Computerlinguistik I

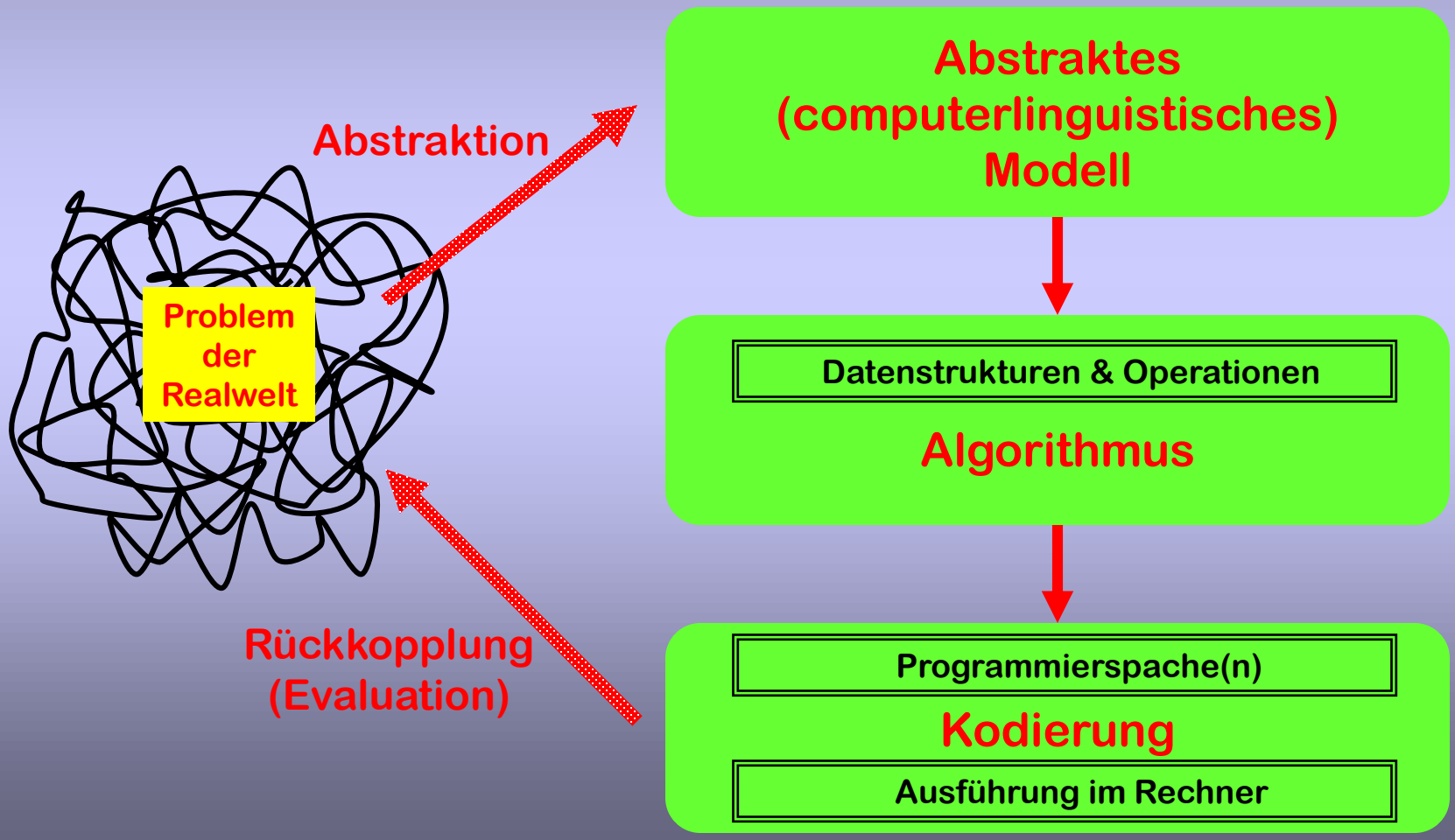
Vorlesung im WiSe 2017/2018
(M-GSW-09)

Prof. Dr. Udo Hahn

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

http://www.julielab.de

Informatischer Problemlösungszyklus



Informatischer Problemlösungszyklus

- **Modellbildung**
  - **Abstraktion von allen unwesentlichen Details der Problemstellung im Hinblick auf die algorithmische Lösung**
  - **Spezifikation der logischen Abhängigkeiten zwischen problemlösungsrelevanten Objekten**
  - **(computer)linguistisches Wissen**

Informatischer Problemlösungszyklus

- Algorithmisierung
  - Übersetzung der modellbezogenen Spezifikation in
    - eine Menge von Objekten (Datenstrukturen) mit bestimmten Eigenschaften und Beziehungen zueinander
    - die erlaubten Operationen auf diesen Objekten
  - Algorithmus: (möglichst präzise) Beschreibung einer Folge zulässiger Operationen auf den Objekten, um das Problem zu lösen
  - Computerlinguistische Kernexpertise

Informatischer Problemlösungszyklus

- Kodierung (Programmierung)
  - Übersetzung der algorithmischen Spezifikation in Konstrukte einer (geeigneten) Programmiersprache
- Ausführung des Programms
  - Hier erst Bezug auf konkrete Maschinen (Datenstrukturen und Algorithmen sind abstrakte Konstruktionen)
  - Test-Modifikationszyklus ... Dokumentation !
  - Informatisches Know-How

Morphologische Prozesse: Flexion - Deflexion

- Kombination von Grundformen mit Flexionsaffixen (Kasus, Numerus, Tempus)
  - Deklination
    - Land: Land, Landes, Lande, Länder, Ländern
  - Konjugation
    - landen: lande, landest, landet, landeten , gelandet
- primär syntaktische, nur minimale semantische Information, keine grundlegenden Wortartwechsel

Morphologische Prozesse: Derivation - Dederivation

- Kombination von Grundformen mit Derivationsaffixen
  - Land: landen, verlanden, anlanden,
  - Land: Landung, Verlandung , Anlandung
  - Land: ländlich, verländlichen, Verländlichung
- modifizierende semantische Information, häufig mit Wortartwechsel verbunden

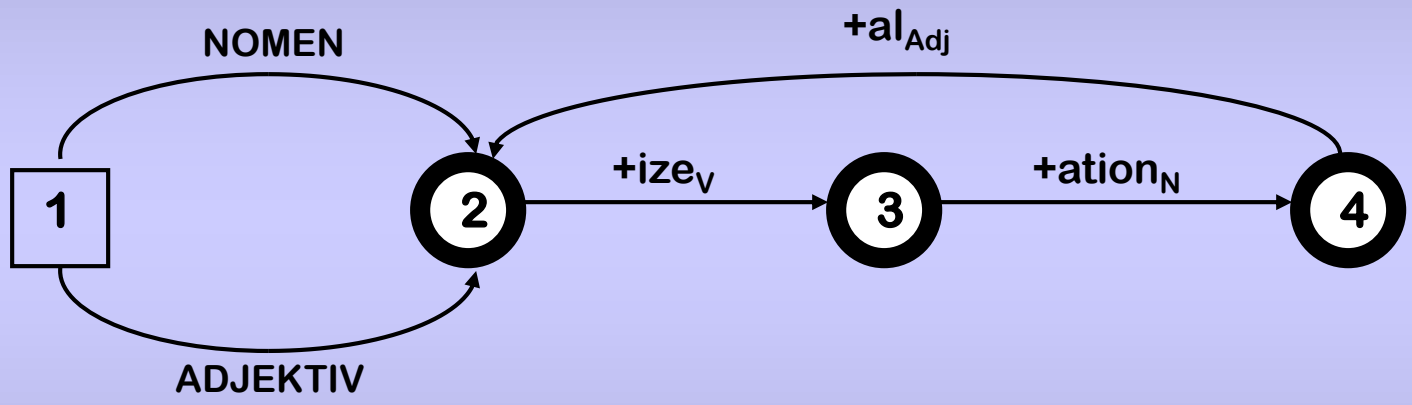
Morphologische Prozesse: Komposition - Dekomposition

- Kombination von Grundformen mit Grundformen (mittels Fugeninfixen)
  - Land: Landnahme, Landflucht, Landgang
  - Land: Heimatland, Ausland, Bauland
  - Land: Landesrekord, Landesverrat, Landsmann
  - Land: Inlandsflug, Landesratspräsidentengattin
- starke semantische Modifikation, fast keine Wortartwechsel
  - … aber: Rotkehlchen, Weichteile

Lemmatisierung vs. Wort-Parsing

Eingabe	Lemma	Wort-Parse
Töchtern	Tochter	Tochter [+N, +FEM, +PL, +DAT]
Hauses	Haus	Haus [+N, +NEU, +SG, +GEN]
sagte	sagen	sagen [+V, +SG, {1P,3P}, +PAST]
Spiegelungen	Spiegelung	$[\text{Spiegel}]_{N}[\text{ung}]_{ds}$ [+N, +FEM, +PL, {NOM,GEN,DAT,AKK}]
leichter	leicht	leicht [+Adj, +POS, +MAS, +SG, +NOM] [+Adj, +KOM]
verlängerte	verlängert	$[\text{ver}]_{dp}[[\text{lang}]_{Adj}[\text{er}]_{ds}]_{Adj}[\text{t}]_{ds}$ [+Part, {MAS,FEM,NEU}, +SG, + DEF, +NOM] [+Part, {FEM,NEU}, +SG, + DEF, +AKK]
	verlängern	$[\text{ver}]_{dp}[[\text{lang}]_{Adj}[\text{er}]_{ds}]_{Adj}[\text{n}]_{ds}$ [+V, +SG, {1P,3P}, +PAST]

Automat für Dederivation



NOMEN: hospital, motor, category, …

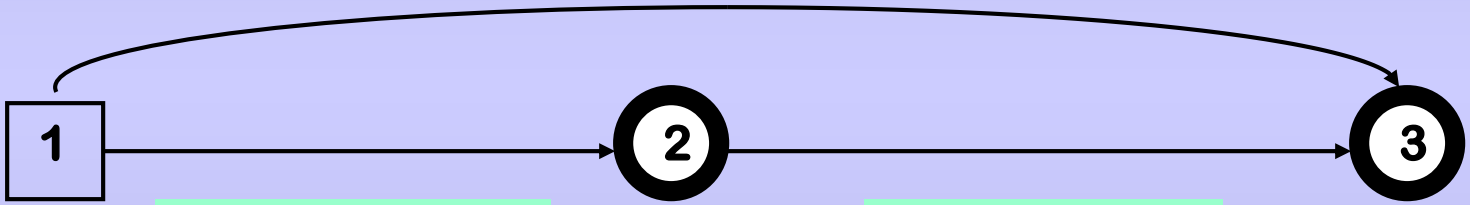
ADJEKTIV: moral, concrete, tender, …

n Anfangszustand

n möglicher Endzustand

Automat für englische Zahlen von 1 bis 99

one, two, three, four, five, six, seven, eight, nine, ten,
eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen



twenty, thirty, forty,
fifty, sixty, seventy,
eighty, ninety

one, two, three,
four, five, six,
seven, eight, nine

Mengentheoretische Grundbegriffe

- Die Zusammenfassung aller Elemente x, die eine Eigenschaft $\mathcal{E}$ haben, wird als **Menge** M bezeichnet:

  $$M := \{x \mid x \text{ hat die Eigenschaft } \mathcal{E}\}$$

  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 | $x \in$ M und $x \in$ N }

  den Durchschnitt

  $M \cup N$:= {x | $x \in$ M oder $x \in$ N }

  die Vereinigung

  von M und N

Mengentheoretische Grundbegriffe

- **Beispiele:**

  LAUF*:= {Laufbahn, laufen, Lauffeuer, Laufmasche, Laufsteg}

  LAUF* ∩ EoR

  = { Lauffeuer, Laufmasche }

  { Lauffeuer, Laufmasche } ∪ { 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

- **Die Menge aller Teilmengen einer Menge M heißt Potenzmenge:**

$$\wp(M) := \{ N \mid N \subseteq M \} = 2^M$$

Beispiel:

LAUFS := { Laufschritt, Laufstall, Laufsteg }

2^{LAUFS} = { Ø, {Laufschritt}, {Laufstall}, {Laufsteg},
{Laufschritt, Laufstall}, {Laufschritt, Laufsteg},
{Laufstall, Laufsteg}, LAUFS }

$| 2^{LAUFS} | = 2^3 = 8$

Mengentheoretische Grundbegriffe

- Das **Kartesische Produkt** von endlich vielen Mengen $M_1,.., M_n$, $n \geq 2$, ist die Menge aller **n-tupel**:

  $M_1 \times M_2 \times ... \times M_n := \{ (m_1,..,m_n) \mid m_i \in M_i , 1 \leq i \leq n \}$

  Beispiel:

  LAUFB := { Laufbahn, Laufbursche }

  LAUFS := { Laufschritt, Laufstall, Laufsteg }

  LAUFB x LAUFS = { (Laufbahn, Laufschritt), (Laufbahn, Laufstall), (Laufbahn, Laufsteg), (Laufbursche, Laufschritt), (Laufbursche, Laufstall), (Laufbursche, Laufsteg) }