Computerlexikographie-Tutorium 04.07.2008

- Themen für heute:
  - GERTWOL und die Zwei-Ebenen-Morphologie
  - (Stemming)
  - Wiederholung II

GERTWOL

- Automatische Wortformerkennung für das Deutsche
- Link mit Demo: http://www2.lingsoft.fi/cgi-bin/gertwol
  - (begrenzte Verwendbarkeit: 20 (oder 100) Versuche pro Tag)
- Name:
  - **GER**man Morphological Analyser
  - Two-Level Model for Morphology: **TWOL** (Zwei-Ebenen-Modell)
    - sprachunabhängige Analysemethode
    - 1983, Kimmo Matti Koskenniemi
- Lexikon: ~ 85000 Wortformen
  - The Collins German Dictionary, 1991 (Neubearbeitung) +
  - unterschiedliche Korpora (verschiedene Textsorten und Themen → ca. 98 % Abdeckung der Daten in einem allgemeinen Text)
- Erkennung der Flexion, Derivation und Komposition, bzw. Konvertierung von Infinitiven, Adjektiven und Partizipien zu Substantiven

GERTWOL

- Eingabe: eine deutsche Wortform (z.B.: *Forschungsabteilungen*)
- Ausgabe: alle formal möglichen Lesungen der Wortform:
  - **"<*forschungsabteilungen>"**

```
"*forsch~ung\s#abt~ei#lunge"  S FEM PL NOM
"*forsch~ung\s#abt~ei#lunge"  S FEM PL AKK
"*forsch~ung\s#abt~ei#lunge"  S FEM PL DAT
"*forsch~ung\s#abt~ei#lunge"  S FEM PL GEN
"*forsch~ung\s#ab|teil~ung"  S FEM PL NOM
"*forsch~ung\s#ab|teil~ung"  S FEM PL AKK
"*forsch~ung\s#ab|teil~ung"  S FEM PL DAT
"*forsch~ung\s#ab|teil~ung"  S FEM PL GEN
```

  - **Stamm** der eingegebenen Wortform mit Kennzeichnung der ermittelten **Morpheme** und
  - **morphologische Informationen** zum Stamm: POS, Genus, Numerus, Kasus, (Stil), (+ zusätzliche wortartenspezifische Angaben)

GERTWOL

- Eingabe: eine deutsche Wortform (z.B.: *Forschungsabteilungen*)
- Ausgabe: alle formal möglichen Lesungen der Wortform:
  - **"<*forschungsabteilungen>"**

```
"*forsch~ung\s#abt~ei#lunge"   S FEM PL NOM
"*forsch~ung\s#abt~ei#lunge"   S FEM PL AKK
"*forsch~ung\s#abt~ei#lunge"   S FEM PL DAT
"*forsch~ung\s#abt~ei#lunge"   S FEM PL GEN
"*forsch~ung\s#ab|teil~ung"    S FEM PL NOM
"*forsch~ung\s#ab|teil~ung"    S FEM PL AKK
"*forsch~ung\s#ab|teil~ung"    S FEM PL DAT
"*forsch~ung\s#ab|teil~ung"    S FEM PL GEN
```

  - **Stamm** der eingegebenen Wortform mit Kennzeichnung der ermittelten **Morpheme** und
  - **morphologische Informationen** zum Stamm: POS, Genus, Numerus, Kasus, (Stil), (+ zusätzliche wortartenspezifische Angaben)

GERTWOL

- Legende:
  - **`*`** = Großschreibung
  - **`#`** = Komposition (strong boundary)

    `*abend#zeit`
  - **`|`** = Präfigierung (auch für Konfixe) (weak boundary)

    `*vor|schule`

    `*biblio|thek`
  - **`~`** = Suffigierung (suffix)

    `*lehr~er~in`

    `brauch~bar`
  - **`\`** = Fugenelement (linking element)

    `zu|griff\s|bereit`

    `*held\en#tat`

GERTWOL

- Erkennung von Portmonteau-Morphemen: **"<zur>"**

```
"zu-die"  PRÄP ART DEF SG DAT FEM
```

- Erkennung von Eigennamen: **"<*goethe>"**

```
"*goethe"  S EIGEN Famname SG NOM
"*goethe"  S EIGEN Famname SG AKK
"*goethe"  S EIGEN Famname SG DAT
"*goethe"  S EIGEN Famname SG GEN
```

- Stilmarkierung: **"<*hause>"**

```
"*haus"  S NEUTR SELTEN SG DAT
"haus~en"  * V IMP PRÄS SG2
"haus~en"  * V IND PRÄS SG1
"haus~en"  * V KONJ PRÄS SG1
"haus~en"  * V KONJ PRÄS SG3
```

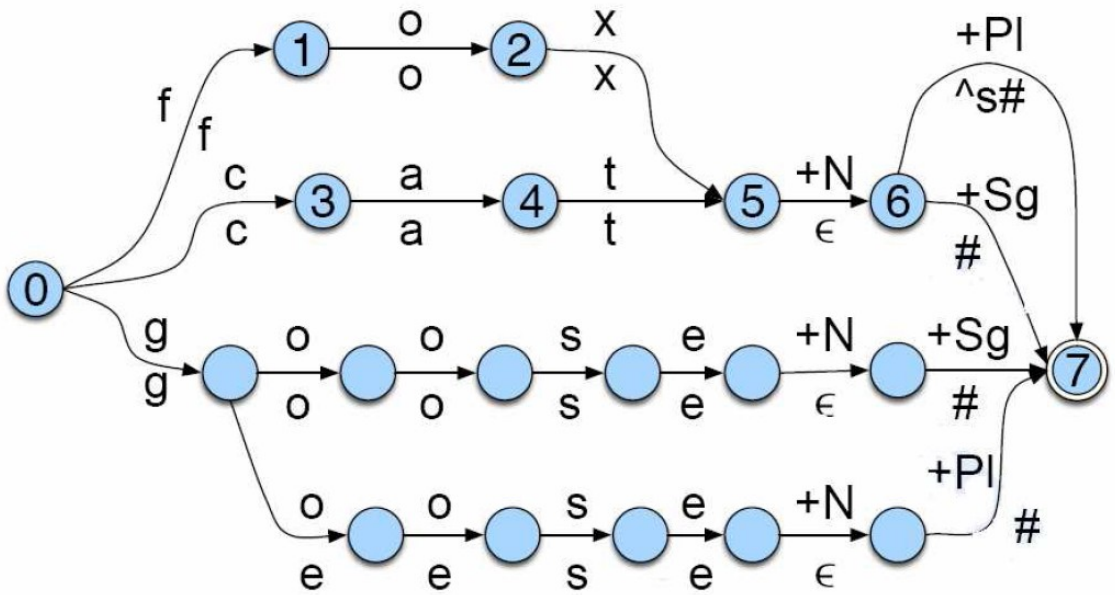
Zwei-Ebenen-Morphologie

- Kimmo Koskenniemi, Dissertation (1983): *Two-Level Morphology: A General Computation Model for Word-Form Recognition and Procuction.*
  - Link: http://www.ling.helsinki.fi/~koskenni/doc/Two-LevelMorphology.pdf
- Lexika:
  - Stammlexika für alle Wortarten (stem formation)
  - Teillexika für Affixe und Flexive (morphotactic structure)
- Regeln definieren phonologische oder morphologische Kontexte für Ersetzung erkannter Einheiten durch eine kodierte Repräsentation dieser Einheit.
  - Die Regeln werden in einem zweibändigen FST (= finite state transducer) implementiert.
- Durch den Transduktoren können die Wortformen analysiert oder generiert werden.

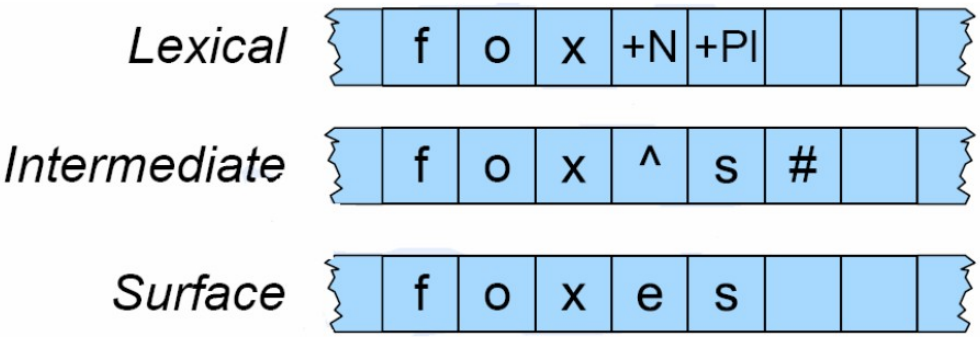
(Nächste Seiten: Frank (2007): S.18, 20, 21)

Von FSAs zu FSTs für morphologische Analyse

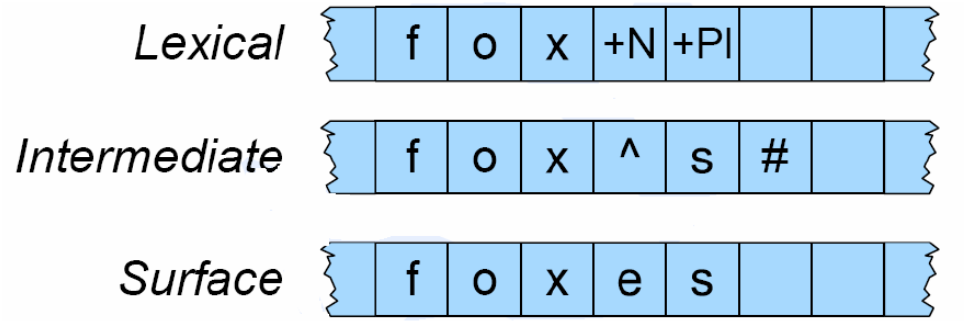
- Resultat



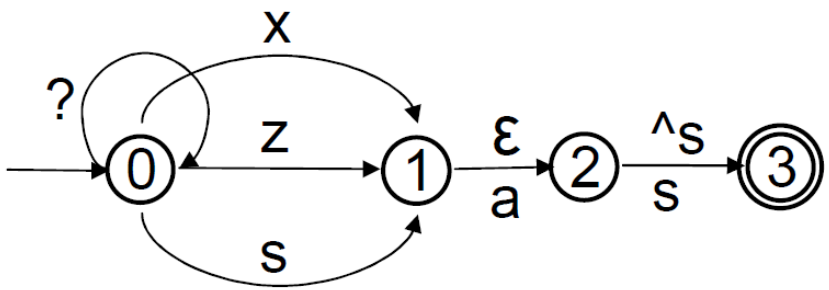
- c:c a:a t:t +N:ε +Pl:^s# ==> cat^s ==> cats
- f:f o:o x:x +N: ε +Pl:^s# ==> fox^s ==> foxes



Transducer für orthographische Regeln



- Komplexe kontextuelle Bedingungen für Ersetzungsregeln:
  z.B. „füge an der Oberfläche ein „e“ ein, wenn ein Morphem mit „x, z oder s“ endet und das nächste Morphem ein „s“ ist



$$\varepsilon \rightarrow e \,/\, \left\{ \begin{matrix} x \\ z \\ s \end{matrix} \right\} \,\hat{}\, \_\_\; s\#$$

- *a* → b/c ___ d : „ersetze *a* durch *b* wenn es zwischen *c* und *d* steht“

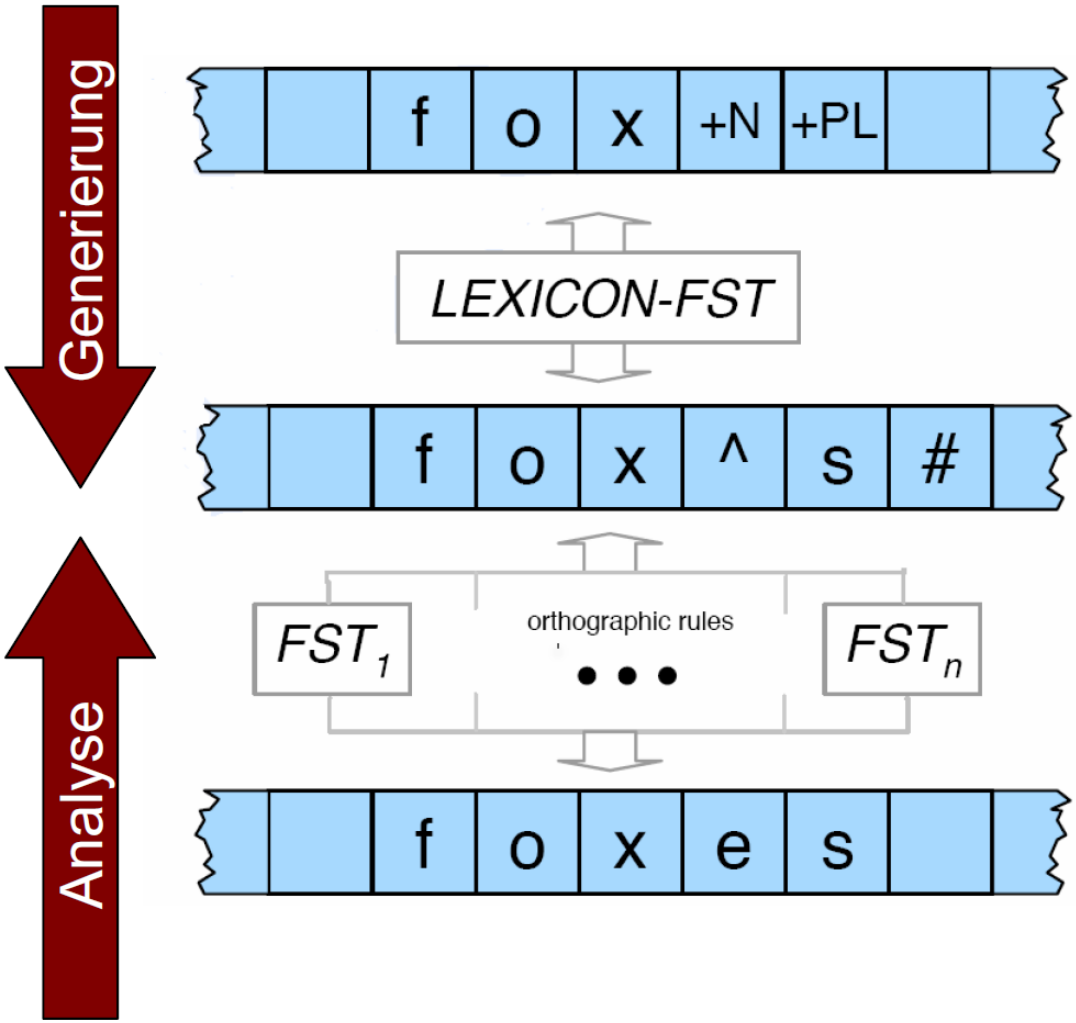
Architektur für morphologische Analyse mit FSTs

Kaskade von Transduktoren

- Lexikalischer Transducer
- Orthographische u. morpho-phonologische Regeln
  - Parallel oder in Kaskade

Komposition von FSTs

- Transformation zu einem einzigen, komplexen FST



Zwei-Ebenen-Morphologie: Beispiele

- Finnische Beispiele: (flektierte Wortformen)

'Dach'

- Eingabe: **`katto`** — **Oberflächenstruktur** (= OS)
  Zwischenebene: **`katTo`** — lexikalische Repräsentation
  Ausgabe: **`"katto" N NOM SG`** — lexikalische Information (**Tiefenstruktur** = TS)

'auf dem Dach'

- Eingabe: **`katolla`**
  Zwischenebene: **`katTo$+llA`**
  Ausgabe: **`"katto" N ADE SG`**

Zwei-Ebenen-Morphologie: Beispiele

- Deutsche Beispiele: (flektierte Wortformen)
  - Eingabe: **`Väter`** Oberflächenstruktur (OS)

    Zwischenebene: **`*vaterI`** lexikalische Repräsentation

    Ausgabe: **`"*vater" S MASK PL NOM`** lexikalische Information (TS)

    `"*vater" S MASK PL AKK`

    `"*vater" S MASK PL GEN`

  - Eingabe: **`rast`**

    Zwischenebene: **`ras00t`** (für die 4. Ausgabe)

    Ausgabe: **`"rast~en" V IMP PRÄS SELTEN SG2`**

    `"ras~en" V IND PRÄS PL2`

    `"ras~en" V IMP PRÄS PL2`

    `"ras~en" V IND PRÄS SG2`

    `"ras~en" V IND PRÄS SG3`

GERTWOL: Anwendungen (z.B.)

- Einsetzung in **Disambiguierung**ssystemen.
  Beispielsatz: *"GERTWOL ist ein System zur automatischen Wortformerkennung deutscher Wörter."* Ausgabe nach der Disambiguierung:

```
"<*g*e*r*t*w*o*l>"      "*g*e*r*t*w*o*l" ABK S EIGEN
"<ist>"                 "sein" V IND PRÄS SG3
"<ein>"                 "ein" ART INDEF SG NOM NEUTR
"<*system>"             "*system" S NEUTR SG NOM
"<zur>"                 "zu-die" PRÄP ART DEF SG DAT FEM
"<automatischen>"       "automat~isch" A POS SG DAT FEM
"<*wortformerkennung>"
                  "*wort#form#er|kenn~ung" S FEM SG DAT
                  "*wort#form~er#kenn~ung" S FEM SG DAT
"<deutscher>"           "deutsch" A KOMP
                        "deutsch" A POS PL GEN
"<*wörter>"             "*wort" S NEUTR PL GEN
"<--punkt>"             ""  PUNKT
```

- **POS-Tagging** (z.B.: „Tagger for German" – Brill-Tagger, Uni Zürich, http://www.ifi.unizh.ch/CL/tagger/index.html)
- …

GERTWOL: Weitere Beispiele

```
 "<*wortformerkennungen>"
    "*wort#form#er|kenn~ung"  S FEM PL NOM
    "*wort#form#er|kenn~ung"  S FEM PL AKK
    "*wort#form#er|kenn~ung"  S FEM PL DAT
    "*wort#form#er|kenn~ung"  S FEM PL GEN
    "*wort#form~er#kenn~ung"  S FEM PL NOM
    "*wort#form~er#kenn~ung"  S FEM PL AKK
    "*wort#form~er#kenn~ung"  S FEM PL DAT
    "*wort#form~er#kenn~ung"  S FEM PL GEN

"<höchstwahrscheinlich>"
    "höchst#wahr|schein~lich"  ADV
    "höchst#wahrschein~lich"  A POS

"<*hausaufgaben>"
    "*haus#auf|gab~e"  S FEM PL NOM
    "*hau#sauf#gab~e"  S FEM PL NOM
```

GERTWOL: Weitere Beispiele

```
"<meine>"
   "ich"  poss PRON PERS SG NOM FEM
   "ich"  poss PRON PERS SG AKK FEM
   "ich"  poss PRON PERS PL NOM
   "ich"  poss PRON PERS PL AKK
   "ich"  poss DET PERS SG NOM FEM
   "ich"  poss DET PERS SG AKK FEM
   "ich"  poss DET PERS PL NOM
   "ich"  poss DET PERS PL AKK
   "mein~en"  V IND PRÄS SG1
   "mein~en"  V KONJ PRÄS SG1
   "mein~en"  V KONJ PRÄS SG3
   "mein~en"  V IMP PRÄS SG2

"<zugriffsbereite>"
   "zug#riff\s|bereit"  A POS SG NOM FEM
   "zu|griff\s|bereit"  A POS SG NOM FEM
```

Stemming; Wiederholung

- **Stemming**: Zurückführung der Wortformen auf einen gemeinsamen Stamm.
  - z.B.: *Brüder, Gebrüder, Bruder, brüderlich → Bruder*

    essen, aß, essbar → ess
  - Die Stammermittlung erfolgt möglichst ohne morphologische Analyse, stattdessen werden Regeln für Entfernung der Affixe oder statistische Methoden eingesetzt, um den Stamm eines Wortes zu ermitteln.
  - Mújdricza, Éva & Ganna Syrota (Folien) (2008): *Stemmingverfahren*. Link: http://www.cl.uni-heidelberg.de/~mujdricz/Referate/IR_Referat_Stemming_MujdriczaSyrota.pdf

- Wiederholung (Fragen)

Quellen

- GERTWOL: http://www2.lingsoft.fi/cgi-bin/gertwol

- Frank, Anette (2007) (Folien): *Morphologie und endliche Transduktoren.* Link: http://www.cl.uni-heidelberg.de/courses/ws07/ecl/transducer_morphII.pdf (Stand: 04.07.2008)
- Haapalainen, Mariikka & Ari Majorin (1994): *GERTWOL: Ein System zur automatischen Wortformerkennung deutscher Wörter.* Link: http://www.ifi.uzh.ch/CL/volk/LexMorphVorl/Lexikon04.Gertwol.html (Stand: 04.07.2008)
- Haapalainen, Mariikka & Ari Majorin (1995): *GERTWOL und Morphologische Disambiguierung für das Deutsche.* Link: http://www2.lingsoft.fi/doc/gercg/NODALIDA-poster.html (Stand: 03.07.2008)
- Herrmann, Teresa, David Beyer & Liu Xin (2004) (Folien): *Das Modell der Two-Level-Morphology.* Link: http://www.cl.uni-heidelberg.de/kurs/ws04/algomorph/twolevel-folien.pdf (Stand: 03.07.2008)

Quellen

- Holler, Anke (2007) (Folien): *Computermorphologie: Item and Process.* Link: http://www.cl.uni-heidelberg.de/kurs/ws03/morph/IandP.pdf, S. 11-19. (Stand: 04.07.2008)
- Koskenniemi, Kimmo (1983) (Dissertation): *Two-Level Morphology: A General Computation Model for Word-Form Recognition and Procuction.* http://www.ling.helsinki.fi/~koskenni/doc/Two-LevelMorphology.pdf (Stand: 03.07.2008)
- Koskenniemi, Kimmo (Bild): http://www.ling.helsinki.fi/~koskenni/kkoskenn.gif (Stand: 03.07.2008)
- Mújdricza, Éva & Ganna Syrota (2008) (Folien): *Stemmingverfahren*. Link: http://www.cl.uni-heidelberg.de/~mujdricz/Referate/IR_Referat_Stemming_MujdriczaSyrota.pdf (Stand: 04.07.2008)
- Volk, Martin (1999): *Zwei-Ebenen-Morphologie.* Link: http://www.ifi.uzh.ch/CL/volk/LexMorphVorl/Lexikon05.html (Stand: 04.08.2008)