

Computerlinguistik II / Sprachtechnologie

Vorlesung im SS 2010
(M-GSW-10)

Prof. Dr. Udo Hahn

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

Allgemeine Hinweise

- Vorlesung: MO, 14-16h (FG 1, SR 259)
- Übung zV: DI, 16-18h (ZwG 4, SR)
 - beginnt am 13.4.
- Vorlesungsmaterialien im Netz
 - <http://www.julielab.de/>
 - ⇒ „Students“
- Sprechstunde: MI, 11-12h (FG 30, R 004)
- Email: udo.hahn@uni-jena.de
- **Fachliteratur ist überwiegend in Englisch**

2

Aufbau der Vorlesung

- Algebraisch-symbolisches Paradigma
 - Graphentheorie: Bäume
 - Syntaxanalyse (Parsing)
- Empirisches Paradigma
 - Tagging und Chunking
 - Korpora und Annotation
 - Lexika und NLP-Software
- Sprachtechnologie
 - Indexing und Klassifikation
 - Textzusammenfassung
 - Frage-Beantwortung und Informationsextraktion

3

Grundbegriffe zu Graphen

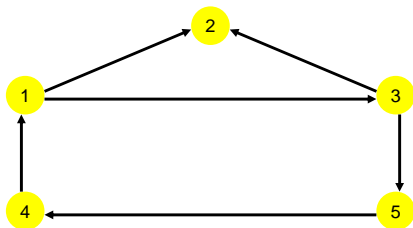
- Ein endlicher gerichteter **Graph** ist ein Paar $\Gamma = (K, \rho)$, wobei
 - K endliche Menge von **Knoten**
 - $\rho \subseteq K \times K$ endliche Menge von **Kanten**

4

Grundbegriffe zu Graphen

Beispiel für einen Graphen

$\Gamma = (\{1, 2, 3, 4, 5\},$
 $\{ (1,2), (1,3), (3,2), (3,5), (4,1), (5,4) \})$



Grundbegriffe zu Graphen

- Ein (endlicher) **Baum** ist ein endlicher gerichteter Graph $\Gamma = (K, \rho)$, wobei
 - K enthält genau einen Knoten k_w (die sog. **Wurzel**) mit der Eigenschaft, dass für jede Kante $(k, k') \in \rho$ gilt: $k_w \neq k'$;
 - Zu jedem Knoten $k \neq k_w$ gibt es genau eine Folge von Knoten $k_w = k_0, k_1, k_2, \dots, k_n = k$, $n \geq 1$, mit $(k_i, k_{i+1}) \in \rho$ für $0 \leq i \leq n-1$.

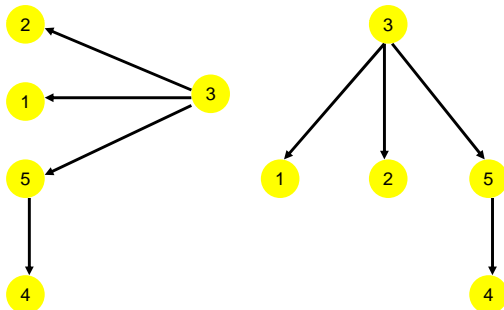
Grundbegriffe zu Graphen

Beispiele für Bäume

$\{ (3,1), (1,4) \}$



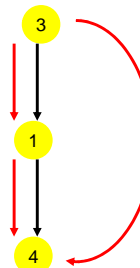
$\{ (3,1), (3,2), (3,5), (5,4) \}$



Grundbegriffe zu Graphen

Beispiele für Graphen, die **keine** Bäume sind

$\{ (3,1), (1,4), (3,4) \}$

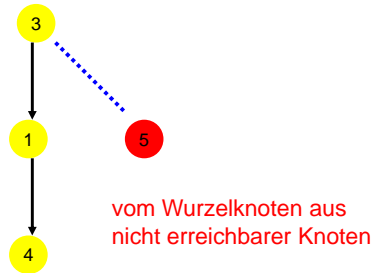


mehrere Wege

Grundbegriffe zu Graphen

Beispiele für Graphen, die **keine** Bäume sind

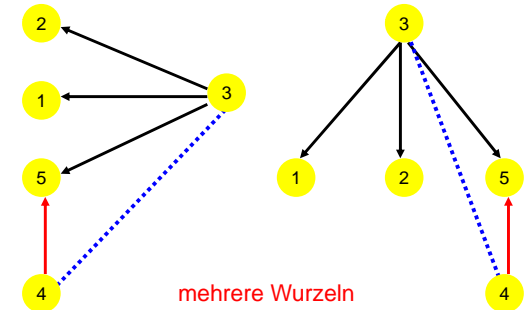
{ (3,1), (1,4) }



Grundbegriffe zu Graphen

Beispiele für Graphen, die **keine** Bäume sind

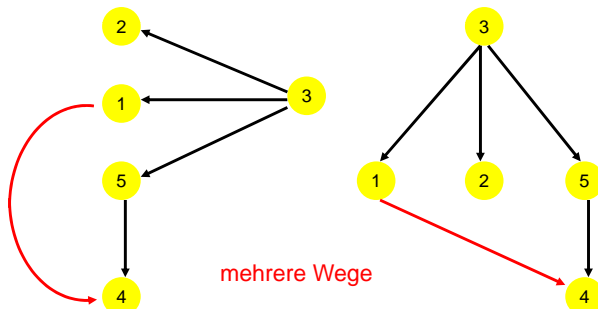
{ (3,1), (3,2), (3,5), (4,5) }



Grundbegriffe zu Graphen

Beispiele für Graphen, die **keine** Bäume sind

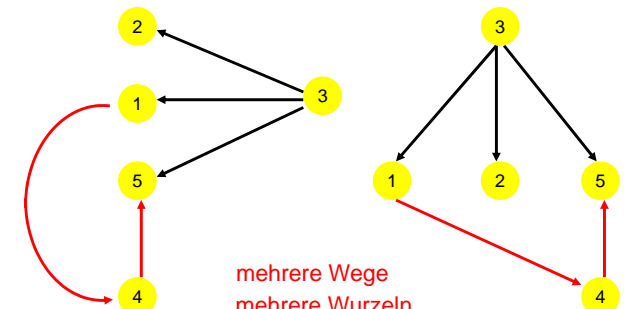
{ (3,1), (3,2), (3,5), (5,4), (1,4) }



Grundbegriffe zu Graphen

Beispiele für Graphen, die **keine** Bäume sind

{ (3,1), (3,2), (3,5), (4,5), (1,4) }



Grundbegriffe zu Graphen

- Eine Folge $k_0, k_1, k_2, \dots, k_n, n \geq 0$, von Knoten mit $(k_i, k_{i+1}) \in \rho$ für $0 \leq i \leq n-1$ heißt **Pfad der Länge n** von k_0 nach k_n .
- Sei $(k, k') \in \rho$, dann heißt k' **direkter Nachfolger** von k und k **direkter Vorgänger** von k' . Wenn k und k' Knoten sind, für die ein Pfad von k nach k' existiert, dann heißt k' **Nachfolger** von k bzw. k **Vorgänger** von k' .
- Ein Knoten k , der keinen Nachfolger hat, heißt **Endknoten** (oder **Blatt**).

13

Grundbegriffe zu Graphen

- Sei $\Gamma = (K, \rho)$ ein endlicher gerichteter Graph. Ein endlicher gerichteter Graph $\Gamma^* = (K^*, \rho^*)$ mit $K^* \subseteq K$ und $\rho^* \subseteq \rho \cap K^* \times K^*$ heißt **Teilgraph** von Γ .

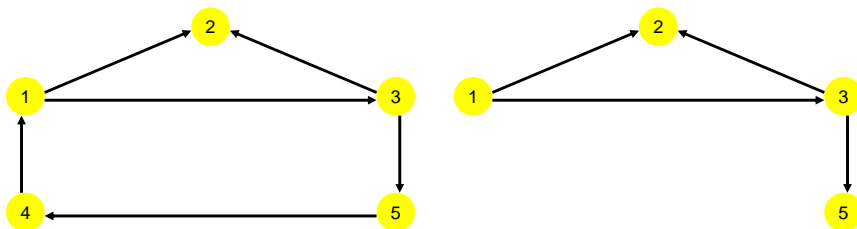
14

Grundbegriffe zu Graphen

Beispiel für einen Teilgraphen

$$\Gamma = (\{1, 2, 3, 4, 5\}, \{(1,2), (1,3), (3,2), (3,5), (4,1), (5,4)\})$$

$$\Gamma^* = (\{1, 2, 3, 5\}, \{(1,2), (1,3), (3,2), (3,5)\})$$



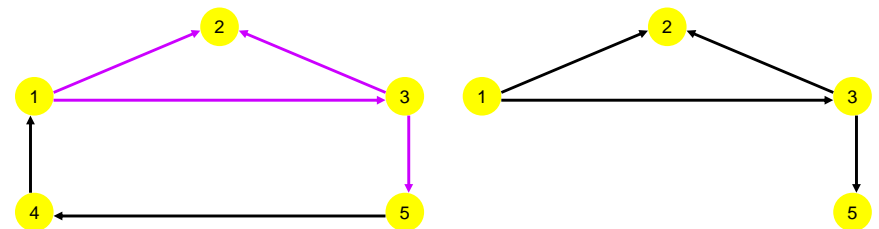
15

Grundbegriffe zu Graphen

Beispiel für einen Teilgraphen

$$\Gamma = (\{1, 2, 3, 4, 5\}, \{(1,2), (1,3), (3,2), (3,5), (4,1), (5,4)\})$$

$$\Gamma^* = (\{1, 2, 3, 5\}, \{(1,2), (1,3), (3,2), (3,5)\})$$



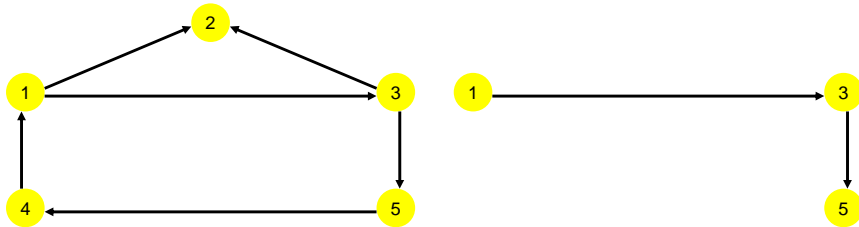
16

Grundbegriffe zu Graphen

Beispiel für einen Teilgraphen

$$\Gamma = (\{1, 2, 3, 4, 5\}, \{ (1,2), (1,3), (3,2), (3,5), (4,1), (5,4) \})$$

$$\Gamma^* = (\{1, 3, 5\}, \{ (1,3), (3,5) \})$$



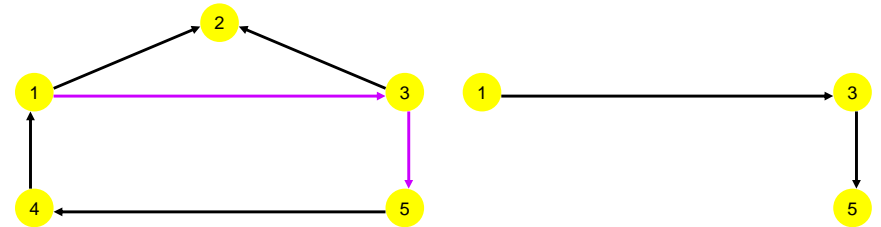
17

Grundbegriffe zu Graphen

Beispiel für einen Teilgraphen

$$\Gamma = (\{1, 2, 3, 4, 5\}, \{ (1,2), (1,3), (3,2), (3,5), (4,1), (5,4) \})$$

$$\Gamma^* = (\{1, 3, 5\}, \{ (1,3), (3,5) \})$$



18

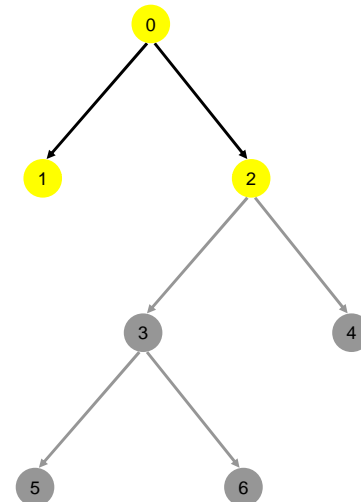
Grundbegriffe zu Graphen

- Sei Γ ein Baum. Ein Teilgraph Γ^* von Γ , der selbst wieder ein Baum ist, heißt **Teilbaum** von Γ .
 - Γ^* heißt **Anfangsteilbaum** von Γ , wenn Γ^* und Γ dieselbe Wurzel haben.
 - Γ^* heißt **Endteilbaum** von Γ , wenn alle Endknoten von Γ^* auch Endknoten in Γ sind.

19

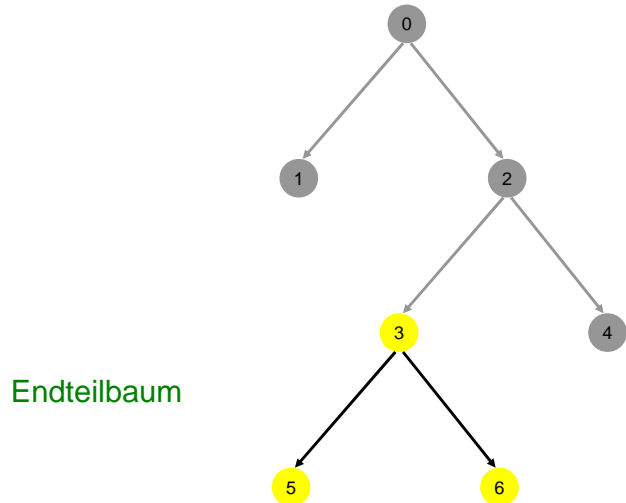
Grundbegriffe zu Graphen

Anfangsteilbaum

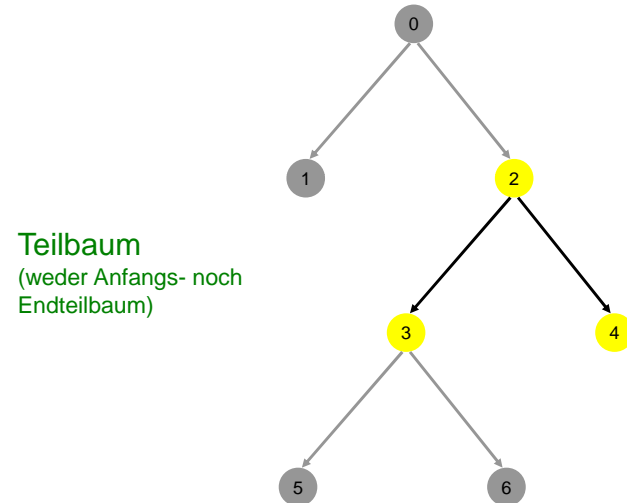


20

Grundbegriffe zu Graphen



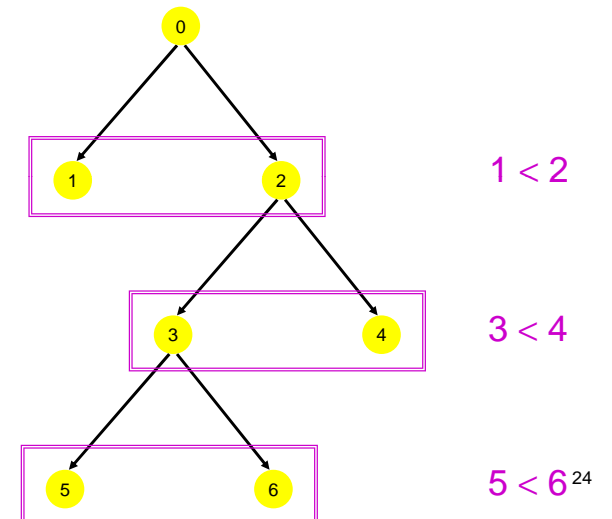
Grundbegriffe zu Graphen



Grundbegriffe zu Graphen

- Ein **partiell geordneter Baum** ist ein Tripel $\Gamma = (K, \rho, <)$, wobei „ $<$ “ eine partielle Ordnung auf K ist, die für jeden Knoten die Menge seiner direkten Nachfolger linear ordnet und auch nur solche Knoten in Bezug zueinander setzt, die direkte Nachfolger ein und desselben Vor-gängerknotens sind.

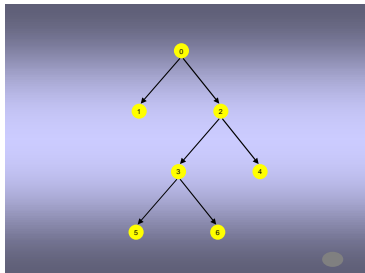
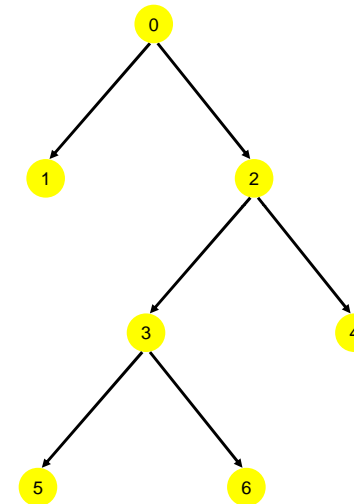
(Partiell) geordneter Baum



Grundbegriffe zu Graphen

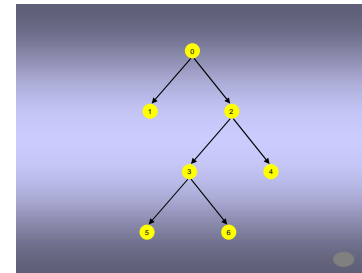
- Ein **vollständig geordneter Baum** ist ein Tripel $\Gamma = (K, \rho, <^*)$, wobei „ $<^*$ “ eine voll-ständige Ordnung auf K ist, die aus der partiellen Ordnung „ $<$ “ durch folgende Ver-einbarung aufgebaut wird. Seien k, k_1 und k_2 Knoten aus K . Dann gilt: $k_1 <^* k_2$ (ge-sprochen: „ k_1 ist vor | links von k_2 “), falls:
 - $k_1 < k_2$ oder
 - k_1 ist Nachfolger von k_2 oder
 - k_1 ist Nachfolger von k^i und $k^i < k_2$ oder
 - k_2 ist Nachfolger von k^i und $k_1 < k^i$.

Berechnung der Relation $<^*$



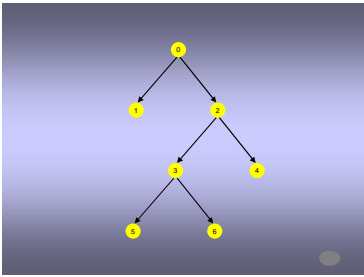
Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
---------	---------	---------



Berechnung der Relation $<^*$

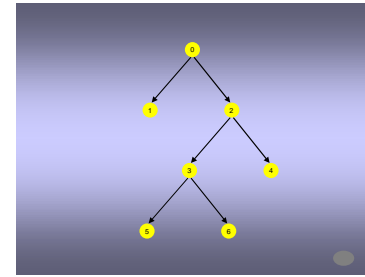
$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$



Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

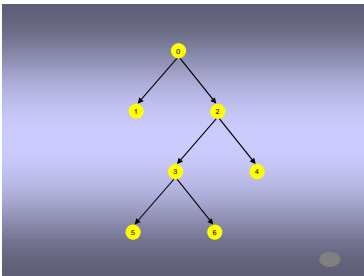
$5 <^* 3$
$5 <^* 2$
$5 <^* 0$



Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

$5 <^* 3$	$6 <^*$
3	
$5 <^* 2$	$6 <^* 2$
$5 <^* 0$	$6 <^* 0$

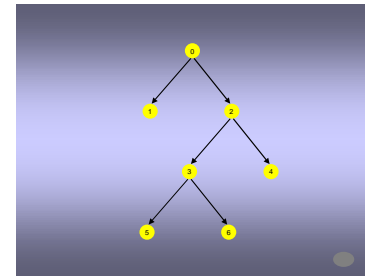


Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

$5 <^* 3$	$6 <^*$
3	
$5 <^* 2$	$6 <^* 2$
$5 <^* 0$	$6 <^* 0$

$3 <^* 2$
$3 <^* 0$

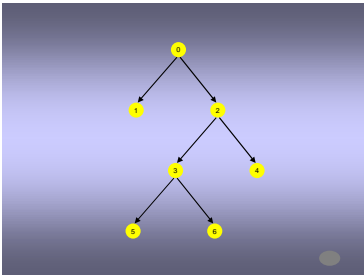


Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

$5 <^* 3$	$6 <^*$
3	
$5 <^* 2$	$6 <^* 2$
$5 <^* 0$	$6 <^* 0$

$3 <^* 2$	$4 <^*$
2	
$3 <^* 0$	$4 <^* 0$



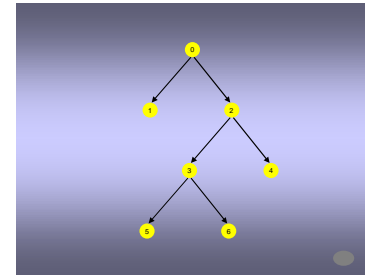
Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

$5 <^* 3$	$6 <^*$
3	
$5 <^* 2$	$6 <^* 2$
$5 <^* 0$	$6 <^* 0$

$3 <^* 2$	$4 <^*$
2	
$3 <^* 0$	$4 <^* 0$

$2 <^* 0$



Berechnung der Relation $<^*$

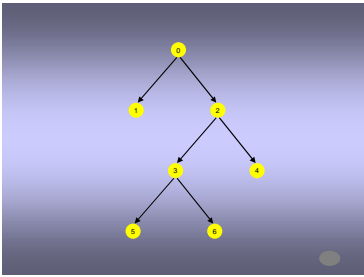
$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

$5 <^* 3$	$6 <^*$
3	
$5 <^* 2$	$6 <^* 2$
$5 <^* 0$	$6 <^* 0$

$3 <^* 2$	$4 <^*$
2	
$3 <^* 0$	$4 <^* 0$

$2 <^* 0$

$1 <^* 0$



Berechnung der Relation $<^*$

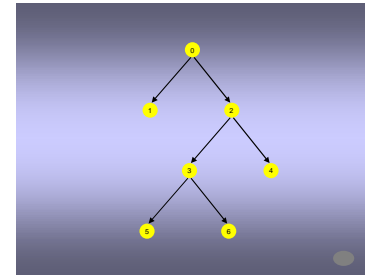
$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

$5 <^* 3$	$6 <^*$
3	
$5 <^* 2$	$6 <^* 2$
$5 <^* 0$	$6 <^* 0$

$3 <^* 2$	$4 <^*$
2	
$3 <^* 0$	$4 <^* 0$

$2 <^* 0$

$1 <^* 0$



Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

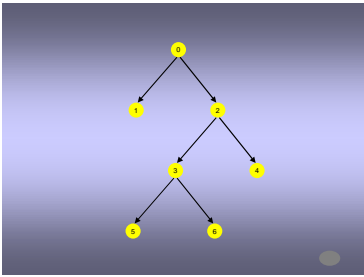
$5 <^* 3$	$6 <^*$
3	
$5 <^* 2$	$6 <^* 2$
$5 <^* 0$	$6 <^* 0$

$3 <^* 2$	$4 <^*$
2	
$3 <^* 0$	$4 <^* 0$

$2 <^* 0$

$1 <^* 0$

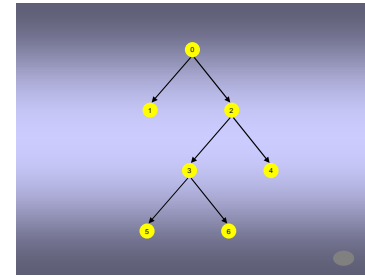
$5 <^* 6$



Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

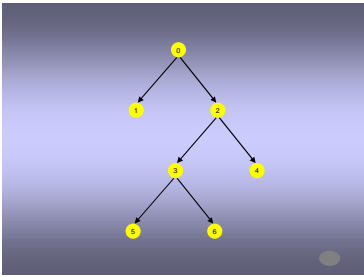
$5 <^* 3$	$6 <^*$	$3 <^* 2$	$4 <^*$	$2 <^* 0$	$1 <^* 0$
3		2			
$5 <^* 2$	$6 <^* 2$	$3 <^* 0$	$4 <^* 0$		
$5 <^* 0$	$6 <^* 0$				
$5 <^* 6 <^* 3$					



Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

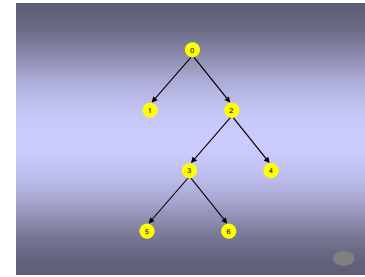
$5 <^* 3$	$6 <^*$	$3 <^* 2$	$4 <^*$	$2 <^* 0$	$1 <^* 0$
3		2			
$5 <^* 2$	$6 <^* 2$	$3 <^* 0$	$4 <^* 0$		
$5 <^* 0$	$6 <^* 0$				
$5 <^* 6 <^* 3 <^* 4$					



Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

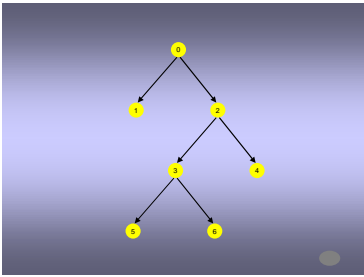
$5 <^* 3$	$6 <^*$	$3 <^* 2$	$4 <^*$	$2 <^* 0$	$1 <^* 0$
3		2			
$5 <^* 2$	$6 <^* 2$	$3 <^* 0$	$4 <^* 0$		
$5 <^* 0$	$6 <^* 0$				
$5 <^* 6 <^* 3 <^* 4 <^* 2$					



Berechnung der Relation $<^*$

$1 < 2$	$3 < 4$	$5 < 6$
$1 <^* 2$	$3 <^* 4$	$5 <^* 6$

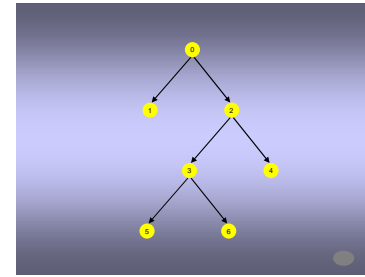
$5 <^* 3$	$6 <^*$	$3 <^* 2$	$4 <^*$	$2 <^* 0$	$1 <^* 0$
3		2			
$5 <^* 2$	$6 <^* 2$	$3 <^* 0$	$4 <^* 0$		
$5 <^* 0$	$6 <^* 0$				
$5 <^* 6 <^* 3 <^* 4 <^* 2 <^* 0$					



Berechnung der Relation $<^*$

$$\begin{array}{ccc} 1 < 2 & 3 < 4 & 5 < 6 \\ 1 <^* 2 & 3 <^* 4 & 5 <^* 6 \end{array}$$

$$\begin{array}{cc} 5 <^* 3 & 6 <^* 3 \\ 3 & \\ 5 <^* 2 & 6 <^* 2 \\ 5 <^* 0 & 6 <^* 0 \\ \hline 5 <^* 6 <^* 3 <^* 4 <^* 2 <^* 0 \end{array}$$

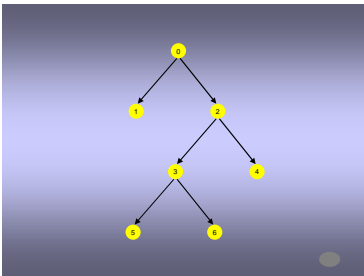


Berechnung der Relation $<^*$

$$\begin{array}{ccc} 1 < 2 & 3 < 4 & 5 < 6 \\ 1 <^* 2 & 3 <^* 4 & 5 <^* 6 \end{array}$$

$$\begin{array}{cc} 5 <^* 3 & 6 <^* 3 \\ 3 & \\ 5 <^* 2 & 6 <^* 2 \\ 5 <^* 0 & 6 <^* 0 \\ \hline 5 <^* 6 <^* 3 <^* 4 <^* 2 <^* 0 \end{array}$$

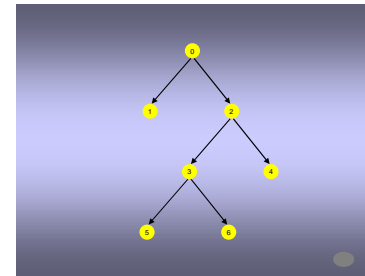
?



Berechnung der Relation $<^*$

$$\begin{array}{ccc} 1 <^* 2 & 3 < 4 & 5 < 6 \\ 1 <^* 2 & 3 <^* 4 & 5 <^* 6 \end{array}$$

$$\begin{array}{cc} 5 <^* 3 & 6 <^* 3 \\ 3 & \\ 5 <^* 2 & 6 <^* 2 \\ 5 <^* 0 & 6 <^* 0 \\ \hline 5 <^* 6 <^* 3 <^* 4 <^* 2 <^* 0 \\ \hline 1 <^* 5 <^* 6 <^* 3 <^* 4 <^* 2 <^* 0 \end{array}$$



Berechnung der Relation $<^*$

$$\begin{array}{ccc} 1 < 2 & 3 < 4 & 5 < 6 \\ 1 <^* 2 & 3 <^* 4 & 5 <^* 6 \end{array}$$

$$\begin{array}{cc} 5 <^* 3 & 6 <^* 3 \\ 3 & \\ 5 <^* 2 & 6 <^* 2 \\ 5 <^* 0 & 6 <^* 0 \\ \hline 5 <^* 6 <^* 3 <^* 4 <^* 2 <^* 0 \\ \hline 1 <^* 5 <^* 6 <^* 3 <^* 4 <^* 2 <^* 0 \end{array}$$

Grundbegriffe zu Ableitungsbäumen

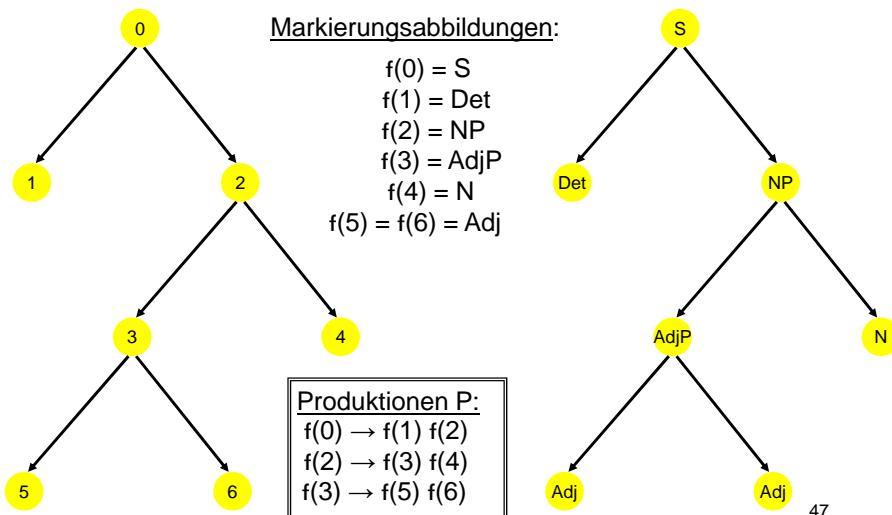
- Ein **vollständig geordneter, markierter Baum** ist ein 4-Tupel $\Gamma = (K, \rho, <^*, f)$, wobei $(K, \rho, <^*)$ ein vollständig geordneter Baum und $f: K \rightarrow M$ eine Abbildung von Knoten K in die Menge M der sog. **Markierungen** (Marken, Etiketten) ist.

45

Grundbegriffe zu Ableitungsbäumen

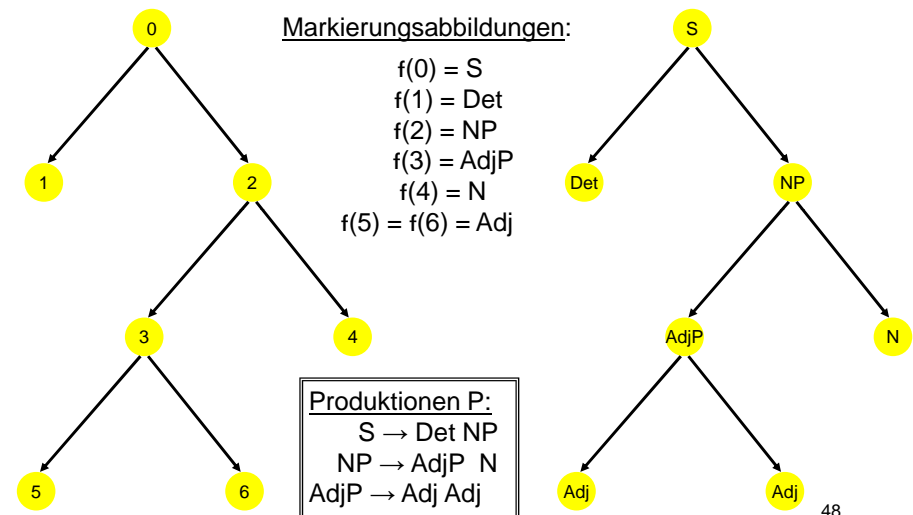
- Sei $G = (N, T, P, S)$ eine kontextfreie Grammatik. Ein vollständig geordneter, markierter Baum $\Gamma = (K, \rho, <^*, f)$ heißt **Ableitungsbaum** zu G , wenn gilt:
 - f bildet die Knotenmenge K auf $M = N \cup T \cup \{\epsilon\}$ ab;
 - für die Wurzel k_w von Γ gilt: $f(k_w) \in N$;
 - Ist $k \in K$ und $\{k_1, k_2, \dots, k_r\}$ die Menge aller seiner direkten Nachfolger mit $k_i < k_{i+1}$ für $1 \leq i < r$, so ist $f(k) \rightarrow f(k_1) f(k_2) \dots f(k_r)$ eine Produktion in P .
- Im Fall $f(k_w) = S$ heißt der **Ableitungsbaum auch Strukturbaum zu G** .

46



47

Strukturbaum

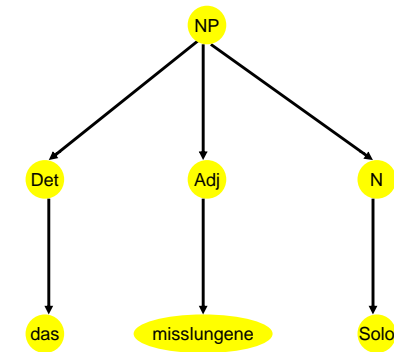


48

Grundbegriffe zu Ableitungsbäumen

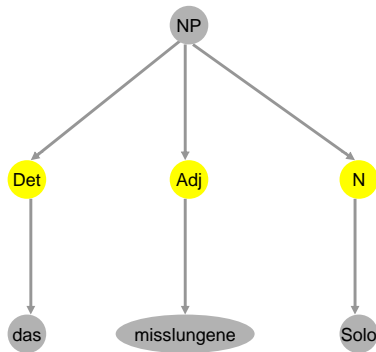
- Sei $\Gamma = (K, \rho)$ ein Baum. Eine Teilmenge $C \subseteq K$ heißt **Schnitt** in Γ , wenn gilt:
 1. es gibt keinen Pfad in Γ , der zwei verschiedene Knoten aus C enthält;
 2. für jede Teilmenge $K^* \subseteq K$, die C echt umfasst (also: $C \subset K^* \subseteq K$), ist Forderung (1) verletzt.

49



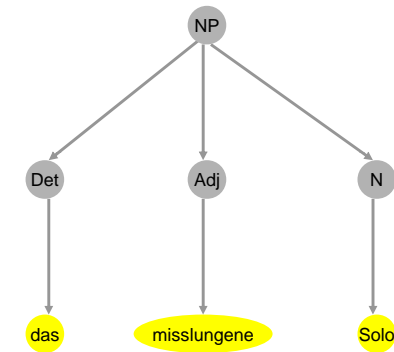
50

Schnitt



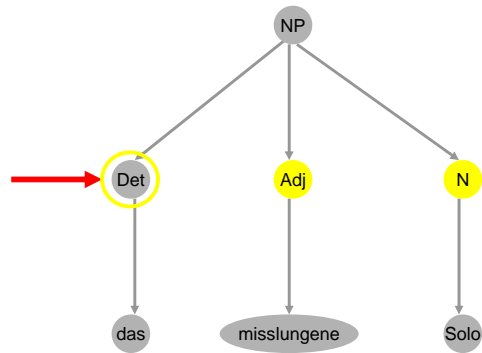
51

Schnitt



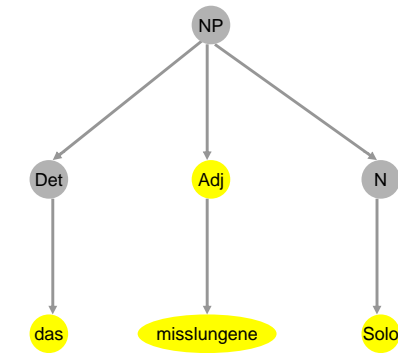
52

Kein Schnitt



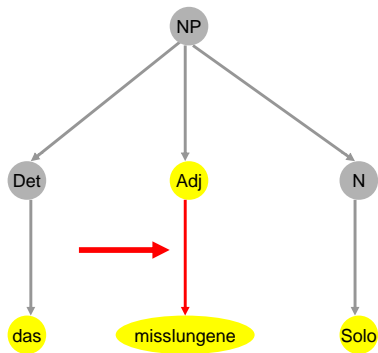
53

Kein Schnitt



54

Kein Schnitt



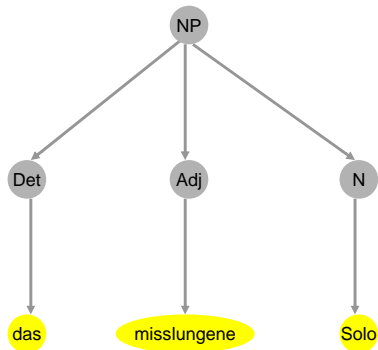
55

Grundbegriffe zu Ableitungsbäumen

- Sei $\Gamma = (K, \rho, <^*, f)$ ein vollständig geordneter, markierter Baum mit $f : K \rightarrow M$. Eine Zeichenfolge $m_1 m_2 \dots m_n \in M^*$ heißt **Schnittbild** in Γ , wenn es einen Schnitt $\{c_1, c_2, \dots, c_n\}$ in Γ gibt mit
 - $c_1 <^* c_2 <^* \dots <^* c_{n-1} <^* c_n$ und
 - $f(c_i) = m_i$ für $1 \leq i \leq n$.
- Falls der zugrundeliegende Schnitt mit der Menge der Endknoten identisch ist, heißt das Schnittbild auch **Endschnittbild**.

56

(End)Schnitt(Bild)



57

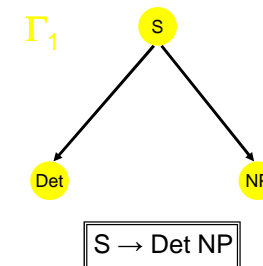
Grundbegriffe zu Ableitungsbäumen

- Ein vollständig geordneter, markierter Baum $\Gamma_n = (K, \rho, <^*, f)$ heißt **Ableitungsbaum** (bzw. **Strukturbaum**) einer **Ableitung** $\Delta = \{\delta_i\}_{i=1}^n$ mit $Q(\Delta) \in N$ (bzw. $Q(\Delta) = S$), wenn es ausgehend vom initialen Baum Γ_0 (bestehend aus dem mit $Q(\Delta)$ markierten Knoten) für jeden Ableitungsschritt $\delta_i, 0 \leq i < n$, eine korrespondierende Erweiterung im Sinne des Ableitungsbaums Γ_i gibt, dessen Endschnittbild $Z(\delta_{i+1})$ ist.

58

Γ_0 **s**

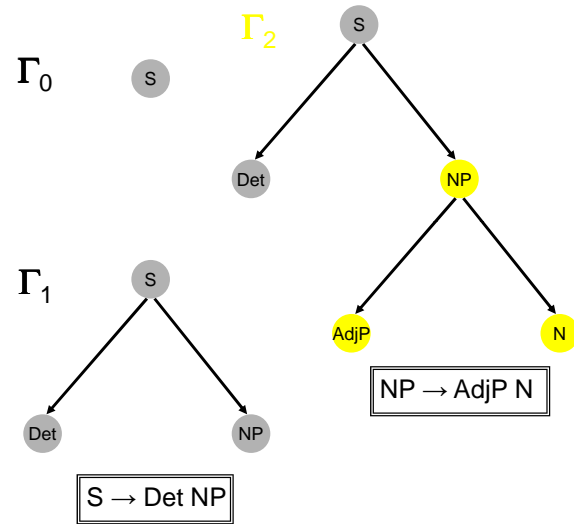
Γ_0 **s**



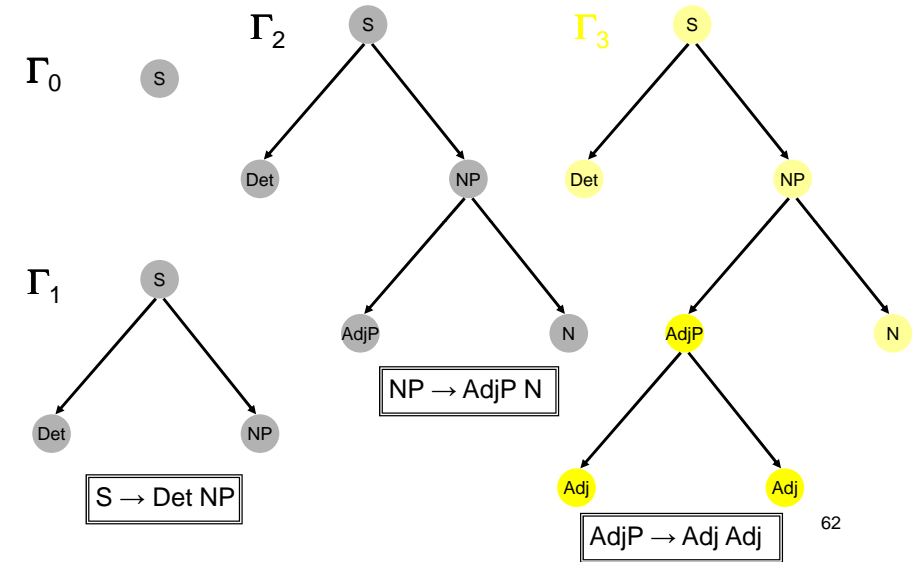
59

60

Strukturbaum einer Ableitung



61



62

Grundbegriffe zu Ableitungsbäumen

- Sei $\Delta = \{ \delta_i \}_{i=1}^n$ eine Ableitung in der Grammatik G mit $Q(\Delta) \in N$. Dann existiert genau ein Ableitungsbaum zu Δ .
- Sei Γ ein Strukturbaum zur Grammatik G , also $Q(\Delta) = S$. Dann existieren im Allgemeinen mehrere verschiedene Ableitungen $\Delta_1 = \{ \delta_{i1} \}_{i1=1}^n$, $\Delta_2 = \{ \delta_{i2} \}_{i2=1}^m$, ... zu Γ .

63

Bemerkung zu Ableitungsbäumen

Bei jedem Erweiterungsschritt im jeweils vorliegenden Anfangsteilbaum Γ_i' bestehen i.A. alternative Möglichkeiten für dessen Expansion, die letztlich zu alternativen Ableitungen führen. Die unterschiedliche Reihenfolge der Anwendung derselben Produktionen führt dennoch zum gleichen Strukturbaum.

Wählt man unter den mit einem Nichtterminalsymbol markierten Endknoten von Γ_i' jeweils den am weitesten links bzw. rechts liegenden, erhält man eine durch den Strukturbaum eindeutig festgelegte Links- bzw. Rechtsableitung.

64

Grundbegriffe zu Ableitungsbäumen

- **Ableitungen**, die denselben Strukturbaum haben, heißen **äquivalent**.
 - Zu jeder Ableitung $\Delta = \{ \delta_i \}_{i=1}^n$ mit $Q(\Delta) \in N$ existiert genau eine äquivalente Links- bzw. Rechtsableitung.
- Jede Klasse von äquivalenten Ableitungen einer Satzform s heißt eine **syntaktische Struktur** von s . Sie kann durch einen Struktur-baum mit Endschnittbild s dargestellt werden.

65

Grundbegriffe zu formalen Grammatiken

- Eine kontextfreie Grammatik G heißt **eindeutig**, wenn jedes Wort $\omega \in \mathcal{L}(G)$ nur eine syntaktische Struktur besitzt.
- Eine Grammatik G , die nicht eindeutig ist, heißt **mehrdeutig (ambig)**.

66

Grundbegriffe zu formalen Grammatiken

- Sei die Grammatik G eindeutig. Dann besitzt jede Satzform s , für die ein $\omega \in T^*$ mit $s \xrightarrow{*} \omega$ existiert, genau eine Rechts- und genau eine Linksableitung.
- Sei die Grammatik G eindeutig und s eine Rechtssatzform mit $s \xrightarrow{*} \omega$ für ein $\omega \in T^*$. Dann ist der Henkel von s und der zugehörige Rechtsreduktionsschritt eindeutig bestimmt.

67

Grundbegriffe zu formalen Sprachen

- Eine kontextfreie **Sprache** \mathcal{L} heißt **eindeutig**, wenn eine eindeutige kontextfreie Grammatik G existiert, die \mathcal{L} erzeugt; sonst heißt \mathcal{L} **mehrdeutig (ambig)**.

Für eine beliebige kontextfreie Grammatik bzw. Sprache ist nicht entscheidbar, ob sie eindeutig ist oder nicht.

68