

Computerlinguistik I

Vorlesung im WiSe 2017/18
(M-GSW-09)

Prof. Dr. Udo Hahn

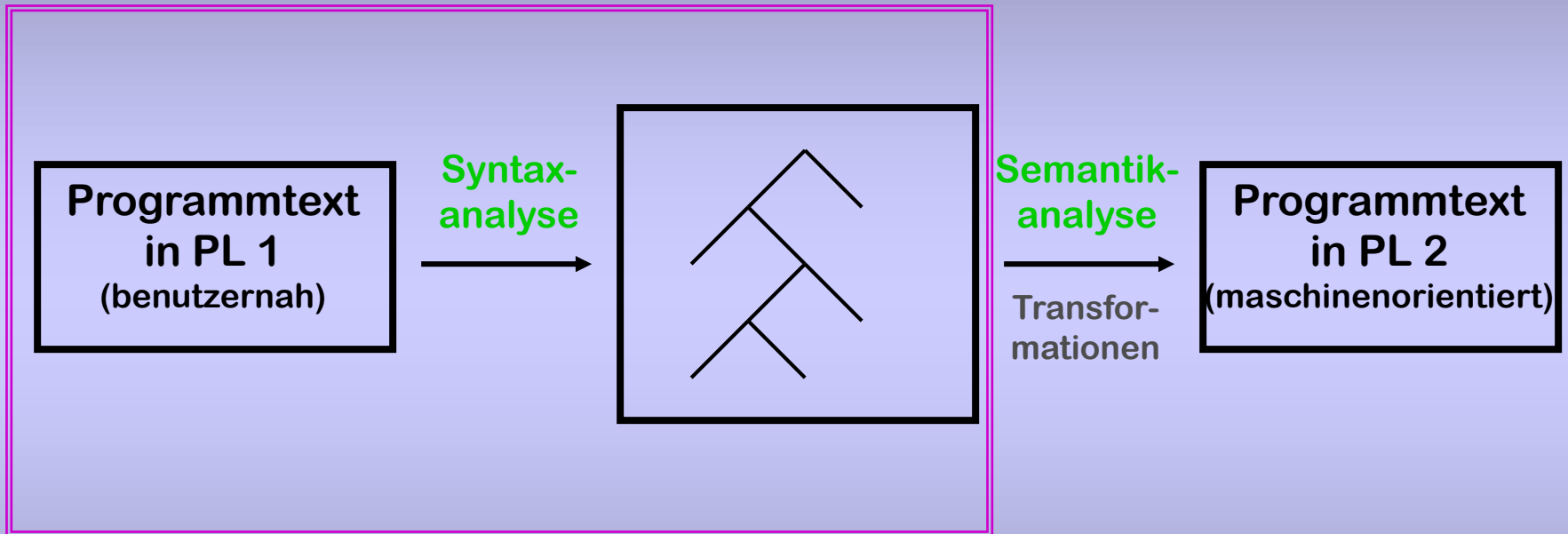
Lehrstuhl für Computerlinguistik
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

<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)

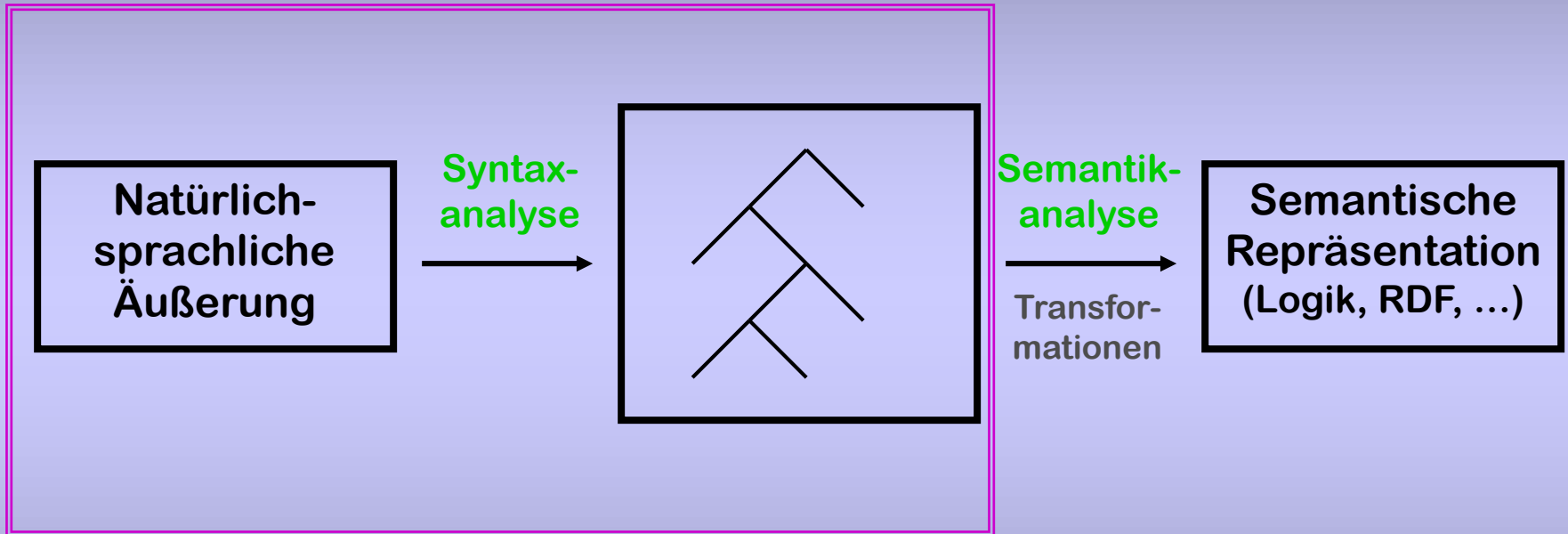
Analyse von Programmen



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

Grundbegriffe zu formalen Grammatiken

- Eine **formale Grammatik** G ist ein 4-tupel

$$G = (N, T, P, S)$$

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 G **erzeugt** eine formale Sprache. Der Erzeugungsprozess ist festgelegt durch eine auf V^* definierte Relation „ \Rightarrow “ (gesprochen: „*ist direkt ableitbar nach*“).
Für $u, v, \gamma \in V^*$ und $\alpha \in V^* N V^*$ gilt:
 $u \alpha v \Rightarrow u \gamma v$ genau dann, wenn $\alpha \rightarrow \gamma \in P$

Grundbegriffe zu formalen Grammatiken

- Die **transitive Hülle** der Relation „ \Rightarrow “ schreibt man
 \Rightarrow^+ (gesprochen: „ist nichttrivial ableitbar nach“)
- Die **reflexive und transitive Hülle** der Relation „ \Rightarrow “ schreibt man
 \Rightarrow^* (gesprochen: „ist ableitbar nach“)

Grundbegriffe zu formalen Grammatiken

- Man schreibt $s \stackrel{n}{\Rightarrow} z$, um auszudrücken, dass s_0, s_1, \dots, s_n existieren mit
$$s = s_0, s_i \Rightarrow s_{i+1} \text{ für } 0 \leq i < n \text{ und } s_n = z$$
- Damit ist also
$$s \stackrel{+}{\Rightarrow} z \text{ g.d.w. } s \stackrel{n}{\Rightarrow} z \text{ für ein } n \geq 1 \quad \text{und}$$
$$s \stackrel{*}{\Rightarrow} z \text{ g.d.w. entweder } s = z \text{ oder } s \stackrel{+}{\Rightarrow} z$$

Grundbegriffe zu formalen Grammatiken

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

$$\mathcal{L}(G) := \{ \tau \mid \tau \in T^*, S \stackrel{*}{\Rightarrow} \tau,$$

$S \text{ ist Startsymbol von } G \}$

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

Grundbegriffe zu formalen Grammatiken

- Zwei formale Grammatiken G_1 und G_2 , $G_1 \neq G_2$, heißen **äquivalent**, wenn sie dieselbe Sprache erzeugen, d.h.:

$$\mathcal{L}(G_1) = \mathcal{L}(G_2)$$

Grundbegriffe zu formalen Grammatiken

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

$$\alpha \rightarrow \gamma$$

mit

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

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

Grundbegriffe zu formalen Grammatiken

- Eine Grammatik 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)₃

Grundbegriffe zu formalen Grammatiken

- Eine Grammatik 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 G heißt **Typ-3-Grammatik** (**reguläre Grammatik**), wenn P nur Produktionen der Gestalt

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

(sog. **links**lineare Produktionen) oder nur Produktionen der Gestalt

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

(sog. **rechts**lineare Produktionen) enthält.

- Man spricht dann auch entsprechend von **linkslinaren** bzw. **rechtslinaren Grammatiken**.
- Eine reguläre Grammatik darf nicht Regeln nach beiden Produktionsregelmustern mischen.

Beispiel einer rechtslinearen Grammatik

$G-3 = (N, T, P, S)$ mit

$$N = \{ S, A, B \}$$

$$T = \{ a, b \}$$

$$P = \{ S \rightarrow aA,$$

$$A \rightarrow aA,$$

$$A \rightarrow bbB,$$

$$B \rightarrow bB,$$

$$B \rightarrow b \}$$

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

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

Beispiel einer rechtslinearen Grammatik

$G-3 = (N, T, P, S)$ mit

$N = \{ S, A, B \}$

$T = \{ a, b \}$

$P = \{ S \rightarrow aA,$

$A \rightarrow aA,$

$A \rightarrow bbB,$

$B \rightarrow bB,$

$B \rightarrow b \}$

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

$\mathcal{L}(G-3) = \{ abbb, aabbb, aaabbb, aaaabbb, abbbb, \dots \}$
 $= a^n b^m, n, m \geq 1$

Beispiel einer kontextfreien Grammatik

$G-2 = (N, T, P, S)$ mit

$$N = \{ S \}$$

$$T = \{ a, b \}$$

$$P = \{ S \rightarrow aSb, \\ S \rightarrow ab \}$$

$$\mathcal{L}(G-2) = \{ ab, aabb, aaabbb, aaaabbbb, \dots \} \\ = a^n b^n, n \geq 1$$

Beispiel einer kontextfreien Grammatik

$G-2 = (N, T, P, S)$ mit

$$N = \{ S \}$$

$$T = \{ a, b \}$$

$$P = \{ S \rightarrow aSb, \\ S \rightarrow ab \}$$

$$\mathcal{L}(G-2) = \{ ab, aabb, aaabbb, aaaabbbb, \dots \} \\ = a^n b^n, n \geq 1$$

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

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

Beispiel einer kontextfreien Grammatik

$G-2 = (N, T, P, S)$ mit

$$N = \{ S \}$$

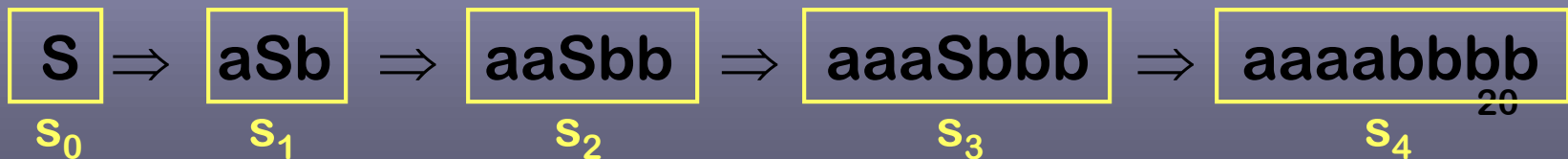
$$T = \{ a, b \}$$

$$P = \{ S \rightarrow aSb, \\ S \rightarrow ab \}$$

$$\mathcal{L}(G-2) = \{ ab, aabb, aaabbb, aaaabbbb, \dots \} \\ = a^n b^n, n \geq 1$$

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

$$S^* \Rightarrow aaaabbbb$$



Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC,$
 $CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC,$
 $aB \rightarrow ab,$
 $bB \rightarrow bb,$
 $C \rightarrow c \}$

$\mathcal{L}(G-1) = \{ abc, aabbcc, aaabbbccc, \dots \}$
 $= a^n b^n c^n, n \geq 1$

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC,$
 $CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC,$
 $aB \rightarrow ab, bB \rightarrow bb, C \rightarrow c \}$

$S^* \Rightarrow aaabbbccc$

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC,$
 $CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC,$
 $aB \rightarrow ab, bB \rightarrow bb, C \rightarrow c \}$

$S \Rightarrow aSBC$

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC,$
 $CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC,$
 $aB \rightarrow ab, bB \rightarrow bb, C \rightarrow c \}$

$S \Rightarrow aSBC \Rightarrow aaSBCBC$

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC, \\ CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC, \\ aB \rightarrow ab, bB \rightarrow bb, C \rightarrow c \}$

$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aaBCBCBC$

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC,$
 $CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC,$
 $aB \rightarrow ab, bB \rightarrow bb, C \rightarrow c \}$

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

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC,$
 $CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC,$
 $aB \rightarrow ab, bB \rightarrow bb, C \rightarrow c \}$

$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aaaBCBCBC \Rightarrow$
 $\dots \Rightarrow aaaBBCBC \Rightarrow \dots \Rightarrow aaaBBCBC$

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

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

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

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

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

$\dots \Rightarrow aaaBBCBC \Rightarrow \dots \Rightarrow aaaBBBCBC \Rightarrow$

$\dots \Rightarrow aaaBBBCCC$

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

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

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

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

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

$\dots \Rightarrow aaaBBCCBC \Rightarrow \dots \Rightarrow aaaBBCCBC \Rightarrow$

$\dots \Rightarrow aa**aB**BBCCC \Rightarrow aa**ab**BBCCC$

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC,$
 $CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC,$
 $aB \rightarrow ab, bB \rightarrow bb, C \rightarrow c \}$

$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aaaBCBCBC \Rightarrow$
 $\dots \Rightarrow aaaBBCCBC \Rightarrow \dots \Rightarrow aaaBBCCBC \Rightarrow$
 $\dots \Rightarrow aaaBBBCCC \Rightarrow aaa**b**BCCC \Rightarrow$
 $aaa**bb**BCCC$

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC,$
 $CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC,$
 $aB \rightarrow ab, bB \rightarrow bb, C \rightarrow c \}$

$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aaaBCBCBC \Rightarrow$
 $\dots \Rightarrow aaaBBCCBC \Rightarrow \dots \Rightarrow aaaBBCCBC \Rightarrow$
 $\dots \Rightarrow aaaBBBCCC \Rightarrow aaabBBCCC \Rightarrow$
 $aaab**b**BCCC \Rightarrow aaab**bb**CCC$

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC,$
 $CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC,$
 $aB \rightarrow ab, bB \rightarrow bb, C \rightarrow c \}$

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

Beispiel einer kontextsensitiven Grammatik

$G-1 = (N, T, P, S)$ mit

$N = \{ S, B, C, X \}$

$T = \{ a, b, c \}$

$P = \{ S \rightarrow aSBC, S \rightarrow aBC,$
 $CB \rightarrow XB, XB \rightarrow XC, XC \rightarrow BC,$
 $aB \rightarrow ab, bB \rightarrow bb, C \rightarrow c \}$

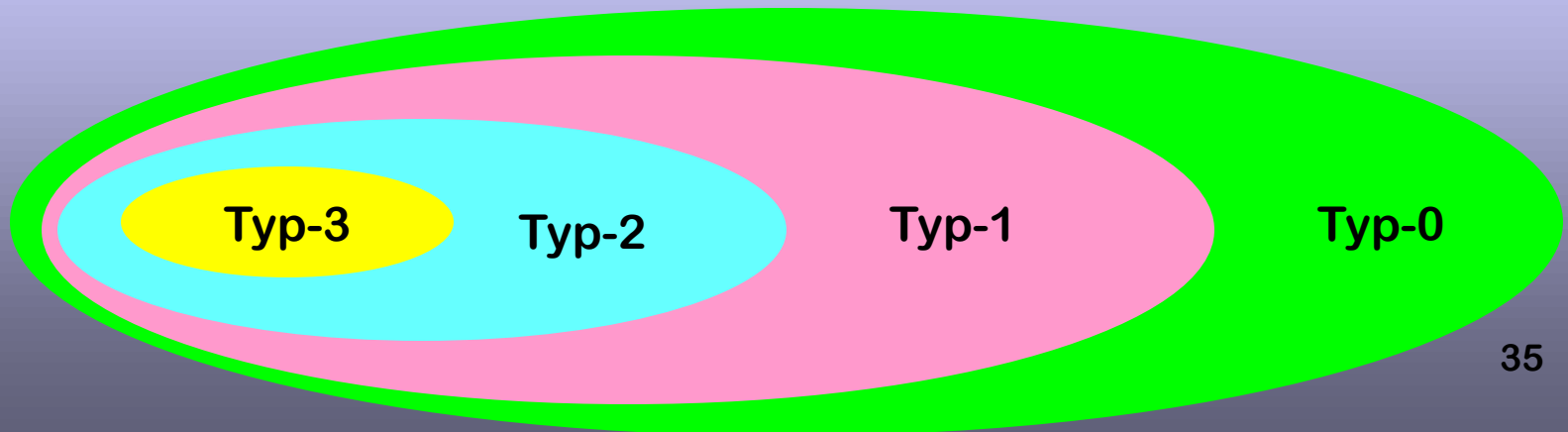
$S \Rightarrow aSBC \Rightarrow aaSBCBC \Rightarrow aaaBCBCBC \Rightarrow$
 $\dots \Rightarrow aaaBBCCBC \Rightarrow \dots \Rightarrow aaaBBCCBC \Rightarrow$
 $\dots \Rightarrow aaaBBBCCC \Rightarrow aaabBBCCC \Rightarrow$
 $aaabbBCCC \Rightarrow aaabbbCCC \Rightarrow \dots \text{aaabbbccc}$

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**.



Formale Eigenschaften natürlicher Sprachen

- Natürliche Sprachen werden im Folgenden als formale Sprachen (Mengen von Wörtern) betrachtet
- Problemstellung:
Welcher Typ formaler Sprachen charakterisiert natürliche Sprachen ?

Natürliche Sprachen als reguläre Sprachen

- Natürliche Sprachen sind ausdrucksstärker als reguläre Sprachen (Typ-3)
 - Beweisbar durch Pumping-Theorem und die Tatsache, dass reguläre Sprachen unter Mengenschnitt abgeschlossen sind ($L_{r1} \cap L_{r2} = L_{r3}$)
 - Illustration durch Sprachdaten

Sind Natürliche Sprachen regulär ?

Satzbeispiele:

(1) The cat waited.

(2) The cat **the dog admired** waited.

(3) The cat **the dog the ant bit admired** waited.

Satzformel:

(„the“ N)ⁿ ($V_{\text{transitiv}}$)ⁿ⁻¹ („waited“ [$V_{\text{intransitiv}}$])¹

bzw. $a^n b^{n-1} x$

Beweisidee (Abschluss unter Schnitt):

Englisch_{reg} [~~a~~ Menge von Ketten] $\cap N^* V^* x = N^n V^{n-1} x$

Aber: $a^n b^n$ ist eine CFL (pumping lemma)!

Natürliche Sprachen als kontextfreie Sprachen

- Einige natürliche Sprachen sind (etwas) ausdrucksstärker als kontextfreie Sprachen
 - Beweisbar durch Pumping Theorem und die Tatsache, dass kontextfreie Sprachen unter Mengenschnitt mit regulären Sprachen abgeschlossen sind ($L_{cf1} \cap L_{r2} = L_{cf3}$)
 - Illustration durch Sprachdaten

Sind alle Natürlichen Sprachen kontextfrei ?

Satzbeispiele (cross-serial dependencies):

- (1) Jan säit.
- (2) Jan säit **das mer em Hans hälfed**.
- (3) Jan säit **das mer em Hans es huus hälfed aastriiche**.
- (4) Jan säit **das mer d'chind em Hans es huus lönd hälfe aastriiche**.
- (5) Jan sagte, dass wir - **die Kinder**_{AKK} - **dem Hans**_{DAT} - **das Haus lassen**_{AKK} - **helfen**_{DAT} - **anzustreichen**

Satzformel:

Jan säit das mer („d'chind“_{AKK})ⁿ („em Hans“_{DAT})^m es huus („lönd“_{AKK})ⁿ („hälfe“_{DAT})^m aastriiche.

bzw. **w aⁿ b^m x cⁿ d^m y**

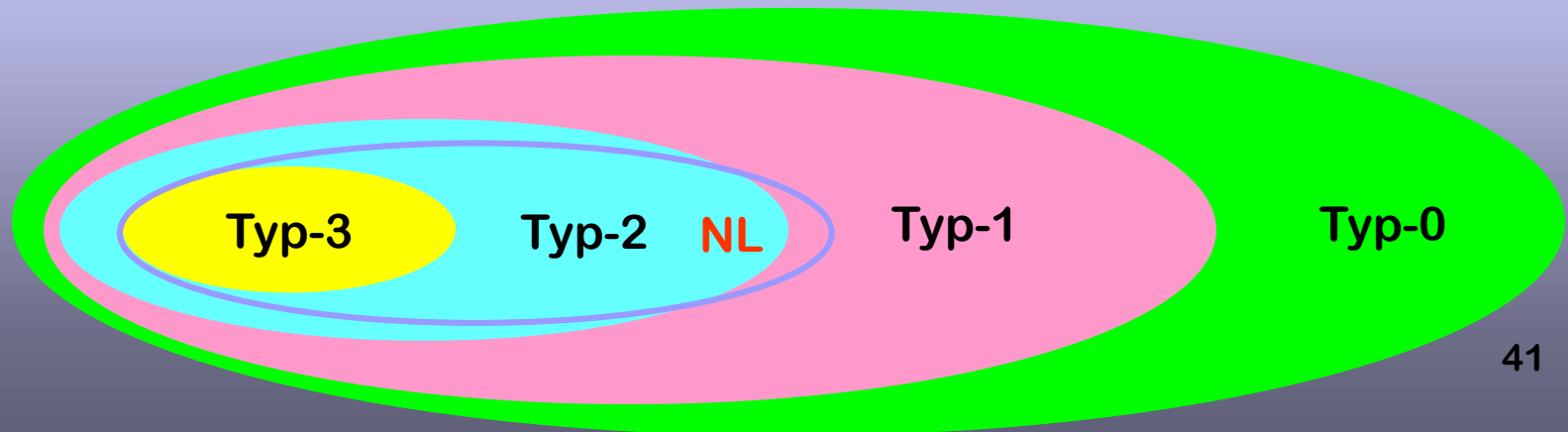
Beweisidee (Abschluss unter Schnitt):

SchweizerDeutsch_{cf} [als Menge von Ketten] \cap Jan säit das mer
(N_{AKK})^{*} (N_{DAT})^{*} es huus (V_{AKK})^{*} (V_{DAT})^{*} aastriiche
= Jan säit das mer (N_{AKK})ⁿ (N_{DAT})^m es huus (V_{AKK})ⁿ (V_{DAT})^m
aastriiche.

Aber: **w aⁿ b^m x cⁿ d^m y** ist eine **CSL** (pumping lemma)!

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



Kostenrechnung für det. FSA- Erkennungsalgorithmus

Funktion D-Erkennen(↓Band, ↓FSA) = „accept“ oder „reject“

Index \Leftarrow Bandanfang

AktualZustand \Leftarrow Anfangszustand des FSA

LOOP

IF Ende der Eingabekette ist erreicht THEN

IF AktualZustand ist ein Endzustand THEN return „accept“

ELSE return „reject“

ELSE-IF Zustandstranstionstabelle[AktualZustand, Band(Index)] = 0 THEN

return „reject“

ELSE AktualZustand \Leftarrow Zustandstranstionstabelle[AktualZustand, Band(Index)]

Index \Leftarrow Index + 1

LOOPEND

1

1

n

1







1

1

1

1

Komplexitätsklassen

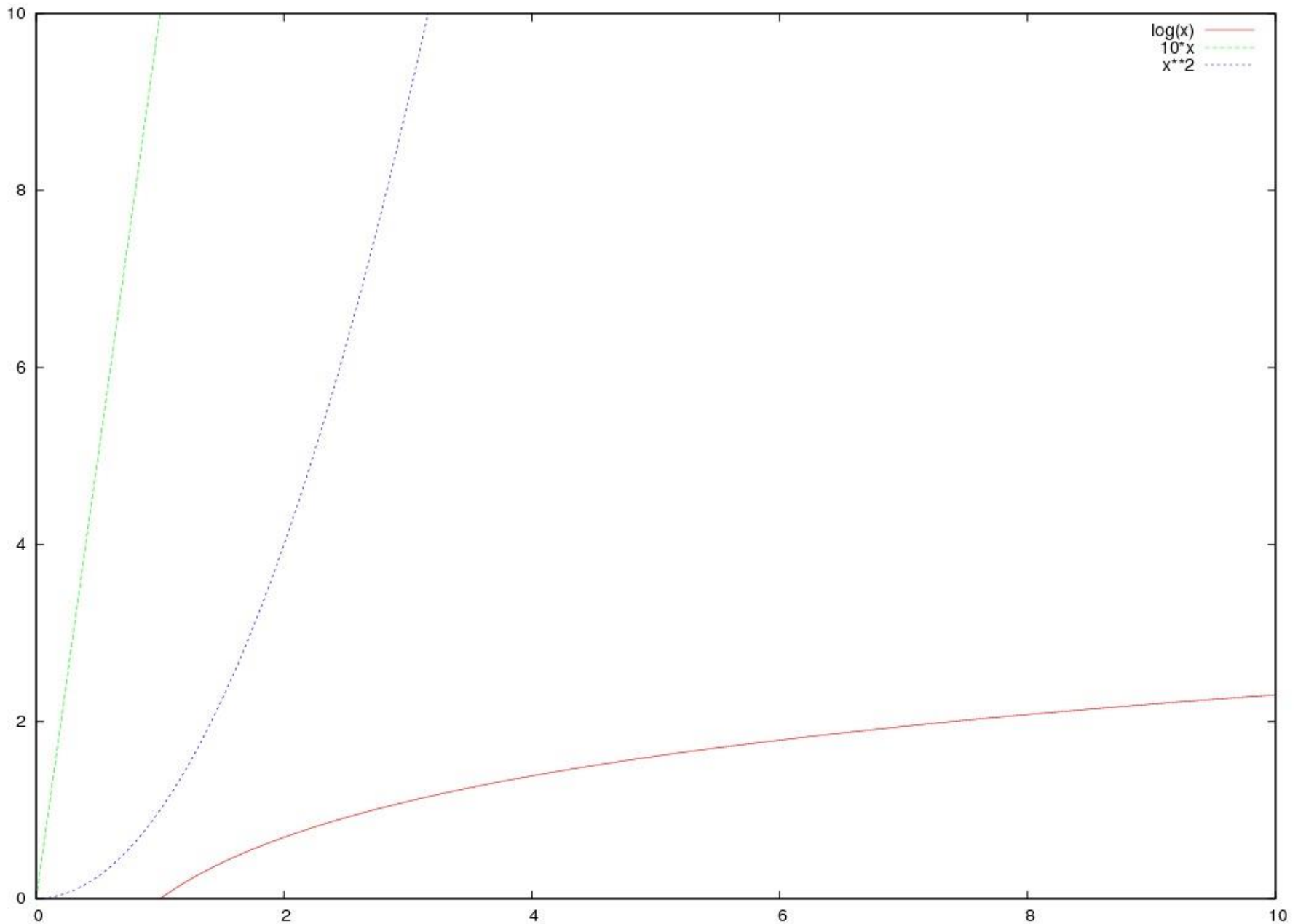
- $O(1)$ konstant 
- $O(\log n)$ logarithmisch 
- $O(n)$ linear 
- $O(n^k)$ polynomial ($k \in [2,4]$) 
- $O(n^k)$ polynomial ($k > 4$) 
- $O(k^n)$ exponentiell 

– wobei n die Problemgröße ist

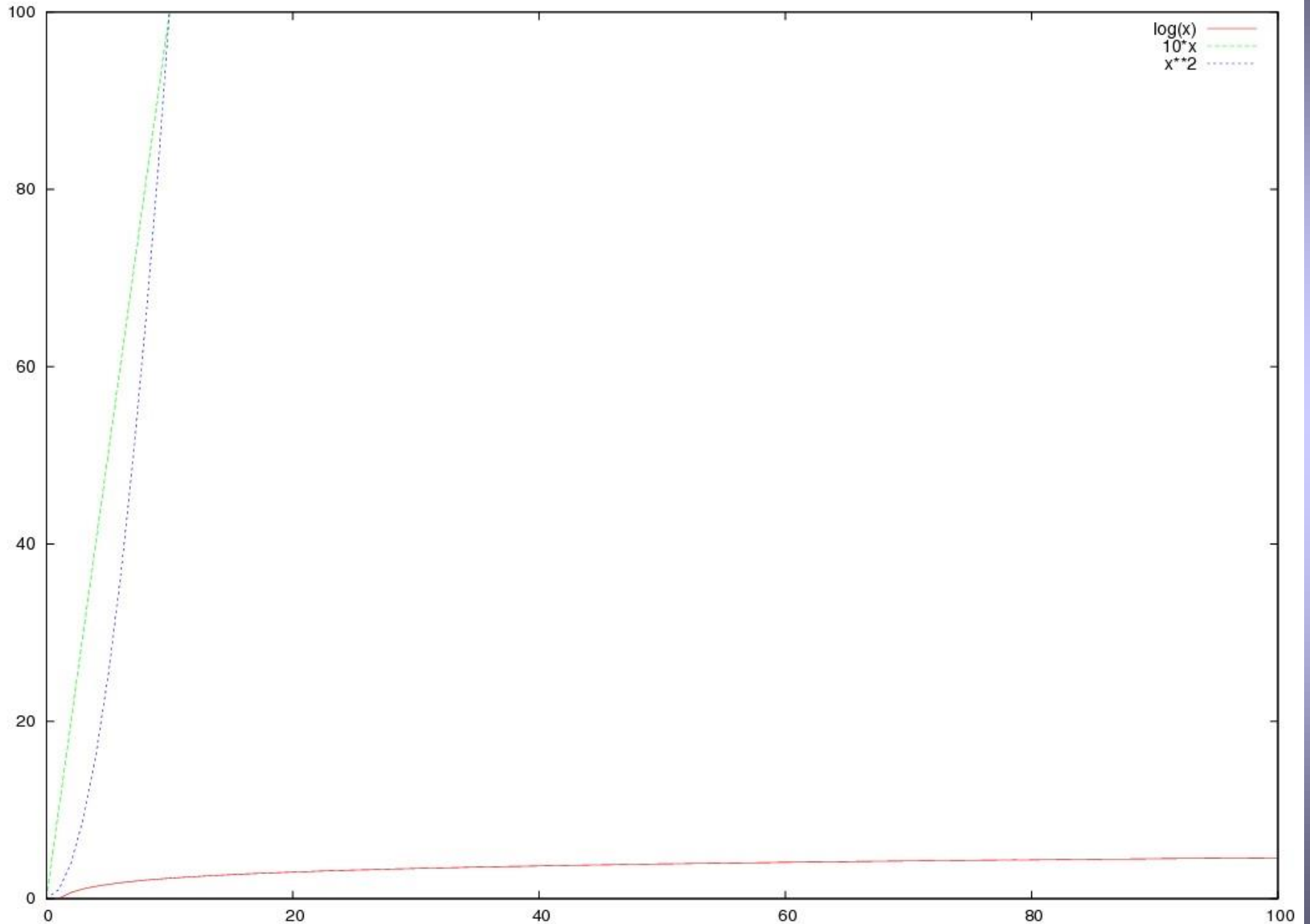
Charakteristische Problemgrößen beim Parsing

- n ist die Länge der Eingabe (Wort, Satz)
- n ist die Kardinalität der Produktionsregelmenge
- Hinweis: n ist vom (Parsing-)Algorithmus abhängig

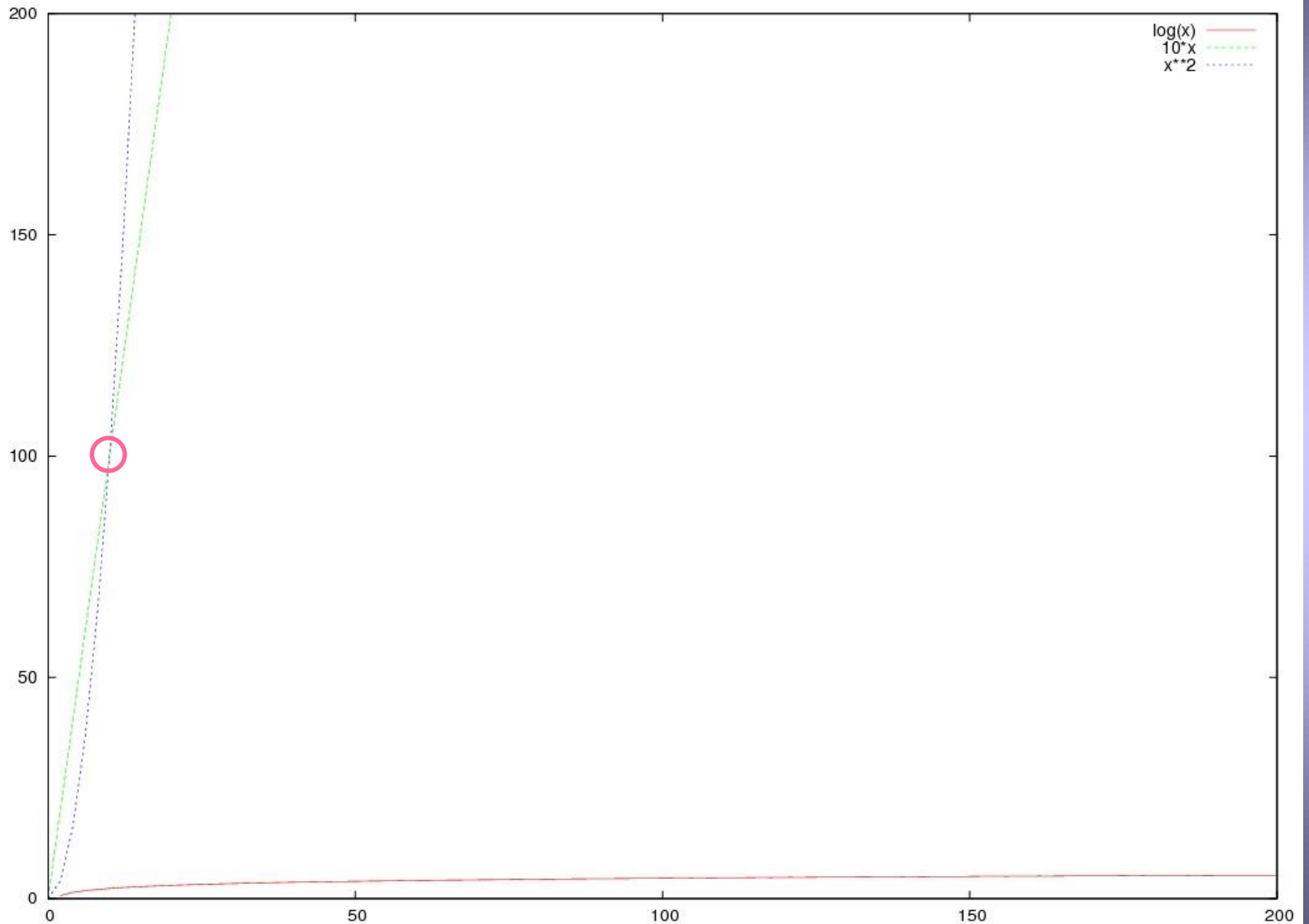
(Lauf-)Zeit-Klassen



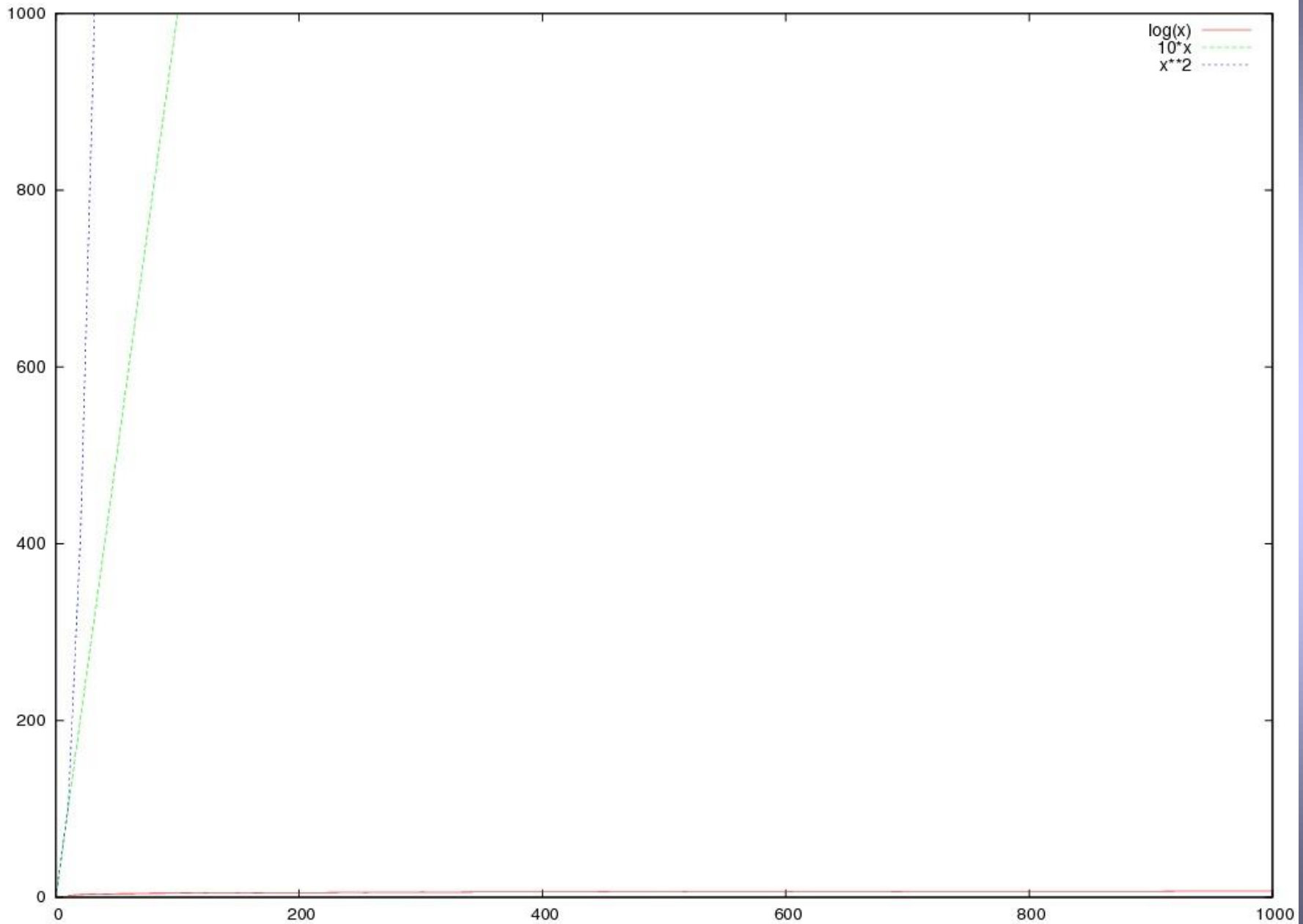
(Lauf-)Zeit-Klassen



(Lauf-)Zeit-Klassen



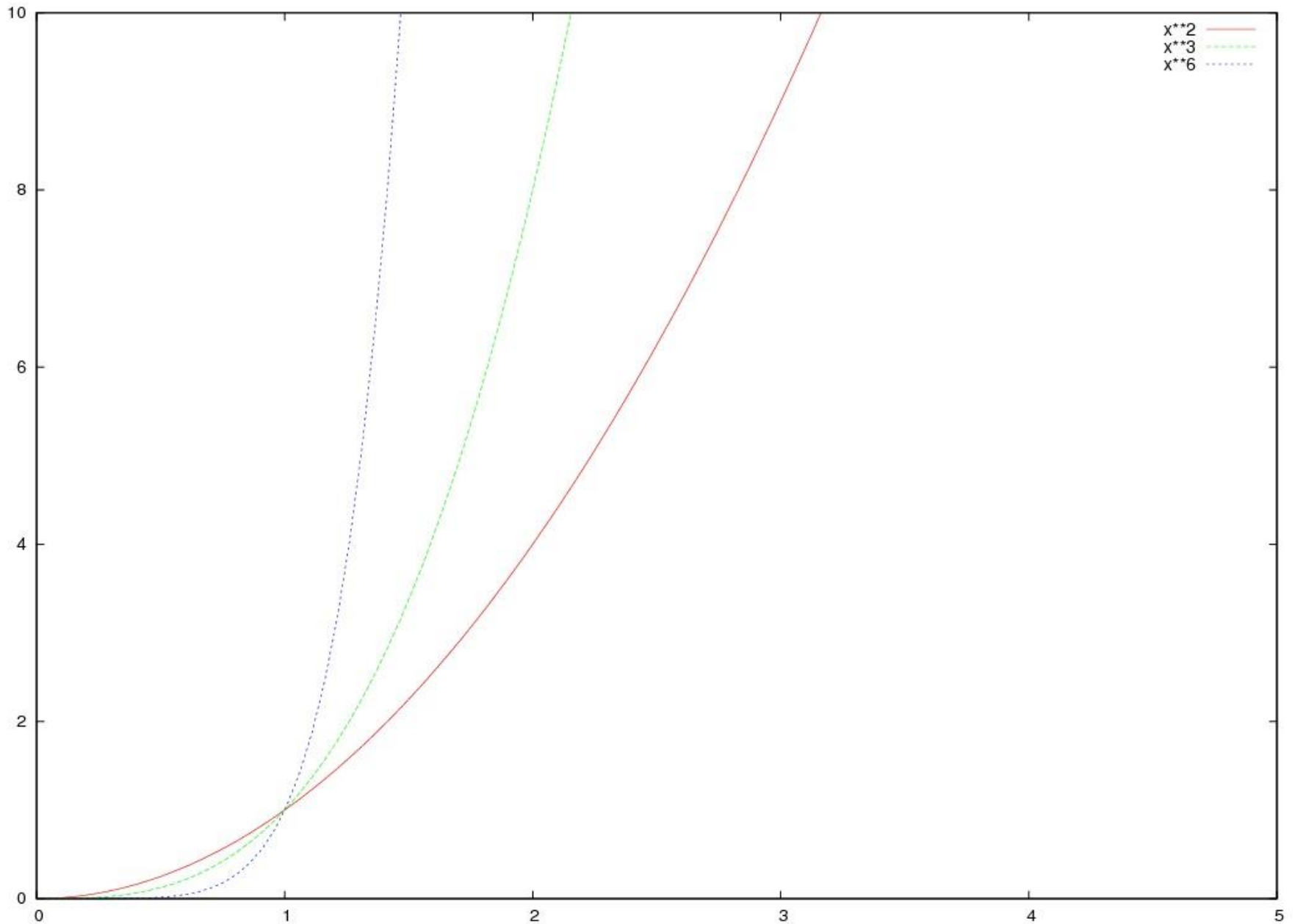
(Lauf-)Zeit-Klassen



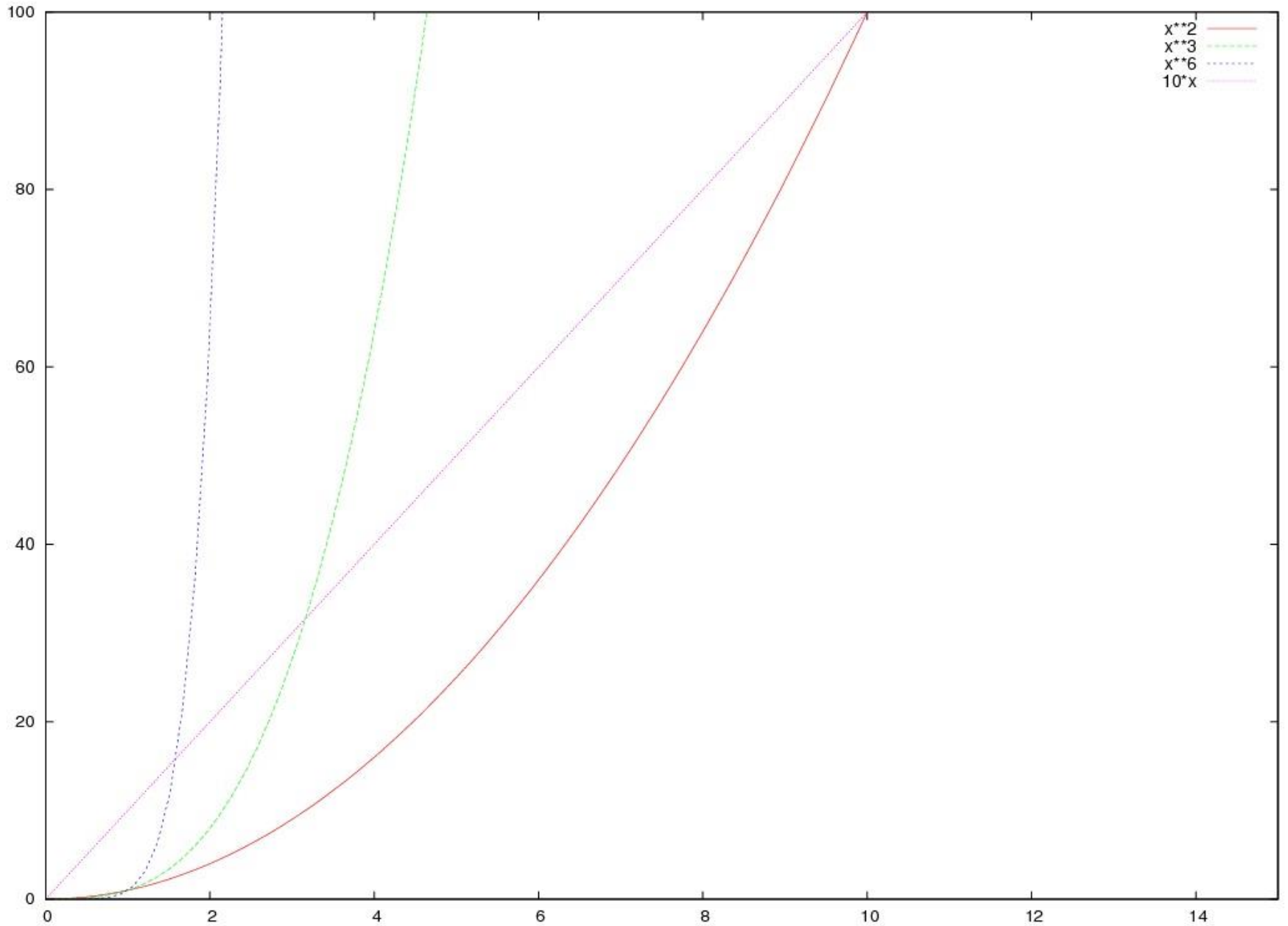
Implikationen für die automatische Sprachanalyse

- NLs sind keine **Typ-3-Sprachen**
 - Trotzdem werden **endliche Automaten (FSA)** für NLP-Analytik eingesetzt
 - **Lineare** Laufzeit ($O(n)$)
- NLs sind überwiegend (Englisch, Deutsch, Französisch, Spanisch, ...) **Typ-2-Sprachen**
 - **Kellerautomaten** als Basismodell
 - Syntaxanalyse in max. **kubischer** Laufzeitkomplexität ($O(n^3)$)
- Einige wenige NLs sind sicher (Schweizer Deutsch) bzw. vermutlich (Niederländisch, Bambara [Mali]) milde **Typ-1-Sprachen**
 - Syntaxanalyse in max. $O(n^6)$ Laufzeitkomplexität
 - Grammatikmodell: Tree Adjoining Grammar (TAG)

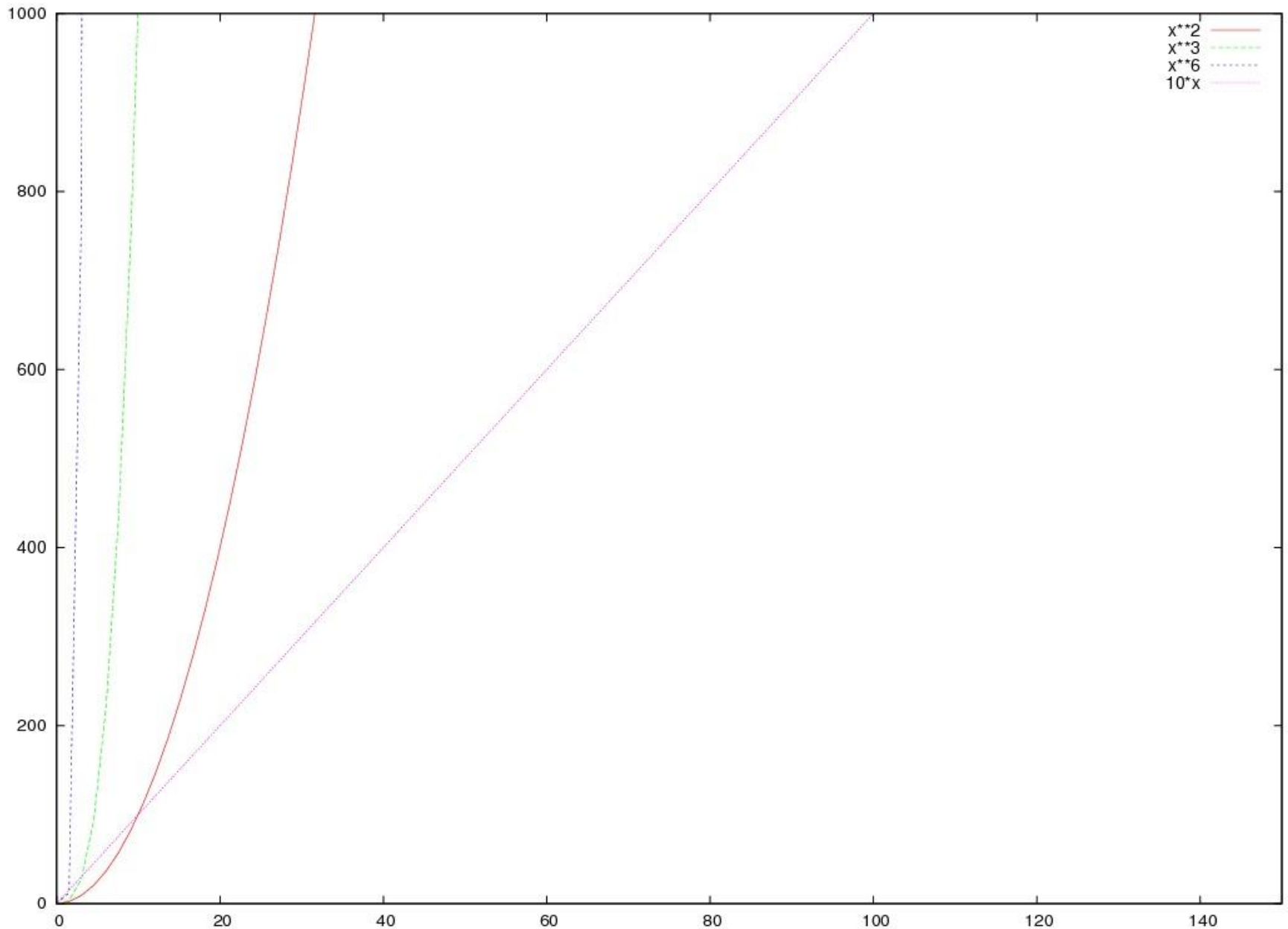
(Lauf-)Zeit-Klassen



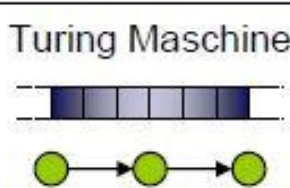
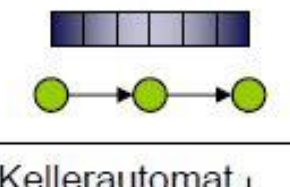


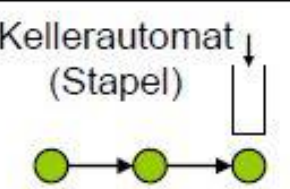





(Lauf-)Zeit-Klassen



(Lauf-)Zeit-Klassen



Chomsky-Hierarchie formaler Sprachen und Automaten

<i>Sprache</i>	<i>Automat</i>	<i>Grammatik</i>	<i>Erkennung</i>	<i>Abhängigkeit</i>
rekursiv aufzählbar	Turing Maschine 	unbeschränkt $Baa \rightarrow s$	unentscheidbar	beliebig
kontext- sensitiv	Linear gebunden 	kontext- sensitiv $At \rightarrow aA$	NP-vollständig 	überkreuzt 
kontext- frei	Kellerautomat (Stapel) 	kontextfrei $S \rightarrow gSc$	polynomiell 	eingebettet 
regulär	Endlicher Automat 	regulär $A \rightarrow cA$	linear 	strikt lokal 

Ableitungen in kontextfreien Grammatiken

Die folgenden Festlegungen zu Ableitungen beruhen auf der Annahme, dass die zugrundeliegende Grammatik kontextfrei ist!

Grundbegriffe zu formalen Grammatiken

- Eine Grammatik G heißt **Typ-2-Grammatik**,
(**kontextfreie Grammatik**), wenn P nur Pro-
duktionen enthält der Gestalt

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

Ableitungen in kontextfreien Grammatiken

- Ein **nicht-identischer Ableitungsschritt** ist definiert als Tripel

$\delta = [\ell, A \rightarrow \gamma, r]$ mit $\ell, r \in \mathcal{V}^*$ und $A \rightarrow \gamma \in P$

δ beschreibt die Beziehung: $\ell A r \Rightarrow \ell \gamma r$

- Als **Quelle** bzw. **Ziel** des Ableitungsschrittes δ gelten $Q(\delta) = \ell A r$ bzw. $Z(\delta) = \ell \gamma r$.

Ableitungen in kontextfreien Grammatiken

- Eine endliche nichtleere Folge

$$\Delta = \{ \delta_i \}_{i=1}^n$$

von Ableitungsschritten δ_i mit $\mathcal{Q}(\delta_{i+1}) = \mathcal{Z}(\delta_i)$ für $1 \leq i < n$ heißt **Ableitung** von $\mathcal{Q}(\delta_1)$ nach $\mathcal{Z}(\delta_n)$.

- **Quelle** von Δ ist $\mathcal{Q}(\Delta) = \mathcal{Q}(\delta_1)$, **Ziel** von Δ ist $\mathcal{Z}(\Delta) = \mathcal{Z}(\delta_n)$.

Ableitungen in kontextfreien Grammatiken

- Als **Länge einer Ableitung** $\Delta = \{ \delta_i \}_{i=1}^n$ ist die Anzahl von nicht-identischen Ableitungsschritten in Δ definiert.
 - Sei Δ eine Ableitung mit $Q(\Delta) = s$ und $Z(\Delta) = z$ der Länge ℓ . Dann gilt: $s \xrightarrow{\ell} z$.
 - Umgekehrt folgt aus $s \xrightarrow{\ell} z$ die Existenz einer Ableitung der Länge ℓ mit $Q(\Delta) = s$ und $Z(\Delta) = z$.

Ableitung für „aaaabbbb“

$$\delta_1 = [\varepsilon, S \rightarrow aSb, \varepsilon]$$

Beispiel einer kontextfreien Grammatik

$\mathcal{G}_2 = (N, T, P, S)$ mit

$N = \{ S \}$

$T = \{ a, b \}$

$P = \{ S \rightarrow aSb,$

$S \rightarrow ab \}$

$\mathcal{L}(\mathcal{G}_2) = \{ ab, aabb, aaabbb, aaaabbbb, \dots \}$

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

$$Q(\delta_1) = \varepsilon S \varepsilon, Z(\delta_1) = \varepsilon a S b \varepsilon$$

Ableitung für „aaaabbbb“

$$\delta_1 = [\varepsilon, S \rightarrow aSb, \varepsilon]$$

$$\delta_2 = [a, S \rightarrow aSb, b]$$

Beispiel einer kontextfreien Grammatik

$\mathcal{G}_2 = (N, T, P, S)$ mit

$N = \{ S \}$

$T = \{ a, b \}$

$P = \{ S \rightarrow aSb,$

$S \rightarrow ab \}$

$\mathcal{L}(\mathcal{G}_2) = \{ ab, aabb, aaabbb, aaaabbbb, \dots \}$

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

59

$$Q(\delta_1) = \varepsilon S \varepsilon, Z(\delta_1) = \varepsilon a S b \varepsilon$$
$$Q(\delta_2) = a S b, Z(\delta_2) = a a S b b$$

Ableitung für „aaaabbbb“

$$\delta_1 = [\varepsilon, S \rightarrow aSb, \varepsilon]$$

$$\delta_2 = [a, S \rightarrow aSb, b]$$

$$\delta_3 = [aa, S \rightarrow aSb, bb]$$

Beispiel einer kontextfreien Grammatik

$\mathcal{G}_2 = (N, T, P, S)$ mit

$N = \{ S \}$

$T = \{ a, b \}$

$P = \{ S \rightarrow aSb,$

$S \rightarrow ab \}$

$\mathcal{L}(\mathcal{G}_2) = \{ ab, aabb, aaabbb, aaaabbbb, \dots \}$

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

59

$$Q(\delta_1) = \varepsilon S \varepsilon, Z(\delta_1) = \varepsilon a S b \varepsilon$$

$$Q(\delta_2) = a S b, Z(\delta_2) = aa S bb$$

$$Q(\delta_3) = aa S bb, Z(\delta_3) = aaa S bbb$$

Ableitung für „aaaabbbb“

$$\delta_1 = [\varepsilon, S \rightarrow aSb, \varepsilon]$$

$$\delta_2 = [a, S \rightarrow aSb, b]$$

$$\delta_3 = [aa, S \rightarrow aSb, bb]$$

$$\delta_4 = [aaa, S \rightarrow ab, bbb]$$

$$Q(\delta_1) = \varepsilon S \varepsilon, Z(\delta_1) = \varepsilon a S b \varepsilon$$

$$Q(\delta_2) = a S b, Z(\delta_2) = aa S bb$$

$$Q(\delta_3) = aa S bb, Z(\delta_3) = aaa S bbb$$

$$Q(\delta_4) = aaa S bbb, Z(\delta_4) = aaa ab^6 bbb$$

Beispiel einer kontextfreien Grammatik

$\mathcal{G}_2 = (N, T, P, S)$ mit

$N = \{ S \}$

$T = \{ a, b \}$

$P = \{ S \rightarrow aSb,$

$S \rightarrow ab \}$

$\mathcal{L}(\mathcal{G}_2) = \{ ab, aabb, aaabbb, aaaabbbb, \dots \}$

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

Ableitung für „aaaabbbb“

$$\begin{aligned}\delta_1 &= [\varepsilon, S \rightarrow aSb, \varepsilon] \\ \delta_2 &= [a, S \rightarrow aSb, b] \\ \delta_3 &= [aa, S \rightarrow aSb, bb] \\ \delta_4 &= [aaa, S \rightarrow ab, bbb]\end{aligned}$$

Beispiel einer kontextfreien Grammatik

$\mathcal{G}_2 = (N, T, P, S)$ mit

$N = \{S\}$

$T = \{a, b\}$

$P = \{ S \rightarrow aSb,$

$S \rightarrow ab \}$

$\mathcal{L}(\mathcal{G}_2) = \{ab, aabb, aaabbb, aaaabbbb, \dots\}$

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

59

$$\begin{aligned}Q(\Delta) &= Q(\delta_1) = S, \\ Z(\Delta) &= Z(\delta_4) = aaaabbbb\end{aligned}$$

$$Q(\delta_1) = \varepsilon S \varepsilon, \quad Z(\delta_1) = \varepsilon aSb \varepsilon$$

$$Q(\delta_2) = aSb, \quad Z(\delta_2) = aaSbb$$

$$Q(\delta_3) = aaSbb, \quad Z(\delta_3) = aaaSbbb$$

$$Q(\delta_4) = aaaSbbb, \quad Z(\delta_4) = aaaabbbb$$

63

Ableitungen in kontextfreien Grammatiken

- Ein Ableitungsschritt $\delta = [\ell, A \rightarrow \gamma, r]$ heißt **Links-** bzw. **Rechtsableitungsschritt**, wenn $\ell \in \mathcal{T}^*$ bzw. $r \in \mathcal{T}^*$.
- Man schreibt $s \Rightarrow_L z$ bzw. $s \Rightarrow_R z$, wenn z aus s durch einen Links- bzw. Rechtsableitungsschritt hervor geht.
 - D.h.: z entsteht aus s durch Ersetzen des in s am weitesten links bzw. rechts stehenden Nichtterminalsymbols gemäß einer Produktion $A \rightarrow \gamma$.

Ableitungen in kontextfreien Grammatiken

- Eine **Links-** bzw. **Rechtsableitung** ist eine Ableitung, die aus einer endlichen nichtleeren Folge $\Delta = \{ \delta_i \}_{i=1}^n$ von nicht-identischen Ableitungsschritten $\delta_i = [\ell_i, A_i \rightarrow \gamma_i, r_i]$ besteht mit $\ell_i \in \mathcal{T}^*$ bzw. $r_i \in \mathcal{T}^*$ für $1 \leq i \leq n$.
- Man schreibt $s \xRightarrow{*}_R z$ bzw. $s \xRightarrow{\ell}_R z$, um auszudrücken, dass eine Rechtsableitung von s nach z bzw. eine Rechtsableitung der Länge ℓ von s nach z existiert (analog für Linksableitungen).

Ableitungen in kontextfreien Grammatiken

- Jede Zeichenfolge s mit $S^* \Rightarrow s$ heißt **Satzform**; im Fall $S^* \Rightarrow_L s$ bzw. $S^* \Rightarrow_R s$ auch **Links-** bzw. **Rechtssatzform**.

Ableitungen in kontextfreien Grammatiken

- Sei Π eine Menge von **Marken**, mit denen die Produktionen P der zugrundeliegenden Grammatik G eindeutig identifiziert werden können.

Eine Folge $\pi = p_1, \dots, p_n$ mit $p_i \in \Pi$, $1 \leq i \leq n$, heißt **Kontrollwort (Parse)** einer Ableitung $\Delta = \{ \delta_i \}_{i=1}^n$, wenn für $1 \leq i \leq n$ die im i -ten Ableitungsschritt angewandte Produktion durch die Marke p_i gekennzeichnet ist.

Ableitungen in kontextfreien Grammatiken

- Um auszudrücken, dass π Kontrollwort einer Ableitung von S nach Z ist, schreibt man auch $S \xRightarrow{\pi} Z$.
- Falls $\omega \in \mathcal{L}(G)$ und $S \xRightarrow{\pi} \omega$, heißt π auch **Kontrollwort (Parse)** für ω .
- Analog gibt es ein **Linkskontrollwort (Links-
Parse)** bzw. ein **Rechtskontrollwort (Rechts-
Parse)**, wenn die dazugehörige Ableitung eine Links- bzw. Rechtsableitung ist, also:
 $S \xRightarrow{\pi}_L Z$ bzw. $S \xRightarrow{\pi}_R Z$.

Kontrollwort für „aaaabbbb“

Beispiel einer kontextfreien Grammatik

$\mathcal{G}_2 = (N, T, P, S)$ mit

$N = \{ S \}$

$T = \{ a, b \}$

$P = \{ S \rightarrow aSb, \quad 1$

$S \rightarrow ab \} \quad 2$

$\mathcal{L}(\mathcal{G}_2) = \{ ab, aabb, aaabbb, aaaabbbb, \dots \}$
 $= a^n b^n, n \geq 1$

59

$$\delta_1 = [\varepsilon, S \rightarrow aSb, \varepsilon]$$

$$\delta_2 = [a, S \rightarrow aSb, b]$$

$$\delta_3 = [aa, S \rightarrow aSb, bb]$$

$$\delta_4 = [aaa, S \rightarrow ab, aaa]$$

$$Q(\Delta) = Q(\delta_1) = S,$$

$$Z(\Delta) = Z(\delta_4) = aaaabbbb$$

$$\pi_{aaaabbbb} = 1112$$

$$S \xrightarrow{\pi} aaaabbbb$$

$$Q(\delta_1) = \varepsilon S \varepsilon, \quad Z(\delta_1) = \varepsilon a S b \varepsilon$$

$$Q(\delta_2) = a S b, \quad Z(\delta_2) = aa S bb$$

$$Q(\delta_3) = aa S bb, \quad Z(\delta_3) = aaa S bbb$$

$$Q(\delta_4) = aaa S bbb, \quad Z(\delta_4) = aaaabbbb$$

69

Ableitungen in kontextfreien Grammatiken

- Eine **Ableitung** $\Delta = \{ \delta_i \}_{i=1}^n$ ist im Allgemeinen durch ihre Quelle und ihr Kontrollwort nicht eindeutig bestimmt.
- Dagegen ist jede **Links- bzw. Rechtsableitung** durch ihre Quelle und ihr Kontrollwort eindeutig bestimmt. Durch einen **Links- bzw. Rechts-Parser** für ein Wort $\omega \in \mathcal{L}(G)$ wird also die zugehörige Links- bzw. Rechtsableitung eindeutig beschrieben.

Ableitungen in kontextfreien Grammatiken

Eine wesentliche Aufgabe der Syntaxanalyse besteht darin, für Wörter $\omega \in \mathcal{L}(G)$ eine Ableitung $S \xRightarrow{*} \omega$ zu bestimmen. Dazu kann die Umkehrung einer Rechtsableitung genutzt werden:

- Sei $\Delta = \{ \delta_i \}_{i=1}^n$ eine Ableitung (Rechtsableitung). Die aus $\Delta^a = [\delta_1, \delta_2, \dots, \delta_{n-1}, \delta_n]$ durch Inversion hervorgehende Folge $\Delta^r = [\delta_n, \delta_{n-1}, \dots, \delta_2, \delta_1]$ heißt **Reduktion** (bzw. **Rechtsreduktion** oder **kanonische Reduktion**) von $\mathcal{Z}(\Delta)$ nach $\mathcal{Q}(\Delta)$.