

Einführung in die Computerlinguistik – Morphologie (1): morphembasierte Morphologie mit Transduktoren

Dozentin: Wiebke Petersen

14.6.2010

Morphologische Grundbegriffe

Lexem / Wort: abstrakte Einheit, die verschiedenen Formen zugrunde liegt.

Wortform: verschiedene einem Lexem zugrundeliegenden Formen

Paradigma: Menge von Wortformen eines Lexems.

Synkretismus: Zusammenfallen verschiedener Wortformen

Beispiel

Lexem: GEHEN

gehe gehen

gehst geht

geht gehen

...

ging gingen

gingst gingt

ging gingen

...

Wortbildung

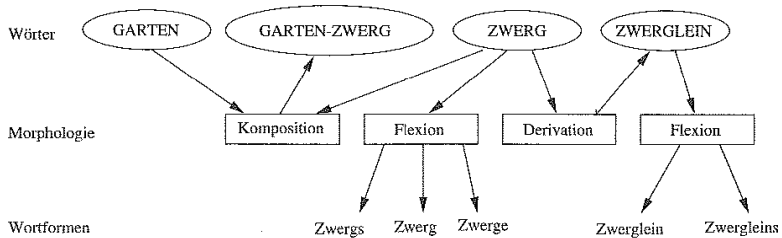


Abbildung 3.4: Flexion und Wortbildung

Komposition: Wort + Wort \mapsto Wort

Derivation: Wort + (gebundenes Morphem) \mapsto Wort

Flexion: Wort + Flexionsmorphem \mapsto Wortform

(Abbildungen aus Carstensen et. al. 2004)

Morpheme

Morphem: Kleinste bedeutungstragende Einheit (genauer, Menge von bedeutungsgleichen Einheiten in komplementärer Verteilung)

Allomorph: Morphe gleicher Bedeutung aber komplementärer Verteilung:
Pluralmorphem $\{-e, -er, -s, \dots\}$ ($-e$, $-er$ und $-s$ sind allomorph)

freies Morphem: frei (ohne weitere Morpheme) äußerbar: Hund, Kind, auf, und, ...

gebundenes Morphem: immer gebunden an andere Morpheme: $-s$, $ver-$, $-lein$, ...

Wurzel(morphem): Ausgangspunkt für Flexion und Derivation (lexikalischer Wortkern): $lauf$, $Kind$, $Hund$, $schwarz$, ...

Affixe: peripher gebundene Morpheme (Suffix, Präfix, Infix, Zirkumfix)

Stamm: Morphemcluster ohne Flexionsaffixe: $vergleich$, $verkauf$, $enthalt$, ...

Regularitäten, Subregularitäten und Ausnahmen

Beispiel: Pluralbildung im Englischen

- regulär**
- door – doors
 - dog – dogs
 - cat – cats

- subregulär**
- fox – foxes
 - ibis – ibises
 - trush – trushes

- Ausnahmen**
- child – children
 - goose – geese
 - mouse – mice

Sprachen mit reicher Morphologie

Türkisch:

uygarlaştıramadıklarımızdanmışsınızcasına

uygar +*laş* +*tır* +*ama* +*dık* +*lar* +*ımız* +*dan* +*mış* +*sınız* +*casına*

civilized +BEC +CAUS +NABL +PART +PL +P1PL +ABL +PAST +2PL +AsIf

“(behaving) as if you are among those whom we could not civilize”

- +BEC “become”
- +CAUS the causative verb marker (‘cause to X’)
- +NABL “not able”
- +PART past participle form
- +P1PL 1st person pl possessive agreement
- +2PL 2nd person pl
- +ABL ablative (from/among) case marker
- +AsIf derivationally forms an adverb from a finite verb

Sprachen mit reicher Morphologie

Walisisch:

LLANFAIRPWLLGWYNGYLLGOGERYCHWYRNDROBWLLLLANTYSILIOGOGOGOCH

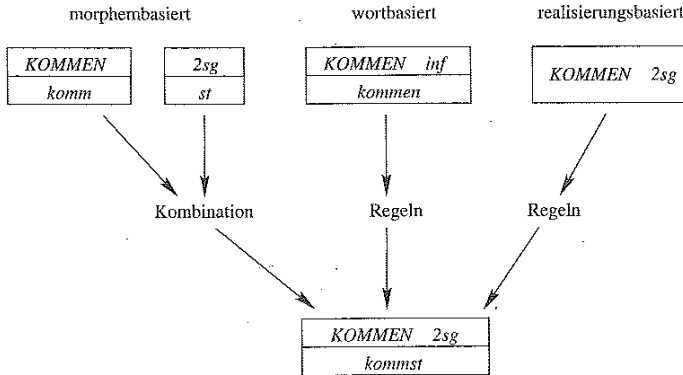
ST MARYS CHURCH IN THE HOLLOW OF THE WHITE HAZEL NEAR TO THE RAPID WHIRLPOOL OF LLANTYSILIO OF THE RED CAVE

St Mary's church in the hollow of the white hazel near to the rapid
whirlpool and the church of St Tysilio of the red cave

Bedeutung der Morphologie in der CL

- Systematische Erfassung der Beziehungen zwischen Wörtern und Wortformen
- Vermeidung von Vollformlexika (Rechtschreibkorrektur, ...)
- Vereinfachung der sprachlichen Verarbeitung (z.B. durch Lemmatisierung)

Grundmodelle der generativen Morphologie



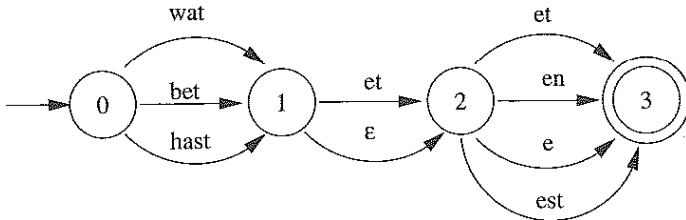
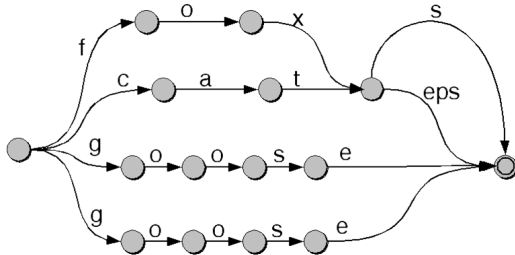
morphemebasiert: Morpheme kombiniert zu Wortformen

wortbasiert: Wortformen werden regelhaft aus Wortformen abgeleitet

realisierungsbasiert: Regeln legen fest, wie Wortform aus Bedeutung und Funktion realisiert werden

Morphologie mit endlichen Automaten

Erkennung von Wortformen



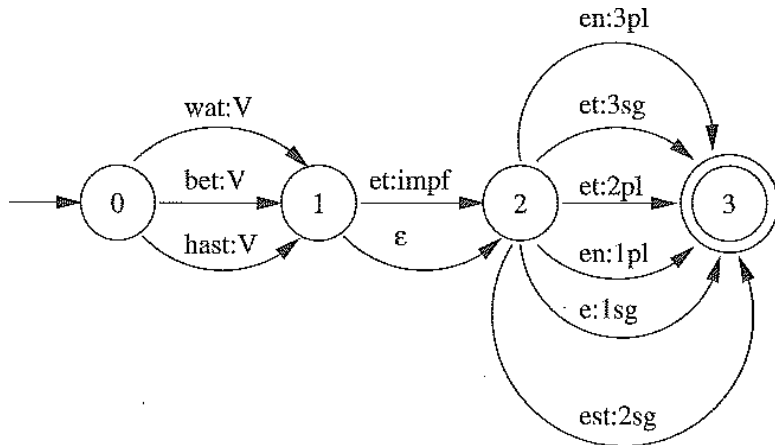
Parsing / Generierung vs. Erkennung

- Mit endlichen Automaten können wir nur Wörter erkennen
- Wir können nicht Parsen
 - Parsen: Aufbau von Struktur
 - Gewöhnlich haben wir einen String einer Sprache und suchen seine Struktur (Parsing)
 - Oder wir haben eine Struktur und möchten die Oberflächenform generieren (Produktion/Generierung)
- Beispiel
 - von “cats” zu “cat +N +PL” (Parsing)
 - von “cat +N +PL” zu “cats” (Generierung)

Lösung: Endliche Transduktoren

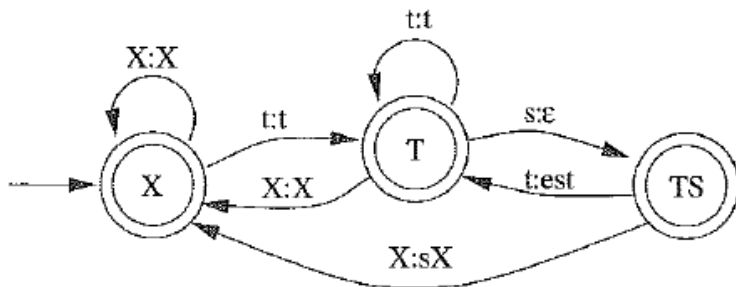
- Füge ein extra Band / Tape hinzu
- Füge extra Symbole zu den Übergängen hinzu
- Auf dem einen Band wird “cats” gelesen, auf das andere wird “cat +N +PL” geschrieben (oder umgekehrt)

Transduktor: Deutsche Verbformen



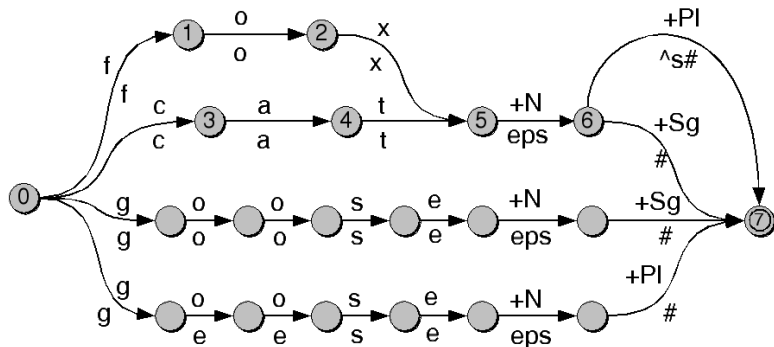
(Carstensen et. al.)

Transduktor: e-Epenthese für deutsche Verbformen



(Carstensen et. al.)

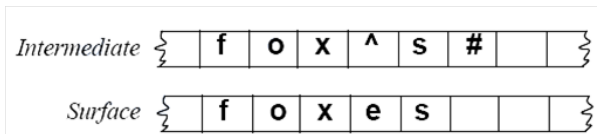
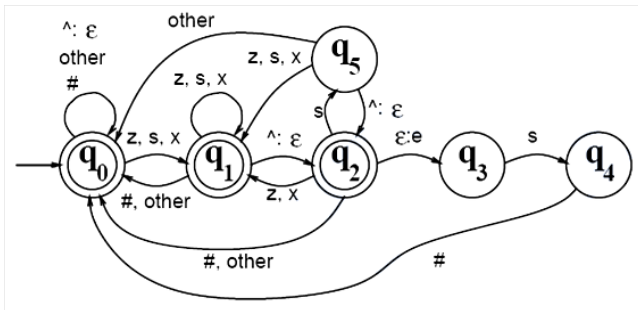
Transduktor: Englischer Plural 1



<i>Lexical</i>	{	f	o	x	+N	+PL	}	
<i>Intermediate</i>	{	f	o	x	^	s	#	}

(Jurafsky & Martin)

Transduktor: Englischer Plural 2 (e-Epenthese)



(Jurafsky & Martin)

Übungsaufgabe

- 1 Erstellen Sie einen Finite State Transducer, der die Steigerungsformen folgender Adjektive erkennt:
schlau – schlauer – schlausten
klein – kleiner – kleinsten
fein – feiner – feinsten
Bei Eingabe von schlauer soll der FST beispielsweise Adj + Komp (für Komparativ) ausgeben, bei Eingabe von feinsten etwa Adj + Sup (für Superlativ).
- 2 Erstellen Sie einen Finite State Transducer, der möglichst viele flektierte Formen des Verbs “lernen” erkennt.
- 3 Erstellen Sie einen Finite State Transducer für die reguläre Flexion der Verben im Englischen.

Hausaufgaben: (Abgabe 21.6.) (BN: Aufgabe 4 und eine der Aufgaben 1-3)

- 1 Erklären Sie anhand der Wörter “fox” und “cat”, wie die Transduktoren von Folie 16 und 17 arbeiten.
- 2 Geben Sie für den Transduktor von Folie 17 die Ableitungen für die Ketten “bus[^] s#” und “hat[^] s#” an.
- 3 Arbeiten Sie **eine** der Aufgaben der vorangegangenen Folie aus.
- 4 Lesen Sie zur Wiederholung bitte das Morphologiekapitel aus Carstensen et. al. (bis einschließlich Kapitel 3.2.4)