Computerlinguistik I

Vorlesung im WiSe 2018/2019
(M-GSW-09)

Prof. Dr. Udo Hahn

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

http://www.julielab.de

Grundbegriffe zur Syntaxanalyse von CFGs

- Das **Wort-** bzw. **Erkennungsproblem** für eine kontextfreie Grammatik $\mathcal{G}$:

  Zeige für $\mathcal{G}$ = (N, T, P, S) und $\omega \in T^*$, dass ω von $\mathcal{G}$ (nicht) erzeugt werden kann (d.h.: $\omega \in \mathcal{L}(\mathcal{G})$ oder $\omega \notin \mathcal{L}(\mathcal{G})$).

  Ein Algorithmus, der dieses Problem löst, heißt **Erkennungsalgorithmus** (oder **Recognizer**).

Grundbegriffe zur Syntaxanalyse von CFGs

- Das Analyseproblem für eine kontextfreie Grammatik $\mathcal{G}$:

  Bestimme für $\mathcal{G}$ = (N, T, P, S) und $\omega \in T^*$ entweder eine syntaktische Struktur von ω bezüglich $\mathcal{G}$ oder zeige, dass $\omega \notin \mathcal{L}(\mathcal{G})$.

  Ein Algorithmus, der dieses Problem löst, heißt Analysealgorithmus (oder Parser).

  Die Bestimmung der syntaktischen Struktur heißt Syntaxanalyse bzw. Parsing.

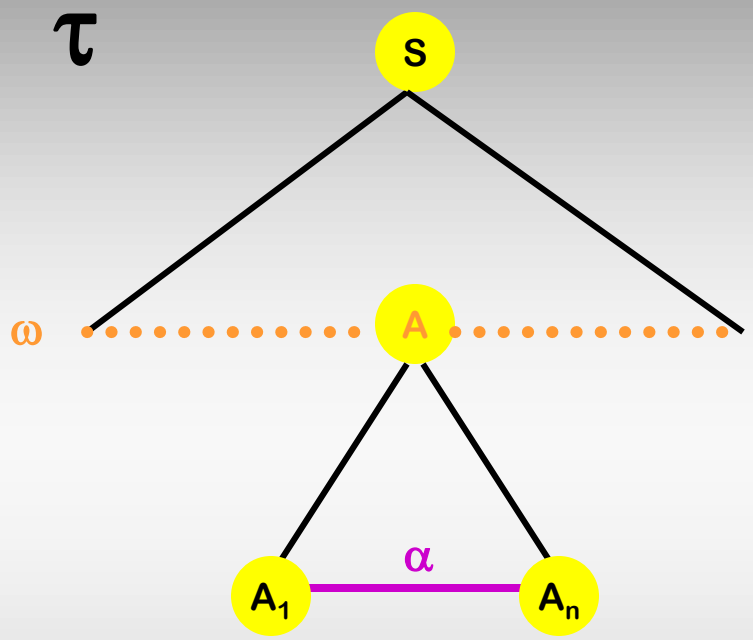
Bemerkungen zur Syntaxanalyse von CFGs

- Ein Analysealgorithmus löst mit der (fehlschlagenden) Bestimmung einer syntaktischen Struktur stets auch das Wortproblem.
- Für Typ-0-Grammatiken ist das Wortproblem unlösbar.
- Für Typ-1-Grammatiken, die bestimmten Beschränkungen unterliegen, und generell für Typ-2-Grammatiken ist das Wortproblem lösbar – wenn auch (für Typ-1) mit z.T. beträchtlicher, aber noch polynomialer Berechnungskomplexität.
- Für Typ-3-Grammatiken ist das Wort- und Analyseproblem einfach lösbar (linear).

Grundbegriffe zur Syntaxanalyse von CFGs

- Sei ω ein von der kontextfreien Grammatik G erzeugtes Wort und τ ein zugehöriger Strukturbaum, der eine feste (beliebig wählbare, aber dann gegebene) Verzweigung besitzt, die aus einem Knoten und seinen direkten Nachfolgern besteht. Diese Verzeigung beschreibt die Anwendung einer Produktion, etwa $A \rightarrow \alpha$ mit $\alpha = A_1 \ldots A_n$ und $A_i \in V$ für $1 \leq i \leq n$.

Gliederung eines Strukturbaums

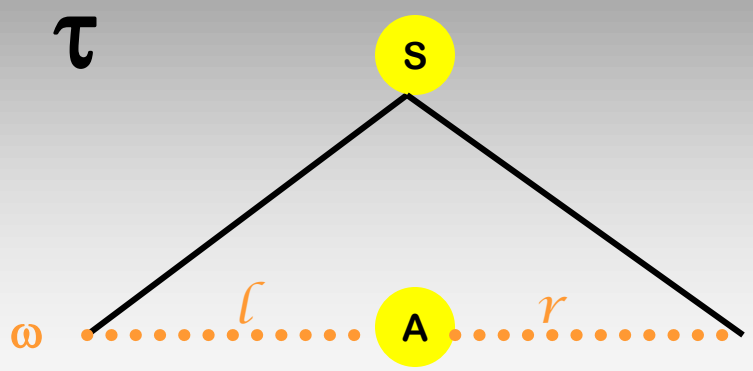


Grundbegriffe zur Syntaxanalyse von CFGs

- Durch die Fixierung einer festen Verzweigung und der zugehörigen Anwendung einer Produktion wird τ in Teilstrukturen zerlegt.

  Dazu betrachten wir die Klasse τ_A aller Strukturbäume zu $\mathcal{G}$, die Anfangsteilbäume von τ sind (d.h. die gleiche Wurzel besitzen) und den fest herausgegriffenen Knoten A als Endknoten haben. Diese haben Endschnittbilder der Form $\ell A r$ mit $\ell, r \in \mathcal{V}^*$ und beschreiben die Ableitung: $S \overset{*}{\Rightarrow} \ell A r$.

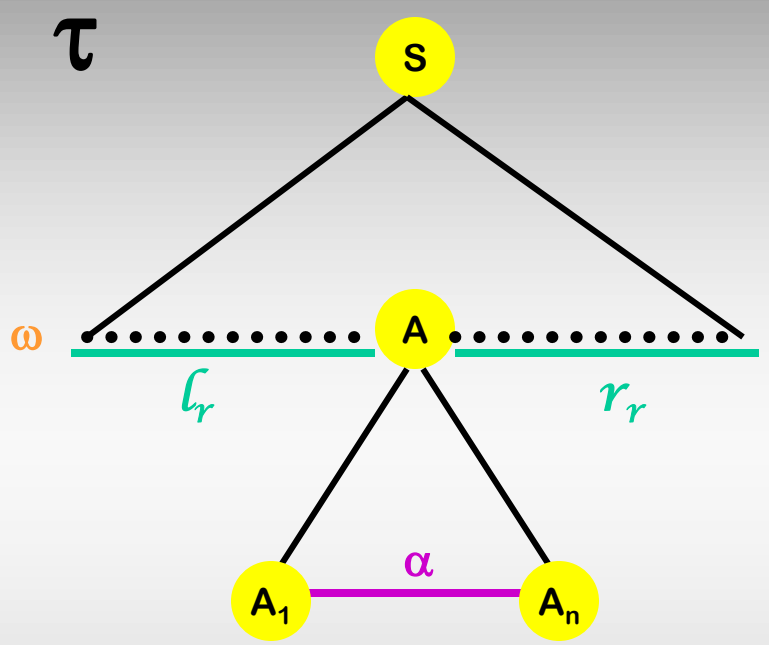
Gliederung eines Strukturbaums



Grundbegriffe zur Syntaxanalyse von CFGs

- Sei τ_{min} der eindeutig fixierte Strukturbaum in τ_A mit minimaler Knotenzahl und $l_r \, A \, r_r$ sein Endschnittbild.
  l_r ist der **reduzierte Linkskontext** und r_r der **reduzierte Rechtskontext** zur betrachteten Anwendung der Produktion $A \rightarrow \alpha$.

Gliederung eines Strukturbaums



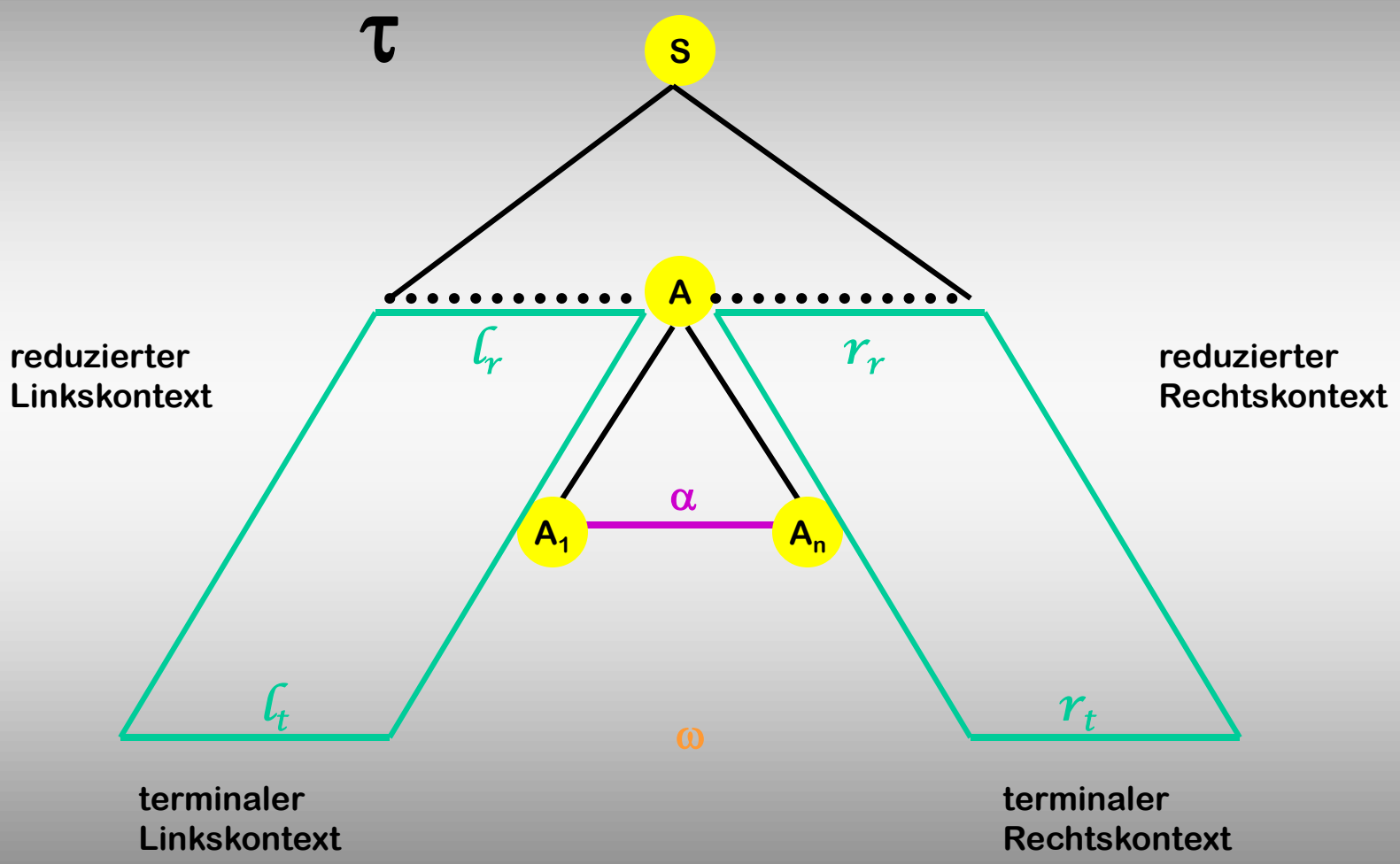
reduzierter Linkskontext

reduzierter Rechtskontext

Grundbegriffe zur Syntaxanalyse von CFGs

- Sei τ_{max} der eindeutig fixierte Strukturbaum in τ_A mit maximaler Knotenzahl und $\ell_t \, A \, r_t$ sein Endschnittbild. Dann sind $\ell_t , r_t \in T^*$.
  ℓ_t heißt dann terminaler Linkskontext und r_t terminaler Rechtskontext zur betrachteten Anwendung der Produktion $A \rightarrow \alpha$.

Gliederung eines Strukturbaums

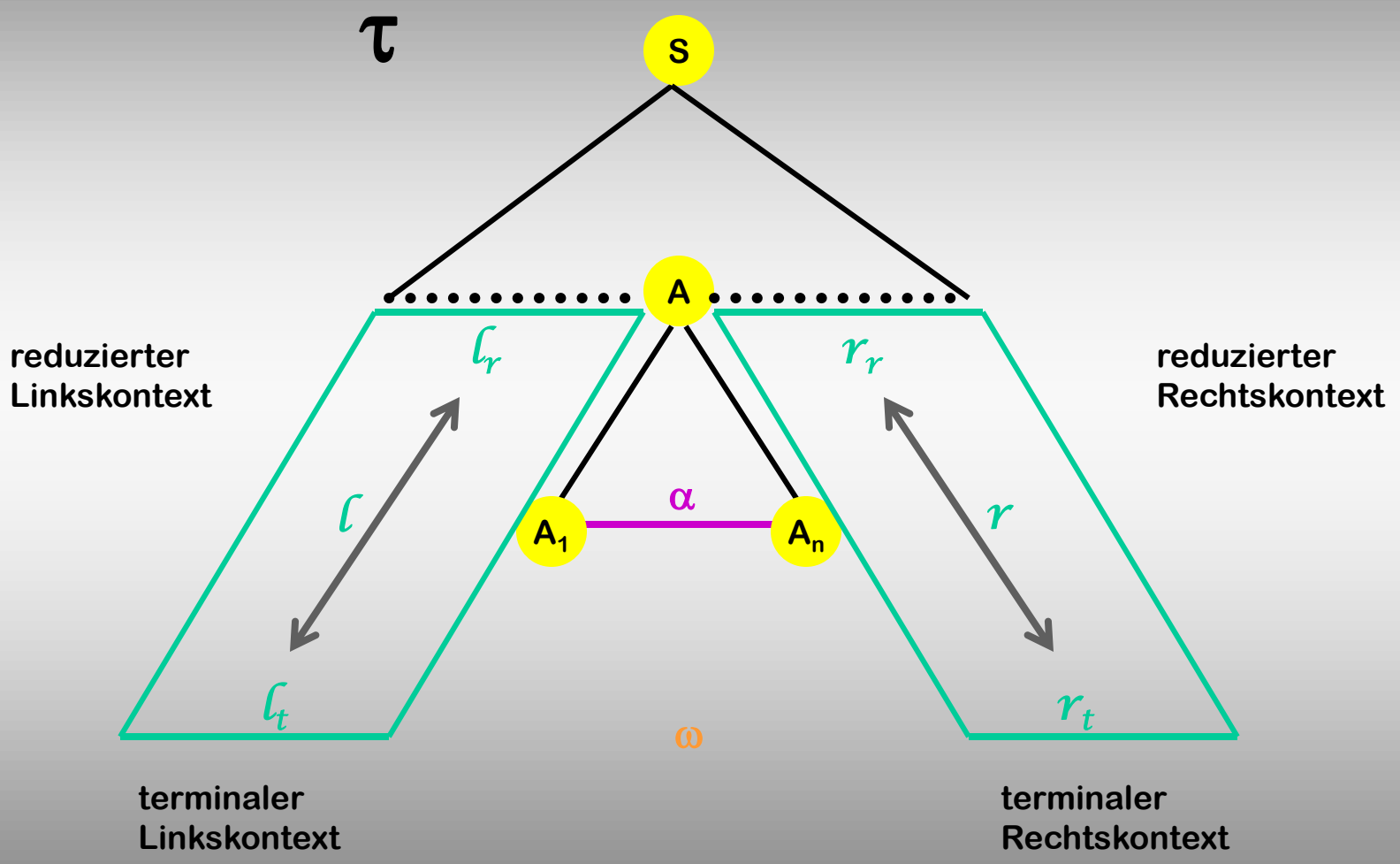


Grundbegriffe zur Syntaxanalyse von CFGs

- Sei τ_{bel} ein beliebig herausgegriffener Strukturbaum in τ_A und sei $\ell \mathrm{A}\, r$ sein Endschnittbild. Dann gilt:

$$\ell_r \overset{*}{\Rightarrow} \ell \overset{*}{\Rightarrow} \ell_t \text{ und } r_r \overset{*}{\Rightarrow} r \overset{*}{\Rightarrow} r_t .$$

Gliederung eines Strukturbaums



Grundbegriffe zur Syntaxanalyse von CFGs

- Die fest herausgegriffene Anwendung der Produktion $A \rightarrow \alpha$ bestimmt in der betrachteten syntaktischen Struktur von ω somit vier Teilstrukturen, die Ableitungen für

  $S \overset{*}{\Rightarrow} \ell_r A r_r$, $\ell_r \overset{*}{\Rightarrow} \ell_t$, $\alpha \overset{*}{\Rightarrow} \alpha_t$ und $r_r \overset{*}{\Rightarrow} r_t$

  mit $\omega = \ell_t \alpha_t r_t$ entsprechen.
- Man nennt diese Teilstrukturen die zur herausgegriffenen Anwendung der Produktion $A \rightarrow \alpha$ gehörige Vorstruktur, Linksstruktur, Nachstruktur und Rechtsstruktur.

Gliederung eines Strukturbaums