Einführung in die Computerlinguistik und Sprachtechnologie

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

Prof. Dr. Udo Hahn

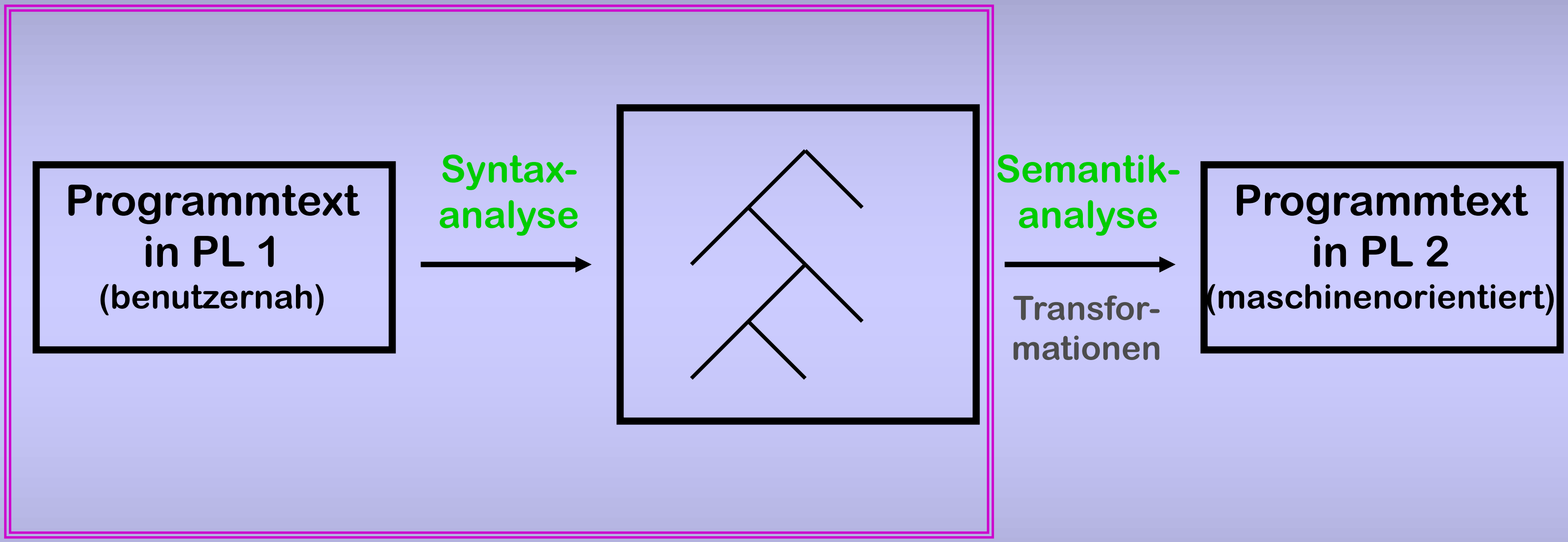
Lehrstuhl für Computerlinguistik
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

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)

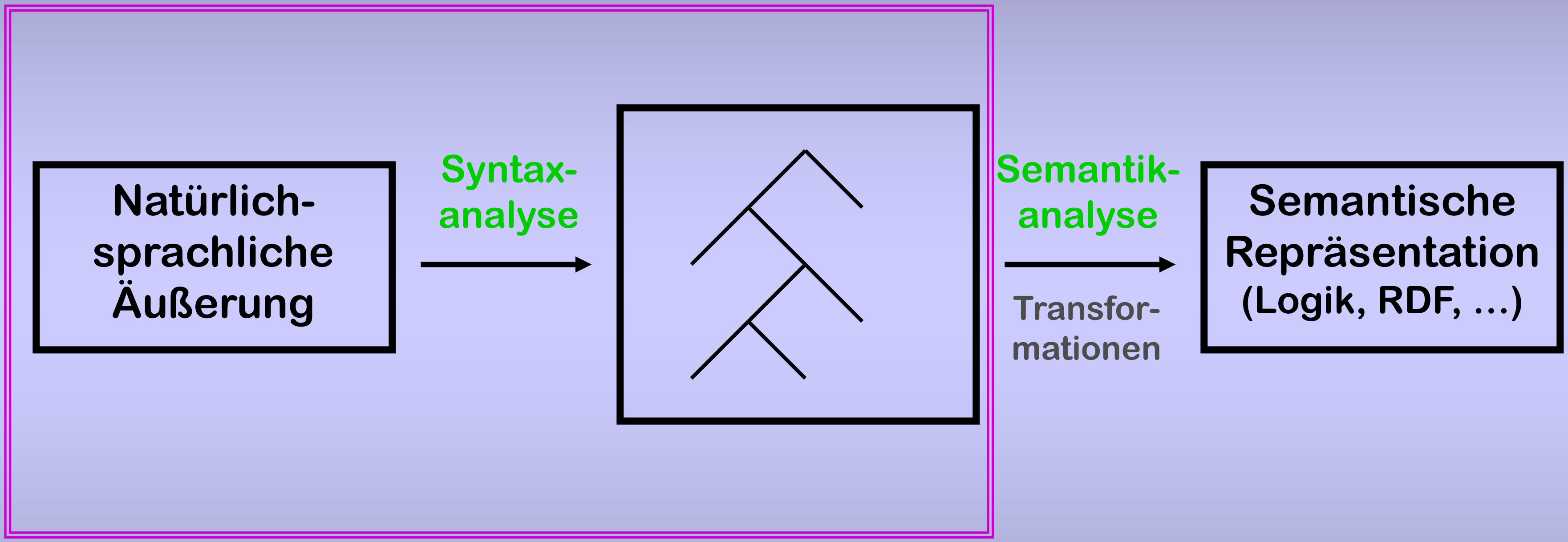
Analyse von Programmen



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* $\cap$ EoR

= { Lauffeuer, Laufmasche }

{ Lauffeuer, Laufmasche } $\cup$ { Lauffeuer, Laufpass }

= { Lauffeuer, Laufmasche, Laufpass }

Mengentheoretische Grundbegriffe

- **Wenn I = {1,…,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 Teil-menge 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