

Morphologie für die Computerlexikographie

Claudia Kunze

kunze8@gmail.com, kunze@c1.uni-heidelberg.de

Computational Lexicography

- Motivation
- einige Grundbegriffe der Morphologie
- Kodierung morphologischer Angaben in Wörterbüchern
- Systeme für die morphologische Analyse

Bedeutung der morphologischen Analyse

Die morphologische Analyse von Wortformen ist für die textbasierte Computerlexikographie aus zwei Gründen besonders wichtig:

- LEMMATISIERUNG: Zuordnung flektierter lexikalischer Einheiten auf ihre jeweilig lexikalisch relevanten Grundformen
- NEUWORTBILDUNG: Produktivität der Sprache greift häufig auf bereits im Lexikon verzeichnete lexikalische Einheiten zurück: morphologische Prozesse der DERIVATION und KOMPOSITION

Einschränkung auf die für uns relevanten Aufgaben der Suche im Wörterbuch und der Lexikonerweiterung

Einige Grundbegriffe

- **TOKEN:** eine Zeichenkette in einem Text, die in den meisten Fällen als **TEXTWORT** interpretiert werden kann. Davon zu unterscheiden sind die sog. **NICHT-WÖRTER** durch z.B. Schreibfehler, fehlerhafte Trennungen, irrtümlich gesetzte oder nicht entfernte Steuerzeichen etc.
- **TYPE:** Repräsentant einer Menge von Token, die gleich sind oder so ähnlich, dass sie zu einer Klasse gehören. Der Satz
(1) Weil sie so rasten, mussten sie oft rasten.
besteht aus 8 Token (oder 10, mit den beiden Interpunktionszeichen), die zu 6 (8) Types zusammengefasst werden: *Weil, sie, so, rasten, mussten, oft, (,), (.)*
- **MORPHEM:** kleinste bedeutungstragende Einheit und kleinster Bestandteil lexikalischer Zeichen, mit eigenständiger Funktion; Morpheme sind systematische Abstraktionen über den real in Wörtern auftretenden Bestandteilen, den **MORPHEN**.
- **MORPH:** Bestandteil eines Wortes, welches einem Morphem zugeordnet werden kann; die Zeichenketten *s* (in: *Garten-s*) und *en* (in: *Soldat-en*) sind Morphe, die das Morphem /Genitiv Singular/ realisieren. Die Zeichenkette *ung* hingegen realisiert in *Sprung* kein Morphem, sondern ist unselbständiger Bestandteil des Wortstamms.
- **STAMM-MORPHEM** bzw. **WORTSTAMM:** zentrales, bedeutungstragendes Element eines Wortes, (letzteres muss mind. einen Wortstamm haben). *seh* ist Stamm von *sehen, ansehenlich, Fernseher* etc. Aufteilung nach primärer Wortklasse in Nominal-, Verbal- und Adjektivstämme usw. Mehrere Wortstämme können zu einem Kompositum verbunden werden. Ein oder mehrere Wortstämme bilden den **WORTKERN**

Grundbegriffe, Forts.

- DERIVATIONSMORPHEM bzw. Affix: Morphem, das zu einem Stamm hinzutreten kann und diesen sowohl hinsichtlich der morphosyntaktischen Klasse (Wortart) als auch der Bedeutung systematisch verändern kann. Suffix *-bar* tritt an Verbalstämme, verändert die Wortklasse zu Adjektiv und fügt der Bedeutung des Verbalstamms die Modalität des „Möglichseins“ hinzu.
- FLEXIONSMORPHEM: tritt zu Stamm, Kompositum oder einem derivierten Wort und markiert dessen morphosyntaktische Kategorie. Mehrere Flexionsmorpheme können zusammen auftreten und z.B. bei Verben Tempus, Numerus und Person markieren (*mach-t-est*).
- FUGENMORPHEM: kann zwischen zwei Stamm-Morpheme treten, meist aus lautlichen Gründen. Die Fugenmorpheme des Deutschen sind zwar in der Form mit Flexiven des Deutschen identisch, nicht jedoch in ihrer Distribution (vgl. *Arbeit-s-amt*).
- GRUNDFORM einer lexikalischen Einheit (auch ZITIERFORM: Form, unter der diese lexikalische Einheit im Wörterbuch verzeichnet wird. Es gibt wortarten- und sprachspezifische Konventionen für die Bildung der Grundform einer lexikalischen Einheit. Die Grundform kann mit dem Wortstamm zusammenfallen (z.B. *rot*, kann aber ebenso über den Stamm hinausgehen (Stamm: *seh*; Grundform: *sehen*).

Morphologische Wortgrammatik

Komplexe Wörter sind nach gewissen Regularitäten aufgebaut, die sich mittels einer “Wortgrammatik” strukturell erfassen lassen:

- Bedingungen für die Aufeinanderfolge der verschiedenen Arten von Morphemen (Stamm, Affix, Flexiv) führen zu folgender kanonischer Wortstruktur:
 - (1) PRÄ* STAMM [FUGE? STAMM]* SUFF* FLEX*
 - PRÄ=Präfix, STAMM = Stammmorphem, FUGE = Fugenmorphem, SUFF = Suffix, FLEX = Flexiv; der Stern kennzeichnet, wie in regulären Ausdrücken, das markierte Element als optional und beliebig oft iterierbar (theoretische Option!). Beispiele:
 - (2) ver-un-treu-en, Un-wäg-bar-keit, Arbeit-s-amt-s-statistik-en
- Zwischen Morphemen verschiedener Klassen bestehen Beziehungen der Rektion oder der Subkategorisierung, ähnlich wie in der Syntax. Das Suffix *-bar* subkategorisiert ausschließlich verbale Stämme und führt einen Kategorienwechsel zum Adjektiv herbei.

Wortgrammatik, Forts.

- RIGHT HAND HEAD RULE (Williams:1981): in morphologisch komplexen Wörtern ist das am weitesten rechts auftretende Element der Head (Kopf) dieses Wortes, d.h. dass dieses Element alle morphologisch relevanten Merkmale determiniert. Beispiel: Phantasiewort *Wrxlbarkeit*:

- Identifizieren des am weitesten rechts stehenden Elements: das Suffix *-keit*
- lexikalische Beschreibung als ein Substantiv mit dem Genus Femininum und einer bestimmten Flexionsklasse
- Aussagen über Wortform(en) ohne Wissen über die Bedeutung des Stammes (*wrxl*) möglich
- Rektionseigenschaften des Suffixes *-bar* weisen auf einen Verbalstamm und die Existenz eines Verbs *wrxln* hin
- (teil)automatischen Analyse unbekannter Wörter in Texten: rechts beginnen und nach dem Verfahren des LONGEST MATCH Flexive und Suffixe identifizieren (die ja geschlossene Klassen sind), danach von links nach Präfixen suchen und den Rest als Stamm des Wortes annehmen

Flexionsmorphologische Analyse

Flexive bilden den rechten Rand von Wörtern, bis auf sehr wenige Ausnahmen, in denen ein Flexiv im Wortinneren stehen kann (*eine Loseblattsammlung* → *einer Lose-n-blatt-sammlung*). Flexive determinieren kongruenzrelevante Merkmale eines Wortes in einer NP oder VP: Kasus, Numerus, Tempus, Person etc. Neben dem System der Flexionsmorpheme, einem Inventar mit nur wenigen Elementen, kennt das Deutsche die Markierung von morphosyntaktischen Merkmalen durch Umlautung (z.B. *Bogen* – *Bögen*) oder Ablautung des Stammes (z.B. *schwimmen* – *schwammen*). Zwei Arten der Ambiguität bei der morphematischen Analyse:

- Zeichenkette wie z.B. *-en* kann entweder Teil eines Wortstamms oder Instanz eines Flexionsmorphems sein. Die Wortform *Rahmen* ist so interpretierbar als eine Instanz des lexikalischen Zeichens */Rahm/* (mit Flexionsendung) oder als Instanz des lexikalischen Zeichens */Rahmen/* (ohne Flexionsendung). Diese Art der Ambiguität verschlechtert das Ergebnis einer Lemmatisierung. Statistische Verfahren können helfen, die im gegebenen Kontext wahrscheinlichere Grundform auszuwählen.
- Die Flexive selbst sind hinsichtlich ihrer Funktion mehrdeutig: Zeichenkette *-en*, an einen verbalen Stamm angehängt, kann Infinitivform oder erste oder dritte Person Plural des Präsens markieren; Problem z.B. beim Part-of-speech-Tagging, wo Unterscheidung zwischen infinitiver und finiter Verbform zu treffen ist. Statistische Verfahren eignen sich auch hier zur Auswahl der kontextuell wahrscheinlicheren Merkmalsausprägung.

Derivationsmorphologische Analyse

Derivationsmorpheme bilden ebenfalls eine relativ stabile, überschaubare Menge von Elementen

- größter Anteil: Suffixe, wenige Präfixe
- deutsche Derivationsmorphologie: keine Infixe (wie z.B. das türkische Verneinungsinfix *-m* in *gel-m-iyor*um (auf Deutsch: "Ich komme nicht.")
- Zwischenklasse der AFFIXOIDE oder KONFIXE:
- Definition von Müller (1998:214): Unter einem Konfix wird ein Wortbildungselement mit lexikalisch-begrifflicher Bedeutung verstanden, das nicht wortfähig ist, sondern nur in Kombination mit anderen Morphemen auftreten kann; entweder nur als Derivationsbasis (*fanat-*) oder nur als Erst- bzw. Zweitglied in Komposita (*bio-*, *-naut*, *-thek*) oder sowohl als Basis als auch als Kompositionsglied (*therm*)
- produktive Konfixe der letzten Zeit, z.B. *cyber-* und *pseudo-*, sind lexikalisch relevant
- automatische Erkennung von Derivationsmorphemen muss mit Ambiguität umgehen können: Zeichenkette *-ung* kann sowohl ein Suffix sein als auch Teil eines Wortstamms (etwa in *Dung* und *Sprung*). Zu dem Suffix *-bar* gibt es einen formgleichen Stamm, wie z.B. in *Cocktailbar*

Kompositionsmorphologische Analyse

Der im Deutschen produktivste Prozess der KOMPOSITION operiert über der offenen Klasse der Wortstämme und der komplexen Wörter. Komposita können deshalb nicht vollständig im Lexikon erfasst werden, aber es ist nützlich, wenn nicht im Lexikon verzeichnete Komposita formal in ihre Bestandteile zerlegt werden, die dann im Lexikon aufgefunden werden können.

Mögliche Ambiguitäten bei der Zerlegung:

- Rein formale Ambiguitäten (sehr selten), z.B.

- (3) Staub-ecke-n vs. Stau-becken

- Strukturelle Ambiguitäten bei drei- und mehrgliedrigen Komposita: eine

*Hofmusik*kapelle als Musikkapelle am Hofe eines Adligen sein oder eine Kapelle, die Hofmusik spielt.

- Aufgabe: Worin besteht die Ambiguität des Wortes *Wachtraum*?

- automatische Analyse der syntaktischen und semantischen Beziehungen der

Bestandteile eines Kompositums schlägt fehl: Beispiel Wortpaar *Schweineschnitzel* –

Jägerschnitzel, formal analoge Formen, deren Bestandteile aber in verschiedenen

inhaltlichen Beziehungen stehen (Schnitzel vom Schwein bzw. Schnitzel nach Jägerart)

- morphologisches Analysesystem GERTWOL:

<http://www2.lingsoft.fi/cgi-bin/gertwol>

Morphologische Angaben in Wörterbüchern

Übersicht über verschiedene morphologische Beschreibungen in Printwörterbüchern und maschinenlesbaren Wörterbüchern für den menschlichen Benutzer mit z.T. sehr unterschiedlichen Verfahren der Darstellung, je nach Phänomen und Zielgruppe

- Angaben zur Flexion
- Angaben zur Derivation
- Angaben zur Komposition

Angaben zur Flexion

- Angabe aller Vollformen einer lexikalischen Einheit ist eher selten
- Printwörterbücher zeigen aus Platzgründen meist nur Lexeme mit unregelmäßigen Paradigmen vollständig an
- elektronische Wörterbücher erwägen Angabe aller Vollformen auch bei regelmäßig flektierenden lexikalischen Einheiten (Platzargument entfällt)

throw

throw (**θrəʊ**) *vb.* **throws, throwing, threw, thrown.**

(*mainly tr.*)

1. (*also intr.*) to project or cast (something) through the air, esp. with a rapid motion of the arm and wrist.
2. (foll. by *in, on, onto*, etc.) to put or move suddenly, carelessly, or violently: *she threw her clothes onto the bed.*

Abbildung 1: Artikel zum Stichwort *throw* aus dem *Collins English Dictionary*

Angaben zur Flexion, Forts.

- deutsche einsprachige Lexikographie: es gibt häufig Eckformen bei Substantiven
- meist verdichtete Angabe von Genitiv Singular und Nominativ Plural (SgBA und PlbA) neben der Zitierform, dem Nominativ Singular: geübter Benutzer kann alle anderen Formen erschließen.

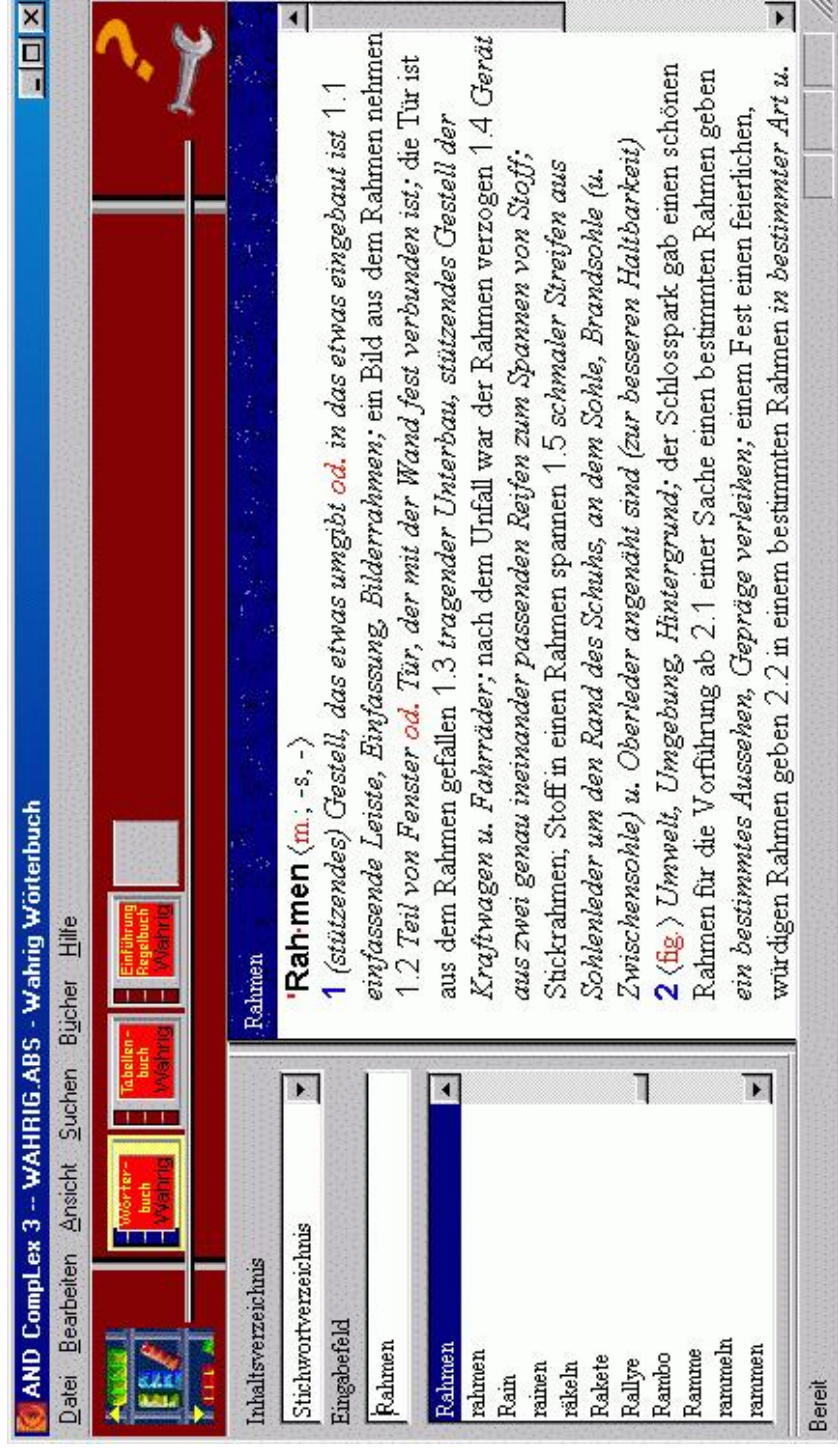


Abbildung 2: Artikel zum Stichwort *Rahmen* aus der elektronischen Version von *Wahrigs Deutschem Wörterbuch*

Angaben zur Flexion, Forts.

- Variante: Angabe von Eckformen unregelmäßiger Verben, entweder in externen Tabellen (der Umtexte) oder in den Artikeln der entsprechenden Verben
- sowohl in einsprachigen als auch in zweisprachigen Wörterbüchern
- Angabe des Infinitivs, der zweiten oder dritten Person Singular Präsens, der ersten Person Singular Präteritum und des Partizip Perfekt (Beispiel: *raten, rätst, riet, geraten*).
- Gebrauch von Verweisen führen entweder aus dem Wörterverzeichnis heraus zu Formentafeln in den Umtexten des Wörterbuchs oder auf einen Eintrag im Wörterverzeichnis, in dessen Formkommentar die benötigten Informationen gegeben werden.
- Abb. 3 zeigt den Eintrag *Rahmen* aus der elektronischen Version des *Bertelsmann Rechtschreibwörterbuchs*. Bei Anklicken einer spezifischen Ziffer gelangt man zur Formentafel, die das komplette Formenparadigma anhand eines kanonischen Beispieleintrags angibt.
- Unterstützung des Benutzers in Produktionssituation zur Bildung der korrekten Vollformen spezifischer lexikalischer Zeichen (alle Verfahren)

Angaben zur Flexion, Forts.



Abbildung 3: Artikel zum Stichwort *Rahmen* aus der elektronischen Version des Bertelsmann Rechtschreibwörterbuchs

Angaben zur Flexion, Forts.

- explizites Verweis-Verfahren vor allem in zweisprachigen Wörterbüchern und in einsprachigen Lernerwörterbüchern: Aufnahme unregelmäßiger Vollformen lexikalischer Zeichen als Verweiseinträge in das Wörterverzeichnis
- Angabe enthält zugehörige Grundform (Basiseintrag) und manchmal zusätzlich morphosyntaktische Merkmale der Vollform
- Unterstützung der Benutzer bei der Lemmatisierung während der Rezeption fremdsprachiger Texte
- Beispiel: Verweiseintrag *cupo* aus dem *Klett Wörterbuch Spanisch-Deutsch*
(4) **cupo** I. 3. *pret de caber* II. ...
- schwierig: Angabe vermuteter Einschränkungen bei der Bildung der Vollformen, z.B. die Angabe eines nicht vorhandenen Plurals (PlbA: o. *Pl.*) oder zur Einschränkung der Steigerbarkeit von Adjektiven.
- Beide Arten von Angaben sind nur in der Produktionssituation von Nutzen
- sollten auf gründlicher Korpusanalyse basiert sein (bislang ist das bei den einsprachigen Wörterbüchern des Deutschen nicht der Fall)

Angaben zur Derivation

Wortbildungselemente wie AFFIXE, AFFIXOIDE und KONFIXE sollten lexikalisch beschrieben werden, ggf. als eigene Einträge oder sonst als Liste in den Umtexten. Müllers (1989)

Vorschlag:

- ausreichende semantische Einordnung, differenziert nach Gebrauchsweisen,
- Charakterisierung der Basiswörter nach Wortart sowie ggf. in grammatischer und semantischer Hinsicht,
- Veränderungen bei der Wortbildung (Umlautung, Fugenmorpheme etc.),
- Restriktionen bei der Auswahl der Wortbasis,
- Hinweise auf Synonyme und Antonyme.

Beispielseintrag zum Wortbildungselement *-able*, *-ible* aus dem *Oxford Advanced Learner's Dictionary of Current English*, der schon einige der Müllerschen Anforderungen bedient:

-able, -ible [...] *suff* **1** (with *ns* forming *adjs*) having or showing the quality of *fashionable* ○ *comfortable*; **2** (with *vs* forming *adjs*) **(a)** that may or must be: *eatable* ○ *payable* ○ *reversible*. **(b)** tending to: *changeable* ○ *perishable*.

Angaben zur Derivation, Forts.

Alternativer Ansatz in Kempckes *Wörterbuch Deutsch als Fremdsprache*, einem Lernerwörterbuch mit nur ca. 20 000 Einträgen:

- zu elementaren lexikalischen Einheiten werden Derivationen und Komposita angegeben (Wortfamilie)
- systematische Unterstützung des Lernalers bei der Erweiterung seines fremdsprachlichen Wortschatzes
- Artikel zum Stichwort *mehr*

¹ **mehr** [...] <Indefinitpron.: für Mask., Fem., Neutr. und Pl.; Komp. zu ¹viel; ² auch *mehrere*; ³ TAFEL X /drückt aus, dass eine Anzahl, eine Menge über ein bestimmtes Maß hinausgeht/: <adj.> *er hat (viel) ~ Geld, Bücher als wir; er will noch ~ Bücher kaufen; es waren ~ Leute da, als du denkst; <subst.> er hat ~ verlangt, als sie leisten konnte; sie hat ~ von ihm erwartet; darf es etwas sein? <2> mehr, Mehr, mehrten, **mehrlei, mehrfach, Mehrheit vermehren – mehrdeutig, Stimmenmehrheit, vielmehr**; vgl. **mehr/Mehr-**; vgl. **viel***

Nur die regelmäßige Analyse von signifikanten Korpora kann eine zuverlässige Grundlage für die Beschreibung von Wortbildungselementen bilden, z.B. neue Wortbildungselemente auffinden und die Produktivität existierender Wortbildungselemente einschätzen.

Angaben zur Komposition

Selektive Aufnahme von Komposita in ein Wörterbuch; mit einer Präferenz für opake (aus den Bestandteilen nicht regelhaft semantisch erschließbare) gegenüber transparenten Komposita.

- Langenscheidts Lernerwörterbuch und das sechsbändige WdG geben transparente Komposita an (s. Beispielseintrag Lehrer aus dem *Langenscheidt Großwörterbuch Deutsch als Fremdsprache*)
- transparente Komposita sind als Sublemmata ohne eigenen semantischen Kommentar am Schluss des Artikels eines der beteiligten Wortstämme aufgeführt
- polyseme Lexeme enthalten Komposita als Sublemmata am Schluss des jeweiligen semantischen Subkommentars
- Angabe der Stellung des Stammes innerhalb des Kompositums als Erst- oder Zweitglied

Leh·rer der; -s, -; j-d, der bes an e-r Schule Unterricht gibt ↔ Schüler <ein strenger, erfahrener L.>: *Er ist L. für Mathematik u. Physik an e-m Gymnasium; Wen habt ihr als L. in Sport?* || K-: **Lehrer-**, -ausbildung, -beruf, -ehepaar, -kollegium, -konferenz, -mangel, -schwemme, -zimmer, -überschuß || -K: Berufsschul-, Grundschul-, Hauptschul-, Sonderschul-, Biologie-, Chemie-, Deutsch-, Englisch-, Französisch-, Latein-, Mathematik-, Physik-, Sport- || *hierzu* **Leh·re·rin** die; -, -nen

Systeme für die morphologische Analyse:

GERTWOL (German Two Level Morphology) wurde in Finnland entwickelt und wird von der Firma *Lingsoft* vertrieben. Die Entwickler charakterisieren GERTWOL so:

GERTWOL ist ein System zur automatischen Wortformerkennung deutscher Wörter. Die theoretische Grundlage des GERTWOL-Systems bildet die sprachunabhängige morphologische Analysemethode ZWEI-EBENEN-MODELL [...] von Professor Kimmo Koskeniemi [...] GERTWOL verfügt über ein Wortformenlexikon von etwa 60 000 Wortformen, deren Geltungsbereich durch eine umfassende Derivationsmorphologie und einen vollständigen Mechanismus zur Kompositabbildung erheblich erweitert wird [...] Bei der morphologischen Analyse gibt GERTWOL die möglichen Lesungen einer deutschen Wortform an. Die Lesungen bestehen aus der Grundform des Wortes und aus morphologischen Daten zur Grundform. (Haapalainen:1995, <http://www2.lingsoft.fi/cgi-bin/gertwol>)

- GERTWOL enthält Regelapparat für die morphologische Analyse und eine umfangreiche lexikalische Ressource.
- sprachunabhängige Basis der Regelkomponente: Zwei-Ebenen-Morphologie (vgl. Koskeniemi:1983)
- Fokus auf die verwendete lexikalische Ressource
- Kodierung morphologischer Information in GERTWOL im Vergleich zu den Kodierungsgepflogenheiten in Printwörterbüchern

Eigenschaften von GERTWOL

- vollständige flexionsmorphologische und eine umfassende derivations- und kompositionsmorphologische Analyse, aber keine semantischen oder syntaktischen Analysen;
- lexikalische Daten basieren auf dem lexikalischen Material eines deutsch-englischen Wörterbuchs von HarperCollins, mit korpusbasierter Erweiterung und Hinzunahme von Eigennamen. Das Wortformenlexikon umfasst insgesamt etwa 85 000 Einträge;
- lexikalische Ressource von GERTWOL umfasst ein Stammformenlexikon und ein Vollformenlexikon sowie weitere Teillexika (s.u.);
- Ein Stammform-Eintrag besteht aus der Stammform des jeweiligen Wortes sowie dem passenden Flexionsklassenkürzel, das die so genannten Fortsetzungsklassen (Endungsparadigmen) spezifiziert;
- Ein Vollform-Eintrag besteht aus einer Wortform und ihren morphosyntaktischen Merkmalen.
- Die in GERTWOL kodierten Flexionsklassen fassen (Gruppen von) Stammformen sowie Fortsetzungsklassen zusammen. Außerdem werden morphophonemische Phänomene wie der schwa-Ausfall berücksichtigt;
- Verben werden hinsichtlich des Modus, Tempus, der Person und des Numerus gekennzeichnet; zwölf Konjugationsklassen für schwach konjugierende Verben; die starken Verben sind in zwölf Hauptkonjugationsklassen eingeteilt;

Eigenschaften von GERTWOL, Forts.

- Substantive werden hinsichtlich Genus, Numerus und Kasus gekennzeichnet. Substantive sind in zehn Hauptdeklinationsklassen eingeteilt; eine weitere Untergliederung führt zu ca. 300 Deklinationsklassen. Jede Deklinationsklasse besteht aus einer Fortsetzungsklasse mit Singularformen und einer mit Pluralformen. Desweiteren werden produktive Ableitungselemente (=Derivationssuffixe) morphosyntaktisch markiert.
- Adjektive werden durch die Merkmale Komparierbarkeit, Numerus, Kasus und Genus charakterisiert: regelmäßige Adjektive werden in siebzehn Deklinationsklassen eingeteilt. Ebenfalls morphosyntaktische Markierung produktiver Ableitungselemente (=Derivationssuffixe);
- Neben Hauptwortklassen werden einige Pronomina, die Artikel, Numeralia und Eigennamen morphologisch kodiert;
- erfasst das System der Präfigierung der Verben, unterscheidet abtrennbare und nicht-abtrennbare Präfixe; beschreibt ebenfalls Präfixe von Substantiven und Adjektiven;
- Bei Suffigierung wird unterschieden zwischen: a) Derivation 1. Grades als der Suffigierung an einen Wortstamm und b) syntagmatischer Derivation als Anfügen eines Suffixes an ein anderes Suffix; enthält ca. 80 Wortbildungsmuster dieses Typs;

Eigenschaften von GERTWOL, Forts.

- Markierung des nicht-markierten Übergangs einer lexikalischen Einheit von einer Wortklasse in die andere (Konversion). Unter dieses Schema werden auch substantivische Zusammenrückungen (z.B. *Stellidichein*) subsumiert;
- Kompositionsmorphologie: Behandlung der Fugenmorpheme; Annahme, dass die Fugenelemente deklinationsklassenspezifisch sind;
- Substantivische Zweitglieder von Komposita, die nicht als selbständige Wörter vorkommen, wurden in ein Teillexikon aufgenommen; ebenfalls Teillexikon für Halbsuffixe und nicht-selbständige adjektivische Zweitglieder;
- Suffixe werden in heimische (native) und fremde Suffixe getrennt. Von dieser geschlossenen Klasse der Suffixe werden die hier als HALBSUFFIXE bezeichneten Affixoide und Konfixe getrennt und in einem anderen Teillexikon beschrieben.

Abbildungen GERTWOL

Die Analyse des Wortes **Arbeitervereinigung**:

```
"<*arbeitervereinigung>"
"*arbeit~er#ver|ein~ig~ung"  S FEM SG NOM
"*arbeit~er#ver|ein~ig~ung"  S FEM SG AKK
"*arbeit~er#ver|ein~ig~ung"  S FEM SG DAT
"*arbeit~er#ver|ein~ig~ung"  S FEM SG GEN
```

Abbildung 4: Informationen zum Stichwort *Arbeitervereinigung*

Die Analyse des Wortes **abzusingen**:

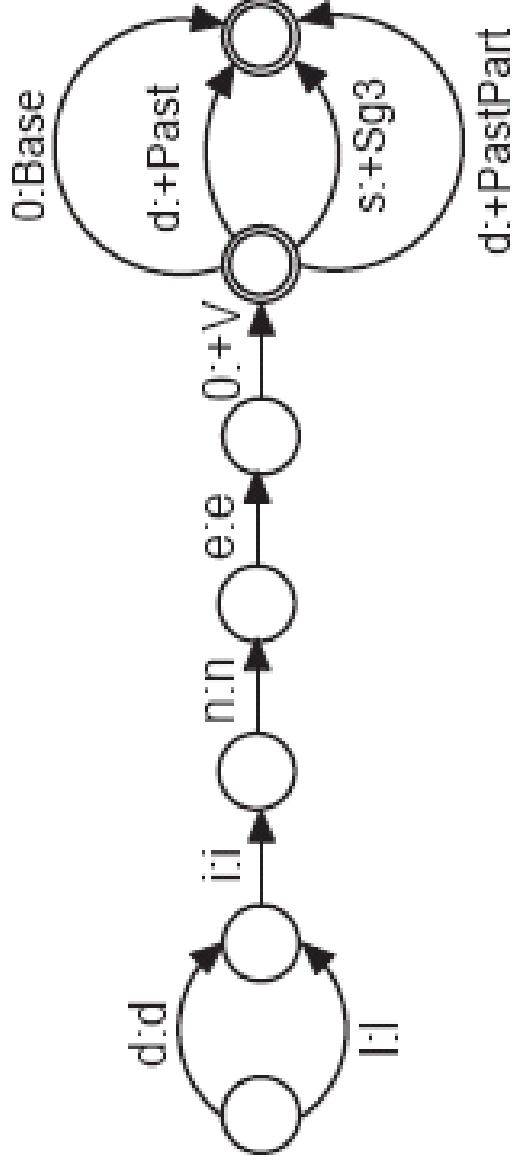
```
"<abzusingen>"
"ab|sing~en"  V TRENNBAR INF zu
```

Abbildung 5: Informationen zum Stichwort *abzusingen*

WordManager und Canoo

Komponenten des Systems WordManager (Fa. Canoo, Basel):

- Lemmatisierer, der Vollformen auf die Basisform(en) reduziert;
- Flexionsanalyse zur Bestimmung der Flexionselemente eines Wortes;
- Wortanalyse: Wohlgeformtheits-Check; Bestimmung der Worтеlemente;
- Wortformengenerator zur Erzeugung aller Vollformen zu einer Basisform.
- zugrunde liegendes Wörterbuch mit ca. 250 000 Basisformen, zu mehr als 3 Mill. Vollformen expandierbar;
- Analyse unbekannter Wörter (nicht fehlerfrei);
- zugrunde liegende Technologie: Finite State Transducer



WordManager, Forts.

Informationssystem auf der Webseite von Canoo: [S. \[www.canoo.net/\]\(http://S.www.canoo.net/\)](http://S.www.canoo.net/) mit Wortabfrage und Links zu wichtigen Bereichen (Wörterbücher, Rechtschreibung, Flexion, Wortbildung, Index und Terminologie) lädt zum Experimentieren ein.

- morphologische Analyse von Wörtern wird gestützt von einer Reihe von Spezialwörterbüchern;
- eigene Spezialwörterbücher für Rechtschreibung, Flexion, Wortbildung und Morphologie; Verlinkung mit externen Ressourcen LEO (Deutsch-Englisches Übersetzungswörterbuch) und der deutschen Wikipedia (allerdings nur für im Canoo-WB vorhandene Stichwörter);
- Spezial-Wörterbücher basieren auf dem Morphologischen Wörterbuch von Canoo:
 - enthält mehr als 200 000 Einträge und generiert mehr als 2 Mill. Wörter (Wortformen);
 - zu jedem Eintrag enthält das morphologische WB alle generierten Wortformen (auch mehrgliedrige lexikalische Einheiten), sowie ihre morphosyntaktischen und schreibungsrelevanten Charakteristika; die aus dem Eintrag generierten Einträge, der oder die Einträge, aus dem der Eintrag generiert wurde, die jeweilige Flexionsregeln, sowie die mit dem Eintrag assoziierten Rechtschreibregeln;

WordManager, Forts.

- Rechtschreibwörterbuch: mehr als 250 000 deutsche Schlagwörter und Redewendungen:
 - alte und neue Schreibung (mit Darstellung des Verhältnisses der beiden Varianten zueinander: Ersetzung oder Alternative) sowie Links zu den Regeln der neuen Rechtschreibung;
 - Angaben zu regionalen Varianten und Schweizer Schreibungsvarianten sowie Angaben zum Gebrauch;
 - bei Verben das für die Bildung der zusammengesetzten Zeitformen verwendete Hilfsverb und bei Substantiven das Genus;
- Flexionswörterbuch: vollständige Flexionstabellen für 200 000 deutsche Stichwörter mit ca. 2 Mill. Wortformen und Wortkategorisierungen sowie Links zu den Flexionsregeln:
 - Deklinationstabellen für Substantive berücksichtigt homographie Einträge wie *Mangel* oder *Service*), sowie Umlautung im Plural unter dem Etikett *Besonderheiten*;
 - Varianten von Flektionsformen (z.B. indigenes und eingedeutsches Paradigma) sind angeführt, z.B. bei *Index*, *Atlas*, *Konto*.
- Wörterbuch der deutschen Wortbildung: Wortbildungszusammenhänge für über 200 000 deutsche Stichwörter, mit Links zu den entsprechenden Wortbildungsregeln: morphologische Wortanalysen in Form von Ableitungsgraphen; weitere Derivationen und Komposita auf Grundlage des Suchwortes; Nennung orthographischer Veränderungen

WordManager, Forts.

- Wörterbuch der Morphologie: morphosyntaktische Klassifizierungen der rund 2 Mill. Wortformen der 200 000 deutschen Stichwörter. Sowohl Wortformen als auch Wortbildungen können interaktiv analysiert und generiert werden;
- zu Vollformen werden die Lexeme (repräsentiert durch Grundform) mit den morphosyntaktischen Merkmalen angezeigt sowie die Positionen, die Vollformen in den Paradigmen dieser Lexeme annehmen; übersichtliche Gestaltung der Tabellen;
- zusätzlich auf Website: Rechtschreibprüfer, in den man Text eingeben und auf korrekte Schreibung prüfen lassen kann; Anzeige der Wörter, deren Schreibung durch die Orthographiereform berührt werden; Anzeige der nicht abgedeckten Wörter; Überprüfung auch nach Orthographiestil (ZEIT vs. NZZ);
- Fazit: System hat eine beeindruckende Abdeckung, besonders in der Kombination eines großen Wörterbuchs mit einer Analysekomponente für unbekannte Wörter

Wortbildung von *Grundversorgung*

Wortbildungsanalyse





Abbildungen Canoo

Flexion von vorsingen

Rechtschreibung

Wortklasse: Verb

Stammformen: vorsingen / vorsang / vorgesungen

Hilfsverb: haben

Flexionsklasse: unregelmäßige Verben, trennbar

Besonderheiten: Trennbare Verben, e-Tilgung im Konjunktiv II, Ablaut in Stammformen

Einfache Zeiten

Präsens Hauptsatz			Präsens Nebensatz		
Indikativ		Konjunktiv I	Indikativ		Konjunktiv I
Person	Verb	Person	Person	Verb	Verb
ich	singe vor	ich	ich	vorsinge	vorsinge
du	singst vor	du	du	vorsingst	vorsingst
er/sie/es	singt vor	er/sie/es	er/sie/es	vorsingt	vorsinge
wir	singen vor	wir	wir	vorsingen	vorsingen
ihr	singt vor	ihr	ihr	vorsingt	vorsinget
sie	singen vor	sie	sie	vorsingen	vorsingen

Präteritum Hauptsatz			Präteritum Nebensatz		
Indikativ		Konjunktiv II	Indikativ		Konjunktiv II
Person	Verb	Person	Person	Verb	Verb
ich	sang vor	ich	ich	vorsang	vorsänge
du	sangst vor	du	du	vorsangst	vorsängest vorsängst
er/sie/es	sang vor	er/sie/es	er/sie/es	vorsang	vorsänge
wir	sangen vor	wir	wir	vorsangen	vorsängen
ihr	sangt vor	ihr	ihr	vorsangt	vorsänget vorsängt
sie	sangen vor	sie	sie	vorsangen	vorsängen

Abbildung 8: Flexionsanalyse von vorsingen



Abbildungen Canoo

canoonet
Powered by Canoo Products

in: ☐ Groß-/Kleinschreibung ☐ ss statt ß

Wörterbücher | Wortgrammatik | Wortbildung | Satzgrammatik | Rechtschreibung | Glossar | Index
Überblick | Rechtschreibregeln | Prüfer | Reform | Wörterbuch | Amtliche Regelung

- Einleitung
- ▼ Stamprinzip
 - Einleitung
 - Umlautschreibung
 - Konsonantenverdoppelung
 - [Ersetzung von ß](#)
 - Drei gleiche Buchstaben
 - Wortangleichung
 - Umwandlung von -ti zu -zi
 - Buchstabenverschmelzung
 - Einzeilfälle
- Fremdwörter
- Getrennt/Zusammen
- Bindestrich
- Groß/Klein
- Zeichensetzung
- Worttrennung

Ersetzung von ß durch ss

Nach kurzem [Vokal](#) wird nicht mehr ß, sondern ss geschrieben.

(In der Schweiz wird wie vor der Reform nie ß, sondern immer nur ss geschrieben.)

Schreibweise Regionaler Filter ☐ richtig ☒ falsch
Neu ohne Variantenwahl

3059 von der Regel betroffene Einträge im Canoo Wörterbuch

Seite: ◀ 21 22 23 24 25 **27** 28 29 30 ▶ Seitengröße: 10 anzeigen

Aufschluß	Aufschluss
Aufschlußarbeiten	Aufschlussarbeiten
aufschlußle	aufschlussle
Aufschlußlösung	Aufschlusslösung
aufschlußreich	aufschlussreich
aufschluß, -t	aufschmiss, -est, -t
aufschloß, -t	aufschloss, -est, -t
aufsproß, -t	aufspross, -est, -t
Augsproß	Augspross
Ausbiß	Ausbiss

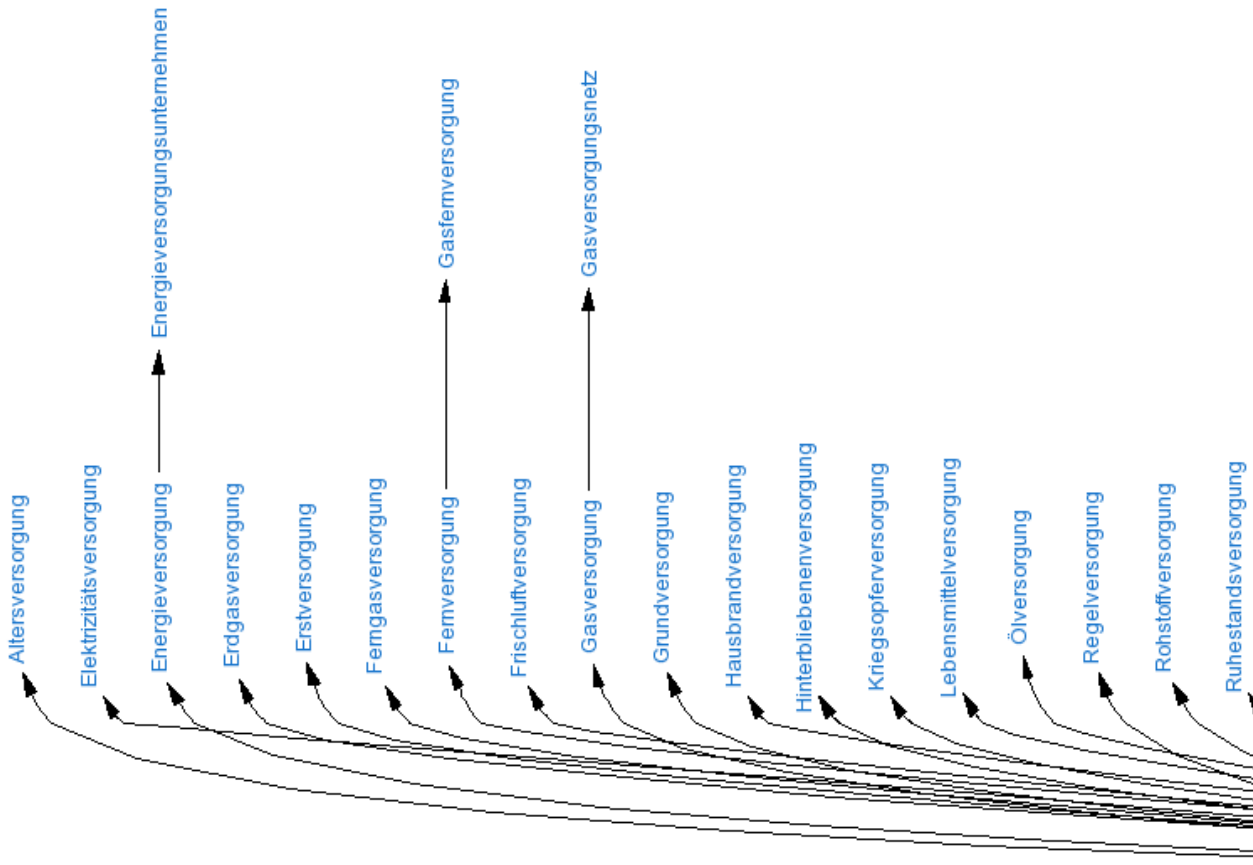
(Amtliches Regelwerk: [§25](#))

Abbildung 9: Informationen zum Stichwort *Aufschluß*



Abbildungen Canoo

Ableitungen



Vergleich GERTWOL - WordManager

Beide Systeme beruhen auf Finite State Transducer Technologie. Ein Analyseversuch mit den Eingabewörtern *Zeitungsangebot* und *dazwischensang*, die beide nicht in den Systemwörterbüchern vorhanden, aber relativ einfach zu analysieren sind, zeigt vergleichbare Resultate:

- Während man bei GERTWOL eine äußerst knapp formulierte Analyse erhält, bekommt man bei CANOO eine weniger aussagekräftige, graphisch aber ansprechender gestaltete Beschreibung; weitere Informationen sind durch Anklicken der Analyseelemente abrufbar;
- GERTWOL: Großschreibung des Wortes weist auf ein Nomen-Vorkommen hin. Bei Canoo spielt Schreibung am Wortanfang keine Rolle. Canoo generiert z.B. für *dazwischensang* eine weitere Analyse mit *sang* als nominalem Kopf (hier nicht zutreffend);
- Beide Systeme versagen, wenn weder das gesamte Wort noch die Bestandteile lexikalisch verzeichnet sind: Analysebeispiel *Cyberwar* (Ende Februar 2007 verzeichnete Google für dieses Wort ca. 62 000 Treffer auf deutschen Webseiten) wird von keinem der beiden Systeme analysiert.

Morphologische Analyse und Präsentation

Flexion

- Printwörterbücher, GERTWOL: Endungsparadigmen werden kodiert (GERTWOL in expliziter Weise);
- Canoo und Morphy: Paradigmen werden komplett kodiert;
- Printwörterbücher: Angabe von Eckformen für Substantivlemmata;
- Berücksichtigung morphophonemischer Phänomene in GERTWOL: große Zahl von Flexionsklassen insb. bei Substantiven; Umlautung durch eine Regel;
- in Printwörterbüchern wird Umlautung an jedem einzelnen Eintrag markiert;
- komplette Auflistung unregelmäßiger Formen in GERTWOL, Canoo etc. als Wortformen;
- explizite Nennung auch in den meisten Printwörterbüchern.

Derivation

- GERTWOL: explizite Listung und morphosyntaktische Klassifikation der Affixe und Affixoide;
- Printwörterbücher haben selten Systematik und Konsistenz von GERTWOL
- GERTWOL hat keine syntaktischen Angaben und damit auch keine Subkategorisierungsangaben von Suffixen, wie z.B. das *Oxford Advanced Learner's Dictionary* (OALD);

Vergleich, Forts.

- morphologische Konversion wird in Printwörterbüchern nur in Zusammenhang mit einhergehender semantischer Veränderung beschrieben - GERTWOL scheint systematischer vorzugehen
- Komposition
 - Auflistung von Komposita, wie in Printwörterbüchern, nicht notwendig bei Systemen, die eine Analysekomponente für unbekannte Wörter haben wie GERTWOL und Canoo;
 - für die Lemmaaufnahme wichtige Unterscheidung zwischen transparenten und opaken Komposita entfällt (bei transparenten Komposita kann man die Gesamtbedeutung aus der Bedeutung der Bestandteile erschließen, z.B. *Apfelbaum*; bei opaken Komposita nicht, z.B. *Bierdeckelsteuer*), da keine semantischen Beschreibungen vorgenommen werden;
 - systematische Behandlung von Fugenmorphemen (die hier den Flexionsklassen der Stämme zugeordnet werden) umso wichtiger bei GERTWOL und WordManager;
 - Fugenmorpheme als Wortbildungselemente ohne eigenen semantischen Status werden in Print-WB nicht behandelt (systematische Beschreibung gehört in (Wörterbuch-)Grammatik).