Warum braucht man
ein System zur Kompositions- und Derivationsmorphologie?

- Parsen: Wortarterkennung (mit Flexionsinformation) und Argumentstrukturinformation
- Informationsrecherche: Suche anhand morphologischer Bausteine:
  forschung*s+projekt, re+****vital****+isier+ung*
- Text-To-Speech-Systeme:
  - Affixe können Einfluß auf die Betonung haben:
    präsid+**ial**+amt
  - Wortstruktur hat Einfluß auf Betonung:
    *((****hund****e$_N$ +kuchen$_N$)+fabrik$_N$)*

Andere Morphologiesysteme

Es gibt bereits eine Reihe von Wortbildungssystemen

WordManager, Projekt Deutscher Wortschatz, Deutsche Malaga Morphologie, Lingsoft, GerTWOL

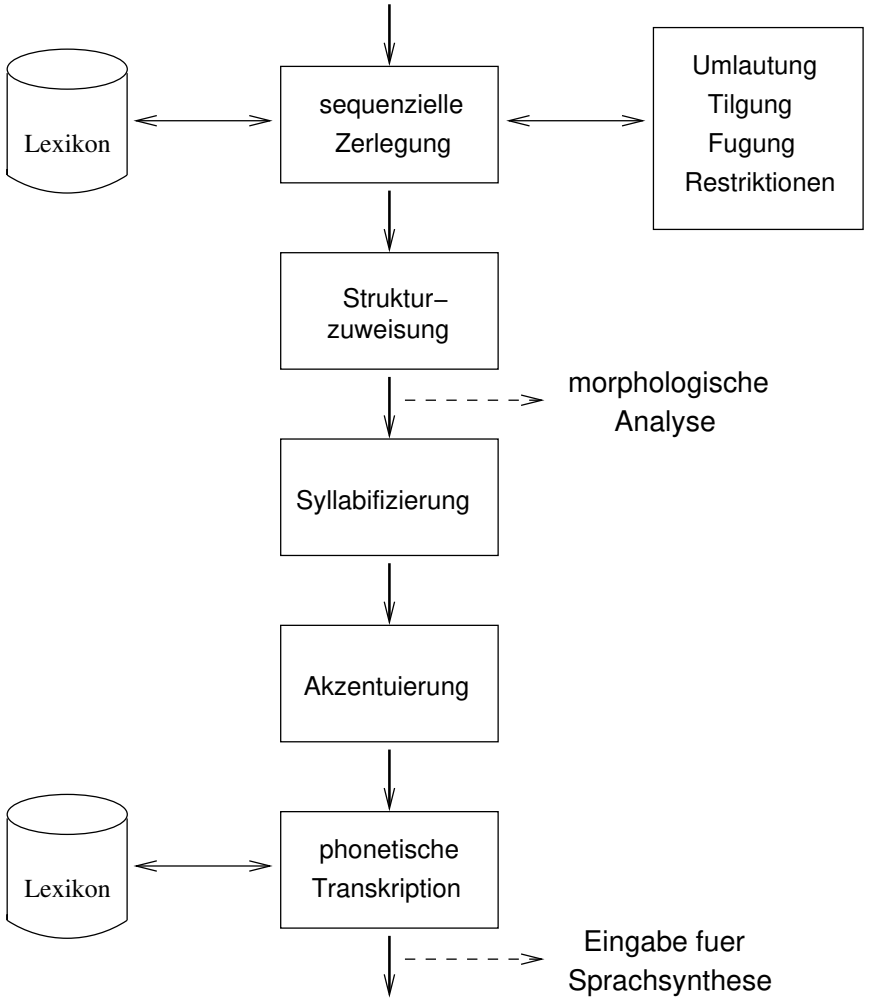
- alle arbeiten mit festen Lexika:
  - manche mit handkorrigierten Daten (WordManager): hohe Qualität
  - manche mit automatisch erzeugten Daten (PDW)
- keines davon gibt Strukturbeschreibungen aus
- man kann nur einzelne Wörter analysieren lassen

Ziele

In jedem Korpus (egal wie groß!) liegt der Median der Lemmahäufigkeiten bei 1.

- man kann mit einem festen Lexikon keine unbekannten Texte verarbeiten
- man braucht ein System, in das man Texte füttern kann
  und das dann unbekannte Wörter “on the fly” analysieren kann
- andererseits möchte man die Qualität verbessern,
  indem man häufige Wörter auflistet
- Analyse (keine Generierung!)

Architektur



Beispiel für die Ausgabe der Komponenten

Eingabe	‘Forschungsprojekt’
Zerlegung	forsch+ungs+projekt
Struktur	Forschungsprojekt$_N$ Forschung$_N$ Projekt$_N$ forsch$_V$ ung$_N$
Betonung	’fo:6$SUNs$pRo:$j@kt

Theoretische Grundannahmen von DeKo

Stämme vs. Affixe

- Es gibt keine grundsätzliche Unterscheidung zwischen Affixen und Stämmen,
  stattdessen hat jedes Element im Lexikon die Attribute
  MORPHOLOGISCHER STATUS: frei/gebunden
  SELEKTION: ja/nein

- Wir arbeiten mit striktem *Item and Arrangement*,
  d.h., wir nehmen an, daß jedes Morphem
  eine Kompositionsstammform und eine Derivationsstammform hat.

Arten komplexer Wörter

- Analyse auf Basis von
  - morphologischen Kriterien
  - semantischen Kriterien
  - Produktivität
- Klassifizierung
  - regelmäßig: morphologisch und semantisch regelmäßig, produktiv
  - semi-regelmäßig: morphologisch regelmäßig,
    aber semantisch nicht kompositionell
    oder regelmäßig aber nicht produktiv
    kränklich, wöchentlich, erwürgen

  Außerdem gibt es Wörter, die (synchron) nicht als komplex wahrgenommen werden: Idiome, werden als Simplexe behandelt
  freilich, Deichsel

Theoretische Grundannahmen und Datenakquisition

Zu erfassende Daten

(A) Derivationsstammmformen und Kompositionsstammformen
für alle freien und gebundenen Nomina, Adjektive und Verben:

	Bauer	-keit
DS	bäuer	keit
KS	bauern	keits

(B) Selektionseigenschaften aller selegierenden Elemente, z.B.

- Wortart

  -bar nur mit Verben: *beweisbar, *tischbar, *grünbar*

- Argumentstruktur

  -bar mit transitiven Verben: *lesbar, *schlafbar*

- Herkunft (fremd oder nativ)

  -abel nur mit klassischen Basen: *akzeptabel, *annehmabel*

- morphologische Beschaffenheit (Abkürzung oder Vollwort)

  -ler mit Abkürzungen: *CDUler*

  -lich nicht mit Abkürzungen: **CDUlich*

- Semantik

  -fach nur mit Kardinalzahlen: *dreifach, *tischfach*

Datenakquisition

Vorgehensweise für die regelmäßige Wortbildung:

- Selegierende Elemente:
  tabellarische Erfassung der Selektionseigenschaften
- Kompositionsmuster:
  tabellarische Erfassung der Eigenschaften
- Modellierung in Lexikon und Regelapparat

Zusätzlich:

- Listen für die semi-regelmäßigen komplexen Wörter und die Idiome
- Stammformen werden halbautomatisch aus Beleglisten akquiriert

Beispieltabelle Derivation

	Affix	erzeugt WBP mit der Kategorie	Vorkommen in 200M Wortformen: – Typen – Tokens – Hapax Legomena	Produktivitäts-rate	nativ oder fremd	semantische Funktionen
AFFIX	-lich	Adjektive	– 873 – 1473983 – 169	0,00011	nativ	relativ, ornativ, . . .

Beispieltabelle Derivation (Fortsetzung)

	Muster	Wortart, morpho-syntaktische Merkmale	konzeptuelle Spezifikation	klassische Basen?	englische Basen?	Abkürzungen?	morphologisch komplexe Basen?
BASIS	*ärztlich, amts-ärztlich, richterlich*	NN, masc	Berufsbezeichnung	nein	nein	nein	ja, Nominalkomposita

Beispieltabelle Derivation (Fortsetzung)

	grammatische Spezifikation	konzeptuelle Spezifikation	Konnotation	Fuge?	Tilgung?	Akzent-wechsel?	Umlaut?
WORT-BILDUNGS-PRODUKT	adverbial, attributiv, prädikativ	Eigenschaft, relativ	neutral	nein	nein	nein	ja

Beispieltabelle Komposition (Ausschnitt)

Muster	Spezifikation	Beispiele	Semantik des Wort-bildungsprodukts	rekursiv?	produktiv?	Wortakzent
Adj + Adj $\rightarrow$ Adj	Simplex + Simplex	hellblau, frühweise	Determinativ-kompositum	nein	ja	Nichtkopf
	Derivat + Simplex	giftiggelb	Determinativ-kompositum	nein	ja	Nichtkopf
	Simplex + Derivat	tiefgläubig	Determinativ-kompositum	nein	ja	Nichtkopf
	Simplex + Partizip	bunt-gestrichen	Determinativ-kompositum	nein	ja	Nichtkopf
	Simplex + Simplex	grüngelb, sauer-scharf	Kopulativ-kompositum	ja	ja	Nichtkopf und Kopf

Lexikon: Morphologisch relevante Informationen

FRAU

KATEGORIE: nomen

FORM: simplex

FLEXIONSSTAMMFORM: frau

DERIVATIONSSTAMMFORM [+UMLAUT]: fräu

DERIVATIONSSTAMMFORM [-UMLAUT]: frau

KOMPOSITIONSSTAMMFORM: frauen

MORPHOLOGISCHER STATUS: frei

SELEKTION: nein

...

-SCHAFT

KATEGORIE: nomen

FORM: simplex

FLEXIONSSTAMMFORM: schaft

DERIVATIONSSTAMMFORM: schaft

KOMPOSITIONSSTAMMFORM: schafts

STATUS: gebunden

SELEKTION: ja, [an_kompos][an_plural][an_nom]

...

Regeln und Lexikon

- Mittelfristige Zielsetzung
  - häufige regelmäßige Wörter:
    einmal mit DeKo analysiert, handkorrigiert, mit Analyse im Lexikon
  - seltenere regelmäßige Wörter: Derivations- und Kompositionsstammformen
    im Lexikon, Analyse durch Regelapparat “on the fly”
  - semi-regelmäßige Wörter: als komplexe Wörter im Lexikon
  - Lexikon-Lookup plus regelbasierte Analyse auf Wunsch
- Derzeitiger Stand
  - Es gibt nicht genug Derivations- und Kompositionsstammformen im Lexikon
  - Alternative: Fugungs- und Tilgungsregeln

Die Ressourcen von DeKo

- Deskriptive Arbeiten
  - 308 Tabellen für selegierende Elemente
  - 20 Tabellen für Kompositionsmuster
  - Listen mit semi-regelmäßigen Elementen zu allen Tabellen (ca. 1000)
  - Derivationsstämme: noch keine (Akquisition läuft an)
  - Kompositionsstämme: ca. 400 Verben, ca. 5400 Nomina
- Lexikonmodell
  - neues integriertes IMSLex
    beruht auf Anforderungsanalyse und Modellierung in DeKo
  - Beschreibung der morphologischen Eigenschaften
    aller selegierenden Elemente und vieler Stämme

Implementierung

Unterschiedliche Anforderungen und Behandlung durch verschiedene Formalismen:

- Segmentierung komplexer Wörter durch deklarative Grammatik
  (gibt Kombinierbarkeit morphologischer Einheiten an)
- hierarchische Strukturinformation durch kontextfreie Grammatik
- phonologische Prozesse durch kontextsensitive Rewrite-Regeln

$\rightarrow$ Gemeinsamkeit: Transformation einer Symbolfolge in eine andere

$\rightarrow$ flexibles Modell zur Konvertierung von Symbolketten:
Technologie endlicher Automaten

$\rightarrow$ in DeKo werden FST-Softwarepakete der AT&T Research Labs verwendet[a]

[a] Sproat (2000)

Stand und Ausblick

Ergebnisse nach 15 Monaten Projektlaufzeit

- Modell für Derivations- und Kompositionsanalyse,
  das auch unbekannte Wörter zerlegen kann (Lexikon + Regeln)
  - theoretisch relevant
  - modular implementiert und nach FST kompiliert
- Komponenten liegen vor (alle nativen Wortbildungsmuster des Deutschen)
- Lexika liegen zu einem kleinen Teil vor
- Testtool liegt vor

Ergänzungsbedarf (06/01 – 12/01)

- Anbindung an die Flexionsmorphologie (dmor)
- Lexikonerweiterung (Kompositions- und Derivationsstammformen)
- Tests und Verbesserungen
- Erkennung von komplexen Wörtern mit Bausteinen, die nicht im Lexikon sind

Literatur

Sproat, R. (2000). Lextools: a toolkit for finite-state linguistic analysis. Technical report.
`http://www.research.att.com/sw/tools/lextools/`.