Mathematische Grundlagen der Computerlinguistik

Bäume

Dozentin: Wiebke Petersen

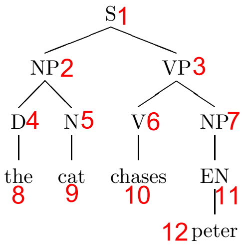
6. Foliensatz
(basierend auf Folien von Gerhard Jäger)

Bäume

Baumdiagramme

Ein Baumdiagramm eines Satzes stellt drei Arten von Information dar:

- die Konstituenten-Struktur des Satzes,
- die grammatische Kategorie jeder Konstituente, sowie
- die lineare Anordnung der Konstituenten.



Bäume

Konventionen

- Ein Baum besteht aus Knoten, die durch
- Kanten verbunden werden.
- Kanten sind implizit von oben nach unten gerichtet (ähnlich zu Hasse-Diagrammen, wo die implizite Richtung aber von unten nach oben ist.)
- Jeder Knoten ist mit einem Etikett (engl. label) versehen.

Bäume

Dominanz

- Ein Knoten x dominiert Knoten y wenn es eine zusammenhängende (möglicherweise leere) Sequenz von abwärts gerichteten Ästen gibt, die mit x beginnt und mit y endet.
- Für einen Baum T bildet

$$D_T = \{\langle x, y\rangle | x \text{ dominiert } y \text{ in } T\}$$

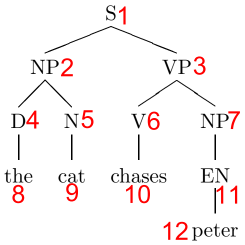
  die zugehörige Dominanz-Relation
- D_T ist eine schwache Ordnung, also reflexiv, transitiv und anti-symmetrisch.

Bäume

Konventionen

- Wenn x nach D_T der unmittelbare Vorgänger von y ist, dann dominiert x y unmittelbar.
- Der unmittelbare Vorgänger von x bzgl. D_T heißt der Mutterknoten von x.
- Die unmittelbaren Nachfolger von x heißen Tochterknoten von x.
- Wenn zwei Knoten nicht identisch sind, aber den selben Mutterknoten haben, heißen sie Schwesterknoten.
- Jeder Baum hat endlich viele Knoten.
- Jeder Baum hat ein Infimum bezüglich der Ordnung D_T. Das Infimum heißt Wurzel oder Wurzelknoten des Baums. Vorsicht: Die Baumdiagramme sind auf den Kopf gestellte Hasse-Diagramme (die Wurzel ist der oberste Knoten des Baumdiagramms, also der Knoten, der als einziges keinen Mutterknoten hat)
- Die maximalen Elemente eines Baumes heißen Blätter (Blätter stehen in einem Baumdiagramm ganz unten. Blätter sind diejenigen Knoten, die keine Töchter haben).

Beispiel



- Knoten 2 dominiert Knoten 8 ($\langle 2,8 \rangle \in D_T$)
- Knoten 2 dominiert Knoten 5 unmittelbar
- Knoten 2 dominiert Knoten 2
- Knoten 2 ist der Mutterknoten von Knoten 5
- Knoten 4 und Knoten 5 sind Schwesterknoten
- Knoten 1 ist der Wurzelknoten des Baums
- Knoten 10 ist ein Blatt des Baums

Bäume

Präzedenz

- Baum-Diagramme beinhalten (anders als Hasse-Diagramm) Informationen über die lineare Abfolge der Knoten.
- Knoten x geht Knoten y voran (engl. x *precedes* y) g.d.w. x links von y steht und keiner der beiden Knoten den anderen dominiert.
- Für einen Baum T bildet

$$P_T = \{\langle x, y\rangle | x \text{ geht } y \text{ voran}\}$$

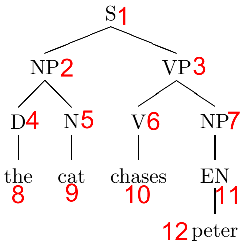
  die zugehörige Präzedenz-Relation.
- P_T ist eine starke Ordnung, also irreflexiv, transitiv und asymmetrisch.

Bäume

Exklusivität

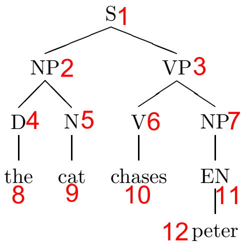
In einem Baum T stehen die Knoten x und y in der Präzedenz-Relation (also $P_t(x, y)$ oder $P_t(y, x)$) g.d.w. sie nicht in der Dominanz-Relation stehen (also weder $D_T(x, y)$ noch $D_T(y, x)$).

Beispiel



- Knoten 7 und Knoten 1 stehen in der
- Knoten 7 und Knoten 2 stehen in der
- Knoten 7 und Knoten 9 stehen in der
- Knoten 7 und Knoten 12 stehen in der
- Knoten 7 und Knoten 10 stehen in der

Beispiel

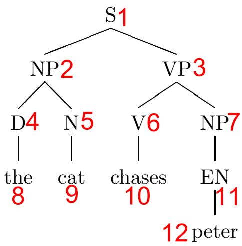


- Knoten 7 und Knoten 1 stehen in der Dominanz-Relation
- Knoten 7 und Knoten 2 stehen in der Präzedenz-Relation
- Knoten 7 und Knoten 9 stehen in der Präzedenz-Relation
- Knoten 7 und Knoten 12 stehen in der Dominanz-Relation
- Knoten 7 und Knoten 10 stehen in der Präzedenz-Relation

Bäume

Nicht-Überkreuzung

Wenn in einem Baum der Knoten x dem Knoten y vorangeht, dann geht jeder Knoten x', der von x dominiert wird, jedem Knoten y' voran, der von y dominiert wird.



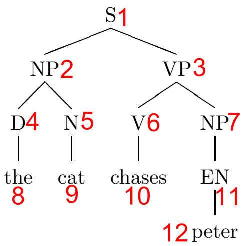
Bäume

Nicht-Überkreuzung

Wenn in einem Baum der Knoten x dem Knoten y vorangeht, dann geht jeder Knoten x', der von x dominiert wird, jedem Knoten y' voran, der von y dominiert wird.

Diese Bedingung schließt aus, dass

- ein Knoten mehrere Mutterknoten hat, oder dass
- sich Äste überkreuzen.

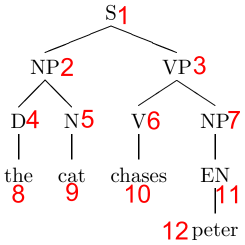


Bäume

Etikettierung

Für jeden Baum T gibt es eine Etikettierungs-Funktion L_T, die jedem Knoten ein Etikett zuweist.

- L_T muss nicht injektiv sein (mehrere Knoten können das selbe Etikett tragen).
- Bei Ableitungsbäumen werden Blätter (auch Terminal-Knoten genannt) auf Terminalsymbole abgebildet und alle anderen Knoten auf Nichtterminal-Symbole.



Bäume

Mit Hilfe dieser Eigenschaften von Bäumen können *Theoreme* bewiesen werden, also Sachverhalte, die für alle Bäume gelten. Zum Beispiel

Theorem 1

Wenn x und y Schwesterknoten sind, dann gilt entweder $P_T(x,y)$ oder $P_T(y,x)$.

Theorem 2

Die Menge der Blätter eines Baumes sind durch P_T total geordnet.

Grammatiken und Bäume

- Bäume repräsentieren die relevanten Aspekte einer Ableitung
- Zusammenhang zwischen Ableitung und Baum am einfachsten, wenn alle Regeln der Grammatik die Form

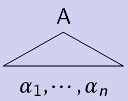
$$A \rightarrow \alpha$$

haben (mit $A \in V_N$ und $\alpha \in (V_T \cup V_N)^*$)

Grammatiken und Bäume

Definition 3

Eine Grammatik $G = (N, T, S, P)$, bei der alle Regeln als linke Seite genau ein Nichtterminal-Symbol haben, generiert einen Baum B genau dann wenn

- die Wurzel von B mit S etikettiert ist,
- die Blätter entweder mit Terminalsymbolen oder it ϵ etikettiert sind, sowie
- es für jeden Teilbaum 

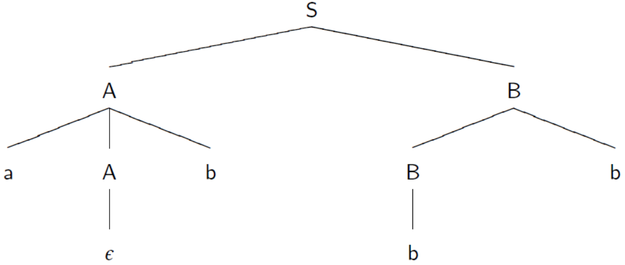
in B eine Regel $A \to \alpha_1, \cdots, \alpha_n$ in P gibt.

Grammatiken und Bäume

Beispiel-Grammatik

$$G = (\{S, A, B\}, \{a, b\}, S, P) \quad P = \left\{ \begin{array}{ll} S \to AB & B \to Bb \\ A \to aAb & B \to b \\ A \to \epsilon & \end{array} \right\}$$

Diese Grammatik generiert z.B. folgenden Baum:



Frage: Welche Sprache wird durch diese Grammatik generiert?