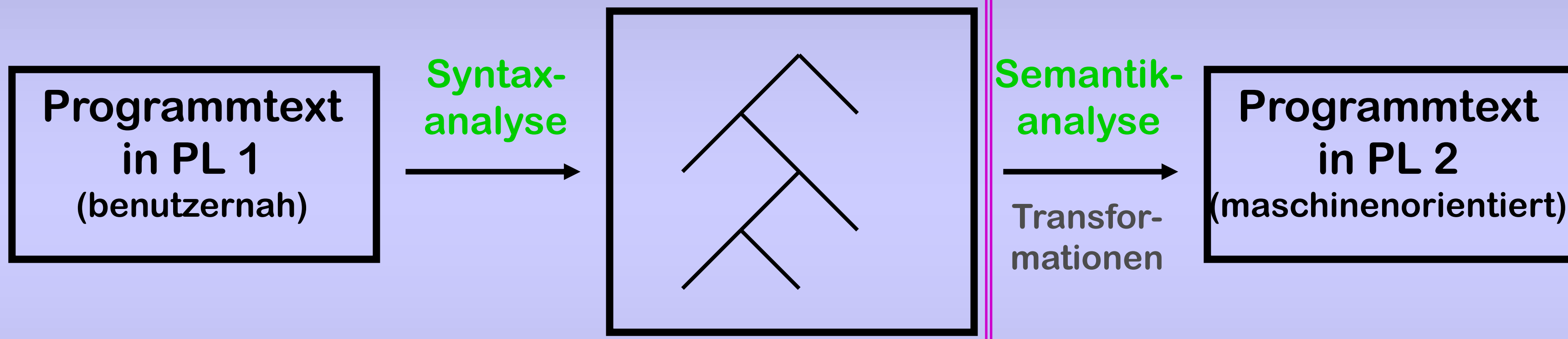
Lehrstuhl für Computerlinguistik
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

http://www.julielab.de

Syntaxanalyse

- Formale Analyse von Ausdrücken einer Sprache
  - Computerlinguistik
    - Formale Analyse von Wörtern oder Sätzen einer natürlichen Sprache (z.B. des Deutschen)
  - Informatik
    - Formale Analyse von Ausdrücken einer formalen Sprache (z.B. einer Programmiersprache)

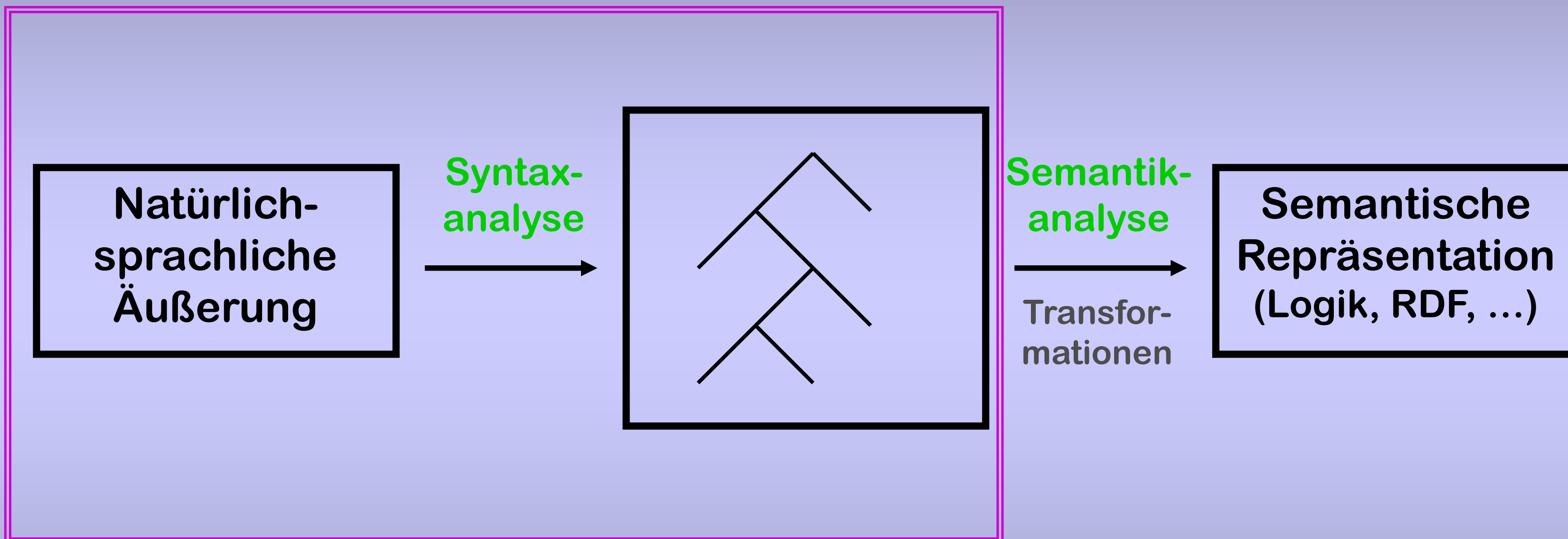
Analyse von Programmen



Aufgaben der Syntaxanalyse:

1. Syntaktisch korrekte Programme werden als korrekt erkannt
2. Syntaktisch unkorrekte Programme werden zurück gewiesen: Fehlererkennung und -diagnose

Analyse von natürlichsprachlichen Äußerungen



Aufgaben der Syntaxanalyse:

1. Syntaktisch korrekte Äußerungen werden als korrekt erkannt
2. Syntaktisch unkorrekte Äußerungen werden zurück gewiesen:

   Aber: Robustheit im Umgang mit paragrammatischen Äußerungen ist wünschenswert !

Beziehung zwischen Informatik und Computerlinguistik

- Informatik besitzt umfangreichen Methodenfundus
  - präzise beschriebene Analyseverfahren
  - Charakterisierung der formalen Eigenschaften dieser Verfahren (Entscheidbarkeit, Berechnungskomplexität)
  - mathematische Beschreibung der „Hintergrundtheorie“ (formale Grammatiken, formale Sprachen, Automaten)
- Übernahme und Adaption an NL in CL

Mengentheoretische Grundbegriffe

- **Die Zusammenfassung aller Elemente x, die eine Eigenschaft $\mathcal{E}$ haben, wird als Menge M bezeichnet:**

$$\mathbf{M} := \{x \mid x \text{ hat die Eigenschaft } \mathcal{E}\}$$

Beispiele:

LAUF := {x | x ist deutsches Lexem, das mit „LAUF" beginnt }

EoR := {x | x ist deutsches Lexem, das auf „E" oder „R" endet }

Mengentheoretische Grundbegriffe

- **Seien M_1 und M_2 Mengen. M_1 ist Teilmenge von M_2, falls aus $x \in M_1$ stets $x \in M_2$ folgt; symbolisch: $M_1 \subseteq M_2$.**
- **Gilt für zwei Mengen, M_1 und M_2 , einerseits $M_1 \subseteq M_2$ und andererseits $M_1 \neq M_2$, dann ist M_1 echte Teilmenge von M_2 ; symbolisch: $M_1 \subset M_2$**

Beispiele:

LAUF* := {Laufbahn, laufen, Lauffeuer, Laufmasche, Laufsteg} $\subseteq$ LAUF

LAUF $\subset$ LA := {x | x ist deutsches Lexem, das mit „LA“ beginnt }

R := {x | x ist deutsches Lexem, das auf „R“ endet } $\subseteq$ EoR

Mengentheoretische Grundbegriffe

- Gilt für zwei Mengen, M_1 und M_2 , sowohl $M_1 \subseteq M_2$ als auch $M_2 \subseteq M_1$, so folgt: $M_1 = M_2$ (Mengengleichheit).
- Die leere Menge ist die Menge, die kein Element enthält; symbolisch: {} oder Ø.
  - Bemerkung: Ø ist Teilmenge jeder Menge.
- Die Kardinalität einer endlichen Menge M ist die Anzahl ihrer Elemente; symbolisch: $|M|$

Mengentheoretische Grundbegriffe

- **Wenn M und N Mengen sind, dann charakterisiert die Menge**

M $\cap$ **N** **:=** $\{x \mid x \in$ **M und** $x \in$ **N** $\}$

den Durchschnitt

M $\cup$ **N** **:=** $\{x \mid x \in$ **M oder** $x \in$ **N** $\}$

die Vereinigung

von M und N

Mengentheoretische Grundbegriffe

- **Beispiele:**

  LAUF*:= {Laufbahn, laufen, Lauffeuer, Laufmasche, Laufsteg}

  LAUF* ∩ EoR

  = { Lauffeuer, Laufmasche }

  { Lauffeuer, Laufmasche } ∪ { Lauffeuer, Laufpass }

  = { Lauffeuer, Laufmasche, Laufpass }

Mengentheoretische Grundbegriffe

- Wenn I = $\{1,\ldots,n\}$ eine nichtleere Indexmenge ist und jedes $i \in$ I für M_i eine Menge charakterisiert, dann gilt als
  - Verallgemeinerung des Durchschnitts

$$\bigcap_{i \in I} M_i := \{x \mid x \in M_i \text{ für alle } i \in I\} = \bigcap_{i=1}^{n} M_i$$

  - Verallgemeinerung der Vereinigung

$$\bigcup_{i \in I} M_i := \{x \mid x \in M_i \text{ f.mind.ein } i \in I\} = \bigcup_{i=1}^{n} M_i$$

Mengentheoretische Grundbegriffe

- Das Kartesische Produkt von endlich vielen Mengen $M_1,..,M_n$, $n \geq 2$, ist die Menge aller n-tupel:

$$M_1 \times M_2 \times \ldots \times M_n := \{ (m_1,..,m_n) \mid m_i \in M_i ,\ 1 \leq i \leq n \}$$

Beispiel:

LAUFB := { Laufbahn, Laufbursche }

LAUFS := { Laufschritt, Laufstall, Laufsteg }

LAUFB x LAUFS = { (Laufbahn, Laufschritt), (Laufbahn, Laufstall), (Laufbahn, Laufsteg), (Laufbursche, Laufschritt), (Laufbursche, Laufstall), (Laufbursche, Laufsteg) }

Grundbegriffe zu Relationen

- **Eine (zweistellige) Relation ρ zwischen zwei Mengen M_1 und M_2 ist eine Teilmenge von $M_1 \times M_2$, d.h. $\rho \subseteq M_1 \times M_2$. Man schreibt auch $m \,\rho\, n$ für $(m,n) \in \rho$.**

Beispiel:

GleicheLänge $\subseteq$ DLexeme x DLexeme

GleicheLänge = { (du, da), (da, Ei), (er, es), (Dom, Bor), (Aal, Tor), (Bild, Tier), (Tiger, Sekte),... }

Grundbegriffe zu Relationen

- Das Produkt zweier Relationen, ρ und σ auf M, ist festgelegt durch

  $\rho\,\sigma := \{ (x,z) \mid (x,y) \in \rho \text{ und } (y,z) \in \sigma \text{ f.e. } y \in M \}$

Grundbegriffe zu Relationen

- Für eine beliebige Relation ρ auf M definiert
  - $\rho^0 := \{ (m,m) \mid m \in M \}$ die Diagonale,
  - $\rho^1 := \rho$ und $\rho^i := \rho^{i-1} \rho$ für $i>1$
  - $\rho^+ := \bigcup_{i \geq 1} \rho^i = \rho^1 \cup \rho^2 \cup \ldots \cup \rho^n$ die transitive Hülle von ρ,
  - $\rho^* := \bigcup_{i \geq 0} \rho^i = \rho^0 \cup \rho^1 \cup \rho^2 \cup \ldots \cup \rho^n$ die reflexive und transitive Hülle von ρ

Transitive Hülle von *„Ist_Unterbegriff“*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

Transitive Hülle von *„Ist_Unterbegriff“*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$\text{Ist\_Unterbegriff}^1 = \text{Ist\_Unterbegriff}$

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

Transitive Hülle von *„Ist_Unterbegriff"*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$\text{Ist\_Unterbegriff}^1 = \text{Ist\_Unterbegriff}$

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$\text{Ist\_Unterbegriff}^2 = \text{Ist\_Unterbegriff}^1 \ \text{Ist\_Unterbegriff}$

= { (VW-Golf, PKW) }

Transitive Hülle von *„Ist_Unterbegriff"*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$\text{Ist\_Unterbegriff}^1 = \text{Ist\_Unterbegriff}$

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$\text{Ist\_Unterbegriff}^2 = \text{Ist\_Unterbegriff}^1 \ \text{Ist\_Unterbegriff}$

= { (VW-Golf, PKW), (VW-PKW, KFZ) }

Transitive Hülle von „*Ist_Unterbegriff*“

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

Ist_Unterbegriff^1 = Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

Ist_Unterbegriff^2 = Ist_Unterbegriff^1 Ist_Unterbegriff

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt) }

Transitive Hülle von *„Ist_Unterbegriff"*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$\text{Ist\_Unterbegriff}^1 = \text{Ist\_Unterbegriff}$

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$\text{Ist\_Unterbegriff}^2 = \text{Ist\_Unterbegriff}^1 \text{ Ist\_Unterbegriff}$

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt) }

Transitive Hülle von „*Ist_Unterbegriff*"

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (<u>Möbel</u>, Artefakt) }

Ist_Unterbegriff^1 = Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, <u>Möbel</u>), (Möbel, Artefakt) }

Ist_Unterbegriff^2 = Ist_Unterbegriff^1 Ist_Unterbegriff

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt) }

Transitive Hülle von „*Ist_Unterbegriff*“

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

Ist_Unterbegriff^1 = Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

Ist_Unterbegriff^2 = Ist_Unterbegriff^1 Ist_Unterbegriff

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt), (Möbel, Objekt) }

Transitive Hülle von *„Ist_Unterbegriff“*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

Ist_Unterbegriff^2 = Ist_Unterbegriff^1 Ist_Unterbegriff

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt), (Möbel, Objekt) }

Ist_Unterbegriff^3 = Ist_Unterbegriff^2 Ist_Unterbegriff

= { (VW-Golf, KFZ) }

Transitive Hülle von *„Ist_Unterbegriff“*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

Ist_Unterbegriff^2 = Ist_Unterbegriff^1 Ist_Unterbegriff

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt), (Möbel, Objekt) }

Ist_Unterbegriff^3 = Ist_Unterbegriff^2 Ist_Unterbegriff

= { (VW-Golf, KFZ), (VW-PKW, Artefakt) }

Transitive Hülle von „*Ist_Unterbegriff*“

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

Ist_Unterbegriff^2 = Ist_Unterbegriff^1 Ist_Unterbegriff

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt), (Möbel, Objekt) }

Ist_Unterbegriff^3 = Ist_Unterbegriff^2 Ist_Unterbegriff

= { (VW-Golf, KFZ), (VW-PKW, Artefakt), (PKW, Objekt) }

Transitive Hülle von *„Ist_Unterbegriff“*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

Ist_Unterbegriff^2 = Ist_Unterbegriff^1 Ist_Unterbegriff

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt), (Möbel, Objekt) }

Ist_Unterbegriff^3 = Ist_Unterbegriff^2 Ist_Unterbegriff

= { (VW-Golf, KFZ), (VW-PKW, Artefakt), (PKW, Objekt), (Schreibtisch, Objekt) }

Transitive Hülle von *„Ist_Unterbegriff“*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$\text{Ist\_Unterbegriff}^2 = \text{Ist\_Unterbegriff}^1\ \text{Ist\_Unterbegriff}$

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt), (Möbel, Objekt) }

$\text{Ist\_Unterbegriff}^3 = \text{Ist\_Unterbegriff}^2\ \text{Ist\_Unterbegriff}$

= { (VW-Golf, KFZ), (VW-PKW, Artefakt), (PKW, Objekt), (Schreibtisch, Objekt) }

$\text{Ist\_Unterbegriff}^4 = \text{Ist\_Unterbegriff}^3\ \text{Ist\_Unterbegriff}$

= { (VW-Golf, Artefakt) }

Transitive Hülle von *„Ist_Unterbegriff“*

Ist_Unterbegriff
= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$\text{Ist\_Unterbegriff}^2 = \text{Ist\_Unterbegriff}^1\ \text{Ist\_Unterbegriff}$
= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt), (Möbel, Objekt) }

$\text{Ist\_Unterbegriff}^3 = \text{Ist\_Unterbegriff}^2\ \text{Ist\_Unterbegriff}$
= { (VW-Golf, KFZ), (VW-PKW, Artefakt), (PKW, Objekt), (Schreibtisch, Objekt) }

$\text{Ist\_Unterbegriff}^4 = \text{Ist\_Unterbegriff}^3\ \text{Ist\_Unterbegriff}$
= { (VW-Golf, Artefakt), (VW-PKW, Objekt) }

Transitive Hülle von *„Ist_Unterbegriff“*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$Ist\_Unterbegriff^2 = Ist\_Unterbegriff^1 \; Ist\_Unterbegriff$

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt), (Möbel, Objekt) }

$Ist\_Unterbegriff^3 = Ist\_Unterbegriff^2 \; Ist\_Unterbegriff$

= { (VW-Golf, KFZ), (VW-PKW, Artefakt), (PKW, Objekt), (Schreibtisch, Objekt) }

$Ist\_Unterbegriff^4 = Ist\_Unterbegriff^3 \; Ist\_Unterbegriff$

= { (VW-Golf, Artefakt), (VW-PKW, Objekt) }

$Ist\_Unterbegriff^5 = Ist\_Unterbegriff^4 \; Ist\_Unterbegriff$

= { (VW-Golf, Objekt) }

Transitive Hülle von *„Ist_Unterbegriff“*

Ist_Unterbegriff

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt) }

$\text{Ist\_Unterbegriff}^2 = \text{Ist\_Unterbegriff}^1\ \text{Ist\_Unterbegriff}$

= { (VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt), (Möbel, Objekt) }

$\text{Ist\_Unterbegriff}^3 = \text{Ist\_Unterbegriff}^2\ \text{Ist\_Unterbegriff}$

= { (VW-Golf, KFZ), (VW-PKW, Artefakt), (PKW, Objekt), (Schreibtisch, Objekt) }

$\text{Ist\_Unterbegriff}^4 = \text{Ist\_Unterbegriff}^3\ \text{Ist\_Unterbegriff}$

= { (VW-Golf, Artefakt), (VW-PKW, Objekt) }

$\text{Ist\_Unterbegriff}^5 = \text{Ist\_Unterbegriff}^4\ \text{Ist\_Unterbegriff}$

= { (VW-Golf, Objekt) }

$\text{Ist\_Unterbegriff}^6 = \text{Ist\_Unterbegriff}^5\ \text{Ist\_Unterbegriff} = \{\}$

Transitive Hülle von „*Ist_Unterbegriff*“

$$\bigcup_{i \geq 1} \text{Ist\_Unterbegriff}^i = \text{Ist\_Unterbegriff}^1 \cup \text{Ist\_Unterbegriff}^2 \cup \ldots \cup \text{Ist\_Unterbegriff}^n$$

= { (VW-Golf, VW-PKW), (VW-PKW, PKW), (PKW, KFZ), (KFZ, Artefakt), (Artefakt, Objekt), (Schreibtisch, Möbel), (Möbel, Artefakt),

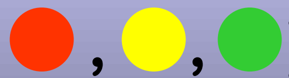
(VW-Golf, PKW), (VW-PKW, KFZ), (PKW, Artefakt), (KFZ, Objekt), (Schreibtisch, Artefakt), (Möbel, Objekt),

(VW-Golf, KFZ), (VW-PKW, Artefakt), (PKW, Objekt), (Schreibtisch, Objekt),

(VW-Golf, Artefakt), (VW-PKW, Objekt),

(VW-Golf, Objekt) }

Grundlagen formaler Sprachen:
Alphabet

- Sei Σ ein beliebiges Alphabet, d.i. eine Menge von Symbolen oder Zeichen
  - Beispiele für verbreitete Alphabete:
    - {A,B,C, ..., X,Y,Z} — lateinisches Alphabet
    - {1,2,3, ..., 7,8,9, 0} — indisch-arabisches Zahlensystem
    - {0,1} — Binärzahlen
    - {  } — internat. Ampelalphabet
    - {A[denin], G[uanin], T[hymin], C[ytosin]} — Basen-Alphabet der DNA

Grundlagen formaler Sprachen:
Wörter

- Seien Wörter (Sätze, Strings, Ketten) über einem Alphabet Σ in der folgenden Weise definiert:

1. ε ist ein Wort über Σ (ε ist das Leerwort, das keine Symbole hat)
2. falls χ ein Wort über Σ und $\alpha \in \Sigma$ ist, dann ist $\chi\ \alpha$ ein Wort über Σ
3. γ ist ein Wort über Σ genau dann, wenn sein Bildung aus (1) oder (2) folgt

Grundlagen formaler Sprachen:
Konkatenation von Wörtern

- **Das Wort $\omega \circ \tau := \omega\tau := \omega_1 \ldots \omega_m \tau_1 \ldots \tau_n$ heißt Konkatenation von ω und τ, falls $\omega = \omega_1 \ldots \omega_m$ und $\tau = \tau_1 \ldots \tau_n$ ($\omega_i, \tau_j \in \Sigma$) Wörter über Σ sind;**
  „o“ (sprich: „Kringel“) ist der Konkatenationsoperator.
  - **Für alle Wörter ω gilt: $\omega \circ \varepsilon = \varepsilon \circ \omega = \omega$**

Beispiele für Konkatenationen:

- ABC o X = ABCX
- 24 o 24 = 2424
- wort o stamm = wortstamm

Grundlagen formaler Sprachen:
Wortmengen

- **Eine Wortmenge (Sprache) $\mathcal{F}$ bezüglich eines Alphabets Σ ist gegeben durch**

$$\mathcal{F} := \{\ \omega \mid \omega \text{ ist Wort über } \Sigma\ \}$$

Grundlagen formaler Sprachen:
Konkatenation von Wortmengen

- Die **Zusammensetzung** (**Konkatenation**) von **Wortmengen** $\mathcal{F}$ und $\mathcal{M}$ ist gegeben durch

$$\mathcal{F} \circ \mathcal{M} := \mathcal{F}\,\mathcal{M} := \{\, \omega\tau \mid \omega \in \mathcal{F}, \tau \in \mathcal{M} \,\}$$

- Dabei gilt:

$$\mathcal{F}\,\{\varepsilon\} = \{\varepsilon\}\,\mathcal{F} = \mathcal{F}$$

$$\mathcal{F}\,\emptyset = \emptyset\,\mathcal{F} = \emptyset$$

Beispiele für Konkatenationen:

- DLex := {Fisch, Fest, Fleck}; DEnd := {e, es, er}
- DLex $\circ$ DEnd = {Fische, Fisches, Fischer, Feste, Festes, Fester, Flecke, Fleckes, Flecker}

Grundlagen formaler Sprachen:

Potenzen von Wörtern bzw. Wortmengen

- **Potenzen** von **Wörtern** ω bzw. **Wort-mengen** $\mathcal{F}$ sind gegeben durch:
  - $\omega^0 := \varepsilon$ $\quad \omega^1 := \omega$ $\quad \omega^i := \omega^{i-1}\,\omega$, für $i \geq 1$
  - $\mathcal{F}^0 := \{\varepsilon\}$ $\quad \mathcal{F}^1 := \mathcal{F}$ $\quad \mathcal{F}^i := \mathcal{F}^{i-1}\,\mathcal{F}$, für $i \geq 1$

Beispiele für Potenzen von Wortmengen:

- DSilbe := {ba, di, ko}
- $\text{DSilbe}^0 = \{\varepsilon\}$, $\text{DSilbe}^1 = \{\text{ba}, \text{di}, \text{ko}\}$
- $\text{DSilbe}^2 = \text{DSilbe}^1\ \text{DSilbe}$

  = { baba, badi, bako, diba, didi, diko , koba, kodi, koko}
- $\text{DSilbe}^3 = \text{DSilbe}^2\ \text{DSilbe}$

Grundlagen formaler Sprachen:
Plus- und Sternhülle – formale Sprache

- Die **Plushülle** bzw. **Sternhülle** einer Wortmenge $\mathcal{F}$ werden definiert durch:

$$\mathcal{F}^{+} := \bigcup_{i \geq 1} \mathcal{F}^{i} \qquad \mathcal{F}^{*} := \bigcup_{i \geq 0} \mathcal{F}^{i}$$

- Ist Σ ein Alphabet, dann ist Σ^* die Gesamtheit aller Wörter über Σ. Jede Teilmenge dieser Sternhülle, $\mathcal{L} \subseteq \Sigma^*$, heißt **formale Sprache** über Σ .

Grundbegriffe zu formalen Grammatiken

- Eine formale Grammatik $\mathcal{G}$ ist ein 4-tupel

$$\mathcal{G} = (N, T, P, S)$$

mit

  - N: das Alphabet der Nicht-Terminalsymbole
  - T: das Alphabet der Terminalsymbole
  - P: eine endliche Menge von Produktionen der Form

$\alpha \rightarrow \gamma$ (gesprochen: „*α produziert γ*“) mit

$\alpha \in (N \cup T)^* \, N \, (N \cup T)^*$ und

$\gamma \in (N \cup T)^*$

  - S: das Startsymbol, $S \in N$

$\mathcal{V} = N \cup T$, bezeichnet das Gesamtalphabet ($N \cap T = \emptyset$)

Beziehung zwischen formalen Grammatiken & formalen Sprachen

- **Eine formale Grammatik $\mathcal{G}$ erzeugt eine formale Sprache. Der Erzeugungsprozess ist festgelegt durch eine auf $\mathcal{V}^*$ definierte Relation „$\Rightarrow$“ (gesprochen: *„ist direkt ableitbar nach“*). Für u, v, $\gamma \in \mathcal{V}^*$ und $\alpha \in \mathcal{V}^*$ N $\mathcal{V}^*$ gilt:**

u α v $\Rightarrow$ u γ v genau dann, wenn $\alpha \rightarrow \gamma \in$ P

Grundbegriffe zu formalen Grammatiken

- Die **transitive Hülle** der Relation „$\Rightarrow$“ schreibt man

  $\overset{+}{\Rightarrow}$ (gesprochen: „ist nichttrivial ableitbar nach“)

- Die **reflexive und transitive Hülle** der Relation „$\Rightarrow$“ schreibt man

  $\overset{*}{\Rightarrow}$ (gesprochen: „ist ableitbar nach“)

Grundbegriffe zu formalen Grammatiken

- Man schreibt $s \overset{n}{\Rightarrow} z$, um auszudrücken, dass Ketten $s_0, s_1, \ldots, s_n$ existieren mit

  $s = s_0$, $s_i \Rightarrow s_{i+1}$ für $0 \leq i < n$ und $s_n = z$

- Damit ist also

  $s \overset{+}{\Rightarrow} z$ g.d.w. $s \overset{n}{\Rightarrow} z$ für ein $n \geq 1$ und

  $s \overset{*}{\Rightarrow} z$ g.d.w. entweder $s = z$ oder $s \overset{+}{\Rightarrow} z$

Grundbegriffe zu formalen Grammatiken

- Die von der formalen Grammatik $\mathcal{G}$ = (N, T, P, S) erzeugte formale Sprache $\mathcal{L}(\mathcal{G})$ ist wie folgt definiert:

  $$\mathcal{L}(\mathcal{G}) := \{ \tau \mid \tau \in T^*, S \overset{*}{\Rightarrow} \tau, \text{ S ist Startsymbol von } \mathcal{G} \}$$

  τ heißt auch Wort (oder Satz) der Sprache $\mathcal{L}(\mathcal{G})$.

Grundbegriffe zu formalen Grammatiken

- **Abhängig von der Form der zugelassenen Produktionen definiert man vier Typen von formalen Grammatiken:**
  - **Eine Grammatik $\mathcal{G}$ heißt Typ-0-Grammatik, wenn die Gestalt der Produktionen nicht weiter eingeschränkt ist. D.h., sie haben die Form**

$$\alpha \longrightarrow \gamma$$

mit

$$\alpha \in (N \cup T)^* \, N \, (N \cup T)^* \text{ und}$$

$$\gamma \in (N \cup T)^*$$

Grundbegriffe zu formalen Grammatiken

– Eine Grammatik $\mathcal{G}$ heißt Typ-1-Grammatik (kontextsensitive Grammatik), wenn P nur Produktionen der Gestalt

$$\alpha \rightarrow \gamma$$

mit

$$\alpha \in (N \cup T)^* \, N \, (N \cup T)^* \text{ und}$$

$$\gamma \in (N \cup T)^*$$

$$|\alpha| \leq |\gamma|$$

(sog. *non-shrinking rules*) und eventuell die Produktion $S \rightarrow \varepsilon$ enthält (wobei letztere nur zugelassen ist, wenn das Startsymbol S in keiner Produktion auf der rechten Seite auftritt).

Grundbegriffe zu formalen Grammatiken

– **Eine Grammatik $\mathcal{G}$ heißt Typ-2-Grammatik (kontextfreie Grammatik), wenn P nur Produktionen enthält der Gestalt**

$$A \rightarrow \gamma \quad \text{mit } A \in N \text{ und } \gamma \in (N \cup T)^*$$

Grundbegriffe zu formalen Grammatiken

- Eine Grammatik $\mathcal{G}$ heißt Typ-3-Grammatik (reguläre Grammatik), wenn P nur Produktionen der Gestalt

$$A \longrightarrow \gamma \quad \text{mit } A \in \mathbf{N} \text{ und } \gamma \in \mathbf{N}\,\mathbf{T}^* \cup \mathbf{T}^*$$

(sog. linkslineare Produktionen) oder nur Produktionen der Gestalt

$$A \longrightarrow \gamma \quad \text{mit } A \in \mathbf{N} \text{ und } \gamma \in \mathbf{T}^*\,\mathbf{N} \cup \mathbf{T}^*$$

(sog. rechtslineare Produktionen) enthält.
- Man spricht dann auch entsprechend von linkslinearen bzw. rechtslinearen Grammatiken.
- Eine reguläre Grammatik darf nicht Regeln nach beiden Produktionsregelmustern mischen.

Beispiel einer rechtslinearen Grammatik

$\mathcal{G}$-3 = (N, T, P, S) mit

N = { S, A, B }

T = { a, b }

P = { $S \longrightarrow aA$,

$A \longrightarrow aA$,

$A \longrightarrow bbB$,

$B \longrightarrow bB$,

$B \longrightarrow b$ }

$S \Rightarrow aA$	mit	$S \rightarrow aA \in P$
$aA \Rightarrow abbB$	mit	$A \rightarrow bbB \in P$
$abbB \Rightarrow abbb$	mit	$B \rightarrow b \in P$

$\mathcal{L}(\mathcal{G}$-3) = { abbb, aabbb, aaabbb, aaaabbb, abbbb, ... }

= $a^n b^m$, $n \geq 1$, $m \geq 3$

Beispiel einer kontextfreien Grammatik

$\mathcal{G}$-2 = (N, T, P, S) mit

N = { S }

T = { a, b }

P = { S $\longrightarrow$ aSb,

S $\longrightarrow$ ab }

$\mathcal{L}(\mathcal{G}$-2) = {ab, aabb, aaabbb, aaaabbbb, … }

= $a^n b^n$, $n \geq 1$

Beispiel einer kontextfreien Grammatik

$\mathcal{G}$-2 = (N, T, P, S) mit

N = { S }

T = { a, b }

P = { S $\longrightarrow$ aSb,

S $\longrightarrow$ ab }

$\mathcal{L}$($\mathcal{G}$-2) = {ab, aabb, aaabbb, aaaabbbb, ... }

= $a^n b^n$, $n \geq 1$

S $\Rightarrow$ aSb mit S $\rightarrow$ aSb $\in$ P

aSb $\Rightarrow$ aabb mit S $\rightarrow$ ab $\in$ P

Beispiel einer kontextfreien Grammatik

$\mathcal{G}$-2 = (N, T, P, S) mit

N = { S }

T = { a, b }

P = { $S \longrightarrow aSb$,

$S \longrightarrow ab$ }

$\mathcal{L}(\mathcal{G}\text{-}2)$ = {ab, aabb, aaabbb, aaaabbbb, … }

= $a^n b^n$, $n \geq 1$

$S \stackrel{4}{\Rightarrow}$ aaaabbbb

$S \stackrel{*}{\Rightarrow}$ aaaabbbb

$\underset{s_0}{\boxed{S}} \Rightarrow \underset{s_1}{\boxed{aSb}}$

Beispiel einer kontextfreien Grammatik

$\mathcal{G}$-2 = (N, T, P, S) mit

N = { S }

T = { a, b }

P = { $S \longrightarrow aSb$,

$S \longrightarrow ab$ }

$\mathcal{L}(\mathcal{G}\text{-}2)$ = {ab, aabb, aaabbb, aaaabbbb, ... }

= $a^n b^n$, $n \geq 1$

$S \overset{4}{\Rightarrow} aaaabbbb$

$S \overset{*}{\Rightarrow} aaaabbbb$

$\underset{s_0}{\boxed{S}} \Rightarrow \underset{s_1}{\boxed{aSb}} \Rightarrow \underset{s_2}{\boxed{aaSbb}}$

Beispiel einer kontextfreien Grammatik

$\mathcal{G}$-2 = (N, T, P, S) mit

N = { S }

T = { a, b }

P = { $S \longrightarrow aSb$,

$S \longrightarrow ab$ }

$\mathcal{L}(\mathcal{G}\text{-}2)$ = {ab, aabb, aaabbb, aaaabbbb, ... }

= $a^n b^n$, $n \geq 1$

$S \stackrel{4}{\Rightarrow} aaaabbbb$ $S \stackrel{*}{\Rightarrow} aaaabbbb$

$\underbrace{\boxed{S}}_{s_0} \Rightarrow \underbrace{\boxed{aSb}}_{s_1} \Rightarrow \underbrace{\boxed{aaSbb}}_{s_2} \Rightarrow \underbrace{\boxed{aaaSbbb}}_{s_3}$

Beispiel einer kontextfreien Grammatik

$\mathcal{G}$-2 = (N, T, P, S) mit

N = { S }

T = { a, b }

P = { S $\longrightarrow$ aSb,

S $\longrightarrow$ ab }

$\mathcal{L}(\mathcal{G}\text{-}2)$ = {ab, aabb, aaabbb, aaaabbbb, … }

= $a^n b^n$, $n \geq 1$

$S \overset{4}{\Rightarrow}$ aaaabbbb $\qquad$ $S \overset{*}{\Rightarrow}$ aaaabbbb

$\boxed{S} \Rightarrow \boxed{aSb} \Rightarrow \boxed{aaSbb} \Rightarrow \boxed{aaaSbbb} \Rightarrow \boxed{aaaabbbb}$

s_0 $\quad$ s_1 $\quad$ s_2 $\quad$ s_3 $\quad$ s_4

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { $S \longrightarrow aSBC$, $S \longrightarrow aBC$,

$CB \longrightarrow XB$, $XB \longrightarrow XC$, $XC \longrightarrow BC$,

$aB \longrightarrow ab$,

$bB \longrightarrow bb$,

$C \longrightarrow c$ }

$\mathcal{L}(\mathcal{G}$-1) = { abc, aabbcc, aaabbbccc, ... }

= $a^n b^n c^n$, $n \geq 1$

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { $S \longrightarrow aSBC$, $S \longrightarrow aBC$,

$CB \longrightarrow XB$, $XB \longrightarrow XC$, $XC \longrightarrow BC$,

$aB \longrightarrow ab$, $bB \longrightarrow bb$, $C \longrightarrow c$ }

$S \stackrel{*}{\Longrightarrow} aaabbbccc$

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { $S \longrightarrow aSBC$, $S \longrightarrow aBC$,

$CB \longrightarrow XB$, $XB \longrightarrow XC$, $XC \longrightarrow BC$,

$aB \longrightarrow ab$, $bB \longrightarrow bb$, $C \longrightarrow c$ }

$S \Rightarrow aSBC$

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { $S \longrightarrow aSBC$, $S \longrightarrow aBC$,

$CB \longrightarrow XB$, $XB \longrightarrow XC$, $XC \longrightarrow BC$,

$aB \longrightarrow ab$, $bB \longrightarrow bb$, $C \longrightarrow c$ }

$S \Rightarrow aSBC \Rightarrow aaSBCBC$

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { S $\longrightarrow$ aSBC, S $\longrightarrow$ aBC,

CB $\longrightarrow$ XB, XB $\longrightarrow$ XC, XC $\longrightarrow$ BC,

aB $\longrightarrow$ ab, bB $\longrightarrow$ bb, C $\longrightarrow$ c }

S $\Rightarrow$ aSBC $\Rightarrow$ aaSBCBC $\Rightarrow$ aaaBCBCBC

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { $S \longrightarrow aSBC$, $S \longrightarrow aBC$,

$CB \longrightarrow XB$, $XB \longrightarrow XC$, $XC \longrightarrow BC$,

$aB \longrightarrow ab$, $bB \longrightarrow bb$, $C \longrightarrow c$ }

$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aaaBCBCBC \Rightarrow$
$\ldots \Rightarrow aaaBBCCBC$

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { $S \longrightarrow aSBC$, $S \longrightarrow aBC$,

$CB \longrightarrow XB$, $XB \longrightarrow XC$, $XC \longrightarrow BC$,

$aB \longrightarrow ab$, $bB \longrightarrow bb$, $C \longrightarrow c$ }

$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aaaBCBCBC \Rightarrow \ldots \Rightarrow aaaBBCCBC \Rightarrow \ldots \Rightarrow aaaBBCBCC$

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { $S \longrightarrow aSBC$, $S \longrightarrow aBC$,

$CB \longrightarrow XB$, $XB \longrightarrow XC$, $XC \longrightarrow BC$,

$aB \longrightarrow ab$, $bB \longrightarrow bb$, $C \longrightarrow c$ }

$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aaaBCBCBC \Rightarrow$
$\ldots \Rightarrow aaaBBCCBC \Rightarrow \ldots \Rightarrow aaaBBCBCC \Rightarrow$
$\ldots \Rightarrow aaaBBBCCC$

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { $S \longrightarrow aSBC$, $S \longrightarrow aBC$,

$CB \longrightarrow XB$, $XB \longrightarrow XC$, $XC \longrightarrow BC$,

$aB \longrightarrow ab$, $bB \longrightarrow bb$, $C \longrightarrow c$ }

$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aaaBCBCBC \Rightarrow$
$\ldots \Rightarrow aaaBBCCBC \Rightarrow \ldots \Rightarrow aaaBBCCBC \Rightarrow$
$\ldots \Rightarrow aaaBBBCCC \Rightarrow aaabBBCCC$

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { S $\longrightarrow$ aSBC, S $\longrightarrow$ aBC,

CB $\longrightarrow$ XB, XB $\longrightarrow$ XC, XC $\longrightarrow$ BC,

aB $\longrightarrow$ ab, bB $\longrightarrow$ bb, C $\longrightarrow$ c }

S $\Rightarrow$ aSBC $\Rightarrow$ aaSBCBC $\Rightarrow$ aaaBCBCBC $\Rightarrow$ … $\Rightarrow$ aaaBBCCBC $\Rightarrow$ … $\Rightarrow$ aaaBBCCBC $\Rightarrow$ … $\Rightarrow$ aaaBBBCCC $\Rightarrow$ aaabBBCCC $\Rightarrow$ aaabbBCCC

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { S $\longrightarrow$ aSBC, S $\longrightarrow$ aBC,

CB $\longrightarrow$ XB, XB $\longrightarrow$ XC, XC $\longrightarrow$ BC,

aB $\longrightarrow$ ab, bB $\longrightarrow$ bb, C $\longrightarrow$ c }

S $\Rightarrow$ aSBC $\Rightarrow$ aaSBCBC $\Rightarrow$ aaaBCBCBC $\Rightarrow$
… $\Rightarrow$ aaaBBCCBC $\Rightarrow$ … $\Rightarrow$ aaaBBCCBC $\Rightarrow$
… $\Rightarrow$ aaaBBBCCC $\Rightarrow$ aaabBBCCC $\Rightarrow$
aaabbBCCC $\Rightarrow$ aaabbbCCC

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { $S \rightarrow aSBC$, $S \rightarrow aBC$,

$CB \rightarrow XB$, $XB \rightarrow XC$, $XC \rightarrow BC$,

$aB \rightarrow ab$, $bB \rightarrow bb$, $C \rightarrow c$ }

$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aaaBCBCBC \Rightarrow$
$\ldots \Rightarrow aaaBBCCBC \Rightarrow \ldots \Rightarrow aaaBBCCBC \Rightarrow$
$\ldots \Rightarrow aaaBBBCCC \Rightarrow aaabBBCCC \Rightarrow$
$aaabbBCCC \Rightarrow aaabbbCCC \Rightarrow \ldots \Rightarrow aaabbbccc$

Beispiel einer kontextsensitiven Grammatik

$\mathcal{G}$-1 = (N, T, P, S) mit

N = { S, B, C, X }

T = { a, b, c }

P = { S $\longrightarrow$ aSBC, S $\longrightarrow$ aBC,

CB $\longrightarrow$ XB, XB $\longrightarrow$ XC, XC $\longrightarrow$ BC,

aB $\longrightarrow$ ab, bB $\longrightarrow$ bb, C $\longrightarrow$ c }

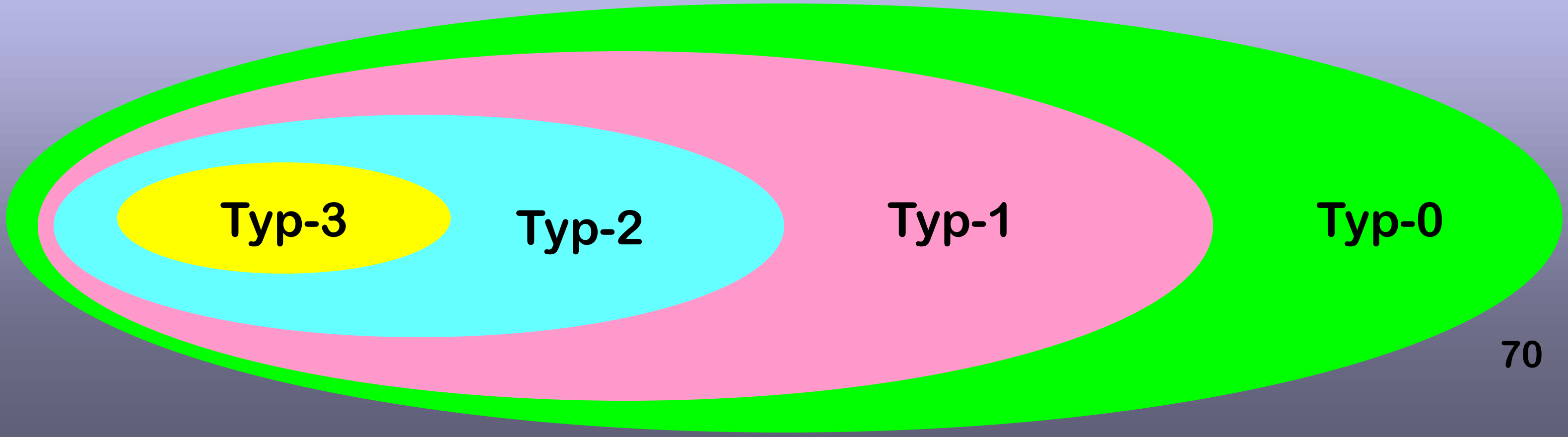
S $\Rightarrow$ aSBC $\Rightarrow$ aaSBCBC $\Rightarrow$ aaaBCBCBC $\Rightarrow$ … $\Rightarrow$ aaaBBCCBC $\Rightarrow$ … $\Rightarrow$ aaaBBCCBC $\Rightarrow$ … $\Rightarrow$ aaaBBBCCC $\Rightarrow$ aaabBBCCC $\Rightarrow$ aaabbBCCC $\Rightarrow$ aaabbbCCC $\Rightarrow$ … aaabbbccc

Typen formaler Sprachen (und ihr Bezug zu Grammatik-Typen)

- Eine formale Sprache heißt vom Typ 0, 1, 2 oder 3, wenn sie von einer Grammatik des entsprechenden Typs erzeugt werden kann.
  - Eine Typ-1-Sprache heißt auch kontextsensitive Sprache.
  - Eine Typ-2-Sprache heißt auch kontextfreie Sprache.
  - Eine Typ-3-Sprache heißt auch reguläre Sprache.

Chomsky-Hierarchie formaler Sprachen

– **Für jedes Alphabet Σ (mit mindestens zwei Symbolen) ist die Menge der Typ-i-Sprachen über Σ für i = 0, 1, 2 jeweils (echte) Obermenge der Typ-[i+1]-Sprachen über Σ.**
Die damit gegebene Hierarchie von formalen Sprachen heißt Chomsky-Hierarchie.



Natürliche Sprachen als formale Sprachen

- NLs sind keine Typ-3-Sprachen
- NLs sind überwiegend (Englisch, Deutsch, Französisch, Spanisch, ...) Typ-2-Sprachen
- Einige wenige NLs sind sicher (Schweizer Deutsch) bzw. vermutlich (Niederländisch, Bambara [Mali]) milde Typ-1-Sprachen