

DeKo – Ein System zur Analyse komplexer Wörter*

Tanja Schmid, Anke Lüdeling, Bettina Säuberlich, Ulrich Heid und Bernd Möbius

Zusammenfassung

DeKo ist ein System zur Analyse von *Derivations-* und *Kompositionsprodukten* im Deutschen.¹ Das System zerlegt komplexe Wörter in Morpheme und weist Struktur(alternativen) zu, wobei es Informationen aus allen linguistischen Ebenen nutzt. Das Ziel von DeKo ist neben dem Aufbau eines Lexikons, das Wortbildungsinformationen enthält, eine Regelbasis, die auch ungesehene komplexe Wörter analysieren kann.

Im Folgenden beschreiben wir die linguistischen Grundannahmen von DeKo sowie die Implementierung mit Finite State Transducern (Lextools Paket von Sproat, 2000).

5.1. Motivation

Die Zerlegung von komplexen Wörtern ist für viele computerlinguistische Anwendungen erforderlich: So können zum Beispiel Informationen über die Argumentsättigung bei Rektionskomposita beim Parsen verwendet werden; Regeln und Lexika für die maschinelle Übersetzung können auf der Grundlage einer Beschreibung und Analyse regelmäßiger Komposition einfacher und robuster gemacht werden (Rackow et al., 1992); bei der Extraktion von Fachwortschatz aus Texten oder bei der Informationsextraktion lassen sich mit einem morphembasierten Ansatz Ausbeute und Präzision steigern. Syllabifizierung und Aussprache, z. B. als Information für ein Sprachsynthesystem, lassen sich zuverlässig nur unter Ausnutzung morphologischer Strukturinformationen bestimmen. Die Analysen des DeKo-Systems können – da sie neben der Zerlegung auch Strukturinformation enthalten (siehe Abschnitt 5.2.3) – im Prinzip für alle diese Applikationen verwendet werden. Im DeKo-Projekt werden die Analysen bereitgestellt und experimentell in das deutsche Sprachsynthesystem Festival² eingebunden.

Da die meisten NLP-Anwendungen mit nicht gesehenen Texten umgehen müssen (maschinelle Übersetzung, TTS-Systeme, textverstehende Systeme) und da die Wortbildung produktiv ist, reicht ein festes Lexikon nicht aus. In den uns bekannten Wortbildungssystemen für das Deutsche werden jedoch (zum Teil sehr große) endliche Lexika ausgewertet, neu gebildete Wortformen können oft nicht analysiert werden.³

* Erschienen in: *Proceedings der GLDV-Frühjahrstagung 2001*, Henning Lobin (Hrsg.), Universität Gießen, 28.–30. März 2001, Seite 49–57. <http://www.uni-giessen.de/fb09/ascl/gldv2001/>

¹ Das System wird seit Januar 2000 im DeKo-Projekt entwickelt (gefördert vom Ministerium für Wissenschaft und Kunst des Landes Baden-Württemberg im Rahmen des Forschungsschwerpunktprogramms).

² <http://www.ims.uni-stuttgart.de/phonetik/synthesis/index.html>

³ Dies bezieht sich zumindest auf die im WWW zugänglichen Versionen dieser Systeme am 26.01.2001: Deut-

Wortbildungsprozesse sind unterschiedlich produktiv. Die Produktivität wird dabei qualitativ von vielen Eigenschaften bestimmt, wie im folgenden Abschnitt beschrieben. Neben der qualitativen Analyse kann man Produktivität auch quantitativ bestimmen: Man ermittelt die Wahrscheinlichkeit, mit der ein neues Wort eines bestimmten Wortbildungsmusters in einem Text auftritt. In DeKo wird die Produktivitätsrate für alle Wortbildungsprozesse berechnet (nach den Methoden von Baayen, 1992, Baayen, 2001, auf Daten aus 200 Millionen Wortformen Zeitungstext, aufbereitet nach Lüdeling et al., 2000a und Evert und Lüdeling, 2001).

5.2. Linguistische Grundannahmen

5.2.1. Verschiedene Typen von komplexen Wörtern

Wir unterscheiden in DeKo zwei Typen von komplexen Wörtern: Einerseits gibt es Wörter, die durch produktive Prozesse gebildet werden und daher morphosyntaktisch und semantisch regelmäßig sind. Diese werden im DeKo-System durch Finite-State-Regeln analysiert, wie unten beschrieben. Andererseits findet man Wörter, die zwar morphologisch komplex sind, aber entweder nicht auf allen Ebenen kompositionell sind oder keinem produktiven Muster angehören. Diese können (bzw. sollen) nicht durch Regeln analysiert werden und sind deshalb mit ihrer Strukturbeschreibung und einem Hinweis auf ihre Nichtkompositionalität im Lexikon gespeichert. Daneben gibt es Wörter, die auf den ersten Blick so aussehen, als entstammten sie Wortbildungsprozessen, die aber (synchron) nicht mehr in ihre Bestandteile zerlegt werden können – diese werden als Simplizia im Lexikon gespeichert. Dies möchten wir am Beispiel von *-lich* illustrieren: (1) gibt einige der produktiven *-lich*-Muster an,⁴ (2) einige der Wörter, die man morphologisch zerlegen möchte, die aber semantisch nicht in ein Muster fallen – *kränklich* bedeutet nicht *ein bisschen krank* – und (3) einige Beispiele für nicht-zerlegbare Wörter.

- (1) a. „im Bezug auf N“: ärztlich, bischöflich, richterlich, ...
b. „ein bisschen Adj“: grünlich, rötlich, säuerlich, ...
c. „jede(s) N“: wöchentlich, monatlich, jährlich, ...
- (2) leiblich, häuslich, wunderlich, rundlich, kränklich, ...
- (3) möglich, gefissentlich, scheußlich, niedlich, ...

Sind keine Lexikoneinträge vorhanden, so werden die Kandidaten anhand der Regeln analysiert, ohne Unterscheidung der Typen. Die Behandlung der regelmäßigen Prozesse wird im nächsten Abschnitt näher beschrieben.

sche Malaga-Morphologie (DMM): <http://www.linguistik.uni-erlangen.de/cgi-bin/orlorenz/dmm.cgi>, Deutscher Wortschatz: http://wortschatz.uni-leipzig.de/index_js.html, WordManager: <http://services.canoo.com/MorphologyBrowser.html>.

⁴ Die semantische Beschreibung ist hier lediglich intuitiv. Die semantische Beschreibung ist nicht Gegenstand des DeKo-Projekts; sie kann auf den von DeKo zur Verfügung gestellten Analysen durchgeführt werden.

5.2.2. Wortbildungsrelevante Informationen

In der empirischen und theoretischen Literatur zur Wortbildung werden seit langem die verschiedenen Arten von Wissen diskutiert, die Wortbildungsprozesse beschränken können.⁵ Affixe wählen die mit ihnen kombinierbaren Basen u. a. nach Wortart (4a), Argumentstruktur (4b), Herkunft (fremd oder nativ, 4c), morphologischer Beschaffenheit (Abkürzung oder Vollwort, 4d) oder Semantik (Konzept, 4e) aus.⁶

- (4) a. *-bar* nur mit Verben: beweisbar, *tischbar, *grünbar
- b. *-bar* primär mit transitiven Verben: lesbar, *schlafbar
- c. *-abel* nur mit klassischen Basen: akzeptabel, *annehmabel
- d. *-ler* mit Abkürzungen: CDUler,
-lich nicht mit Abkürzungen: *CDUlich
- e. *-fach* nur mit Kardinalzahlen: dreifach, *tischfach

In den vorhandenen computerlinguistischen Systemen zur Wortbildungsanalyse (z. B. Domenig und ten Hacken, 1992, Guenther, 1996, Fischbach und Kilbury, 1999) wird bisher meistens nur die Wortarteninformation verwendet, aber nicht in ausreichendem Maße auf das übrige Wissen zurückgegriffen.

Das liegt zum einen daran, dass es noch wenig computerlesbare Lexika gibt, in denen die entsprechenden Informationen zu den Stämmen und Affixen kodiert sind. Daher wird parallel zum DeKo-Regelmechanismus ein solches Lexikon entwickelt (Lüdeling et al., 2000b)⁷; wo notwendig, werden die Informationen mit in DeKo entwickelten Werkzeugen (halb)automatisch aus Text akquiriert.

Zum anderen fehlt für viele Wortbildungsprozesse eine wirklich systematische Beschreibung, die so standardisiert ist, dass man die relevanten Informationen leicht in Regeln fassen kann. In DeKo werden daher die Eigenschaften der einzelnen Wortbildungsprozesse nach genau definierten Kriterien untersucht und beschrieben. Das ermöglicht eine systematische Kodierung der oben genannten Beschränkungen in Regeln für die Zerlegung und die Strukturbeschreibung. Tab. 5.1 zeigt einen Ausschnitt aus einer DeKo-Beschreibungstabelle: hier wird eines der produktiven Muster des Suffixes *-lich* beschrieben. Informationen zum Affix, zur Basis und zum Wortbildungsprodukt (WBP) werden dabei getrennt aufgeführt.

Ähnliche Tabellen werden auch für Kompositionsmuster erarbeitet.⁸ Tab. 5.2 stellt einen Ausschnitt der Beschreibungstabelle für Adjektiv+Adjektiv-Komposita dar.

⁵ Zur deutschen Wortbildung siehe dazu vor allem die empirischen Beschreibungen in Fleischer und Barz (1992) und den Bänden *Deutsche Wortbildung* (Kühnhold und Wellmann, 1973, Wellmann, 1975, Kühnhold et al., 1978, Ortner et al., 1991, Pümpel-Mader et al., 1992).

⁶ In allen Beispielen werden nur die noch produktiven Muster angesprochen; oft gibt es einige vermeintliche Gegenbeispiele, die aber nicht mehr produktiv sind.

⁷ Dabei werden zu jedem Lemma Kompositions- und Derivationsstämme im Sinne von Fuhrhop (1998) angegeben.

⁸ Auf dem Stand von Ende März 2001 gibt es in DeKo Tabellen für ca. 270 Affixe und Kompositionsmuster.

Affix	erzeugt WBP mit der Kategorie	Vorkommen in 200 M Wortformen (Typen)	Vorkommen in 200 M Wortformen (Tokens)	davon Hapax Legomena ^a	Produktivitätsrate ^b	nativ oder fremd	semantische Funktionen
Affix	-lich	Adjektiv	873	1 473 983	169	0,00011	nativ relativ, ornativ, ...
Muster	Wortart, morpho-syntaktische Merkmale	konzeptuelle Spezifikation	klassische Basen?	englische Basen?	sonstige fremde Basen?	Abkürzungen?	morphologisch komplexe Basen?
Basis	<i>bischöflich, ärztlich, amtsärztlich, richterlich</i>	NN, masc	Berufsbezeichnung	nein	nein	nein	ja, Nominal-komposita
grammatische Spezifikation	konzeptuelle Spezifikation	Konnotation	Fuge?	Tilgung?	Akzentwechsel?	Umlaut?	
Wort-BILDUNGS-PRODUKT	adverbial, attributiv, prädiaktiv	Eigenschaft, relativ	neutral	nein	nein	nein	ja

^a Hapax Legomena sind solche Typen, die nur einmal im Korpus vorkommen. Diese sind gute Indikatoren für Produktivität, da man annehmen kann, dass lexikalisierte Wörter von verschiedenen Autoren in verschiedenen Texten verwendet werden (und somit mehr als einmal vorkommen), während spontan gebildete Wörter – da sie ja nicht der Sprechergemeinschaft zur Verfügung stehen – nur einmal auftreten. Zur Diskussion siehe Baayen (1992) und Baayen und Lieber (1991).

^b Die Produktivitätsrate berechnen wir hier nach Baayen (1992): Anzahl der Hapax Legomena/Anzahl Tokens. Dies ist nur für den ersten Eindruck ausreißend, für angemessene Berechnungen müssen LNRE-Modelle benutzt werden, siehe Baayen (2001).

Tabelle 5.1.: Tabelle für ein produktives Muster von *-lich*

Muster	Spezifikation	Beispiele	Semantik des Wortbildungsprodukts	rekursiv?	produktiv?	Wortakzent
Adj + Adj → Adj	Simplex + Simplex	hellblau, frühweise	Determinativkompositum	nein	ja	Nichtkopf
	Derivat + Simplex	giftiggelb	Determinativkompositum	nein	ja	Nichtkopf
	Simplex + Derivat	tiefgläubig, schwerverständlich	Determinativkompositum	nein	ja	Nichtkopf
	Simplex + Partizip	buntgestrichen, gutorganisiert	Determinativkompositum	nein	ja	Nichtkopf
...						
	Simplex + Simplex	grüngelb, sauerscharf	Kopulativkompositum	ja	ja	Nichtkopf und Kopf
...						

Tabelle 5.2.: Tabelle für ein Kompositionsmuster (Adj + Adj)

5.2.3. Hierarchische Strukturen

Für einige computerlinguistische Anwendungen reicht die reine Zerlegung der Wörter in ihre Bestandteile; andere Anwendungen, wie zum Beispiel semantische Desambiguierung oder Sprachsynthese, brauchen zusätzlich die hierarchische Strukturbeschreibung. Zur Erstellung von hierarchischen Beschränkungen verwenden wir wieder das in den Tabellen gesammelte linguistische Wissen. Aus den Beschränkungen in (5) folgt, dass *unregierbar* nur die Struktur (6a) haben kann. Die Struktur (6b) ist nicht möglich, da hier *un-* mit einem Verb zu *unregier-* verbunden würde, was durch (5a) ausgeschlossen ist.

- (5) a. *-un* verbindet sich produktiv nur mit Adjektiven
b. *-bar* verbindet sich produktiv nur mit transitiven Verben

- (6) a. $(un \cdot (regier_V \cdot bar_{Adj})_{Adj})_{Adj}$
b. $*((un \cdot regier_V)_V \cdot bar_{Adj})_{Adj}$

Kompositionsmuster können weit weniger beschränkt werden als Derivationsmuster. Dies führt zu einer großen Anzahl von möglichen Strukturen bei mehr als zweigliedrigen Komposita. Beschränkungen der Produktivität und Rekursivität können hier helfen, Ambiguitäten zu vermeiden. So ist beispielsweise der Kopf bei Adjektiv-Determinativkomposita fast nie komplex. Das heißt, wenn man ein Adjektiv+Adjektiv+Adjektiv-Kompositum oder ein Nomen+Nomen+Adjektiv-Determinativkompositum findet, ist fast immer das Erstglied komplex, wie in (7) illustriert⁹:

- (7) a. ((rot-grün)blind), %(rot-(grünblind))
b. ((sonnen(unter+gangs))rot), %(sonnen((unter+gangs)rot))

5.3. Implementierung

Die Modellierung erfolgt mit Werkzeugen auf der Grundlage der *Finite State Transducer* (FST) (van Leeuwen, 1990, Roche und Schabes, 1997). Die verschiedenen Aspekte der linguistischen Analyse und Beschreibung der Prozesse, die bei der Derivation und Komposition auftreten, erfordern unterschiedliche Repräsentationen und Formalismen:

- Die sequenzielle Zerlegung morphologisch komplexer Wörter lässt sich durch eine deklarative Grammatik beschreiben, die die Kombinierbarkeit morphologischer Einheiten (Basen, Affixe) angibt.
- Hierarchische Strukturinformation lässt sich durch eine kontextfreie Grammatik beschreiben.
- Phonologische Prozesse (etwa zur Behandlung des Umlauts oder des Wortakzentes) werden als kontextsensitive Rewrite-Regeln modelliert.

⁹ Mit einem „%“ markierte Strukturen sind nicht ungrammatisch, sondern nur nicht präferiert.

Trotz des inhärent heterogenen Charakters dieser verschiedenen Ebenen der linguistischen Beschreibung lässt sich ein homogener Implementierungsansatz finden. Bei allen Teilproblemen handelt es sich nämlich um die Transformation einer Symbolfolge in eine andere Symbolfolge. Ein flexibles Modell für die Konvertierung von Symbolketten beruht auf der Technologie der endlichen Automaten.

In DeKo werden die FST-Softwarepakete *FSM Library* (Mohri et al., 2000) und *Lextools* (Sproat, 2000) der AT&T Research Labs verwendet. Mit Hilfe dieser FST-Compilersuite können unterschiedliche linguistische Beschreibungsformalisten in ein einheitliches Format von gewichteten endlichen Transducern (weighted FST, WFST) konvertiert werden (Mohri, 1997). Diese können durch mathematische Operationen miteinander verknüpft werden.

Im Folgenden werden (vereinfachte) Beispiele für die verschiedenen Regeltypen gegeben.

- kontextsensitive Rewrite-Regel: In der Beispielregel wird das Graphem „e“ am Wortende von Nomina im Singular getilgt, wenn das Suffix *-lich* folgt (z. B. für *stündlich*). Lexikalische und morphologische Merkmale stehen in eckigen Klammern, „+“ markiert die Morphemgrenze:

$$e \rightarrow \epsilon / _ [nomen] [sg] + lich [suff]$$

- deklarative Grammatik zur sequenziellen Zerlegung morphologisch komplexer Wörter: Dargestellt ist ein Ausschnitt der Grammatik, der *unregierbar* wie folgt sequenziell zerlegt: un[adj.pref] + regier[verb] + bar[suff][adj]

Dabei bezeichnet die erste Spalte der folgenden Beispielgrammatik den Ausgangszustand, die zweite Spalte den Zielzustand und die dritte Spalte die Symbolfolge auf dem Übergang zwischen den beiden Zuständen in einem endlichen Automaten:

START	PREF	un[adj.pref] +
	PREF	STEM regier[verb]
	STEM	SUFF + bar[suff][adj]

- kontextfreie Grammatik zur hierarchischen Strukturierung: Hier wird der Zerlegung für *unregierbar* die hierarchische Struktur (markiert durch geschweifte Klammern) zugewiesen:

$$\begin{aligned} ADJ &\rightarrow \{ADJ.PREF + ADJ\} \\ ADJ &\rightarrow \{VSTEM + ADJ.SUFF\} \\ &\rightarrow \{un + \{regier + bar\}_{Adj}\}_{Adj} \end{aligned}$$

Abb. 5.1 gibt einen schematischen Überblick über die Module in DeKo. Zunächst findet eine Zerlegung der Eingabe in Morpheme (Lexikonzugriff) statt, wobei evtl. noch Umlautungs-, Tilgungs- oder Fugungsregeln angewendet werden. Die verschiedenen möglichen Zerlegungen werden durch Restriktionsregeln disambiguiert (Beschränkungen über die Kombinierbarkeit bestimmter Affixe mit umgelauteten bzw. nicht umgelauteten Stämmen, Beschränkungen über Affixkombinationen etc.). Der zerlegten Symbolfolge wird (mindestens) eine hierarchische Struktur zugewiesen. Die so analysierten komplexen Wörter können in vielen computerlinguistischen Anwendungen weiterverarbeitet werden.

Für das Text-to-Speech-System Festival wird auf der Grundlage der morphologischen Information nun die Syllabifizierung (nach Müller et al., 2000) durchgeführt. Zum Schluss erfolgen

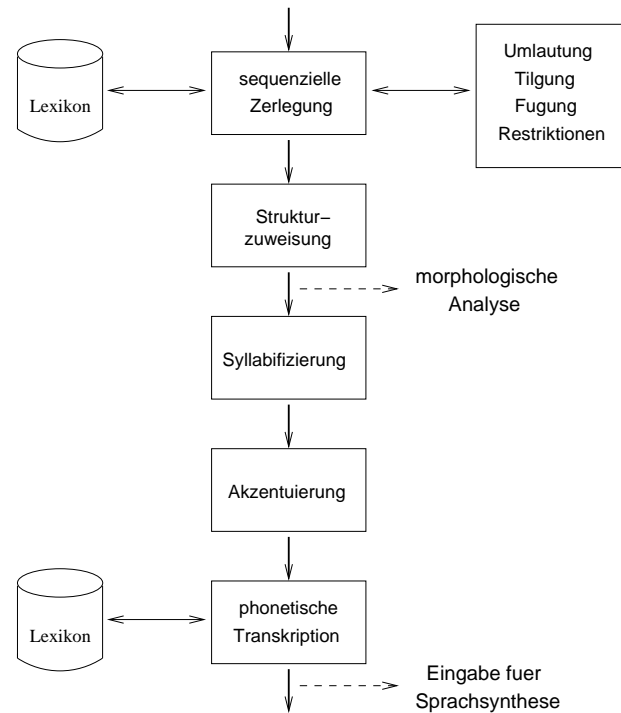


Abbildung 5.1.: Schematisierter Überblick über die Automatenarchitektur

noch die Akzentzuweisung (manche Affixe beeinflussen den Akzent) und die phonetische Transkription (durch ein Aussprachelexikon oder Letter-to-Sound-Regeln). Silbeninformation und Akzent werden z. B. für Sprechrhythmus und Betonung in der Sprachsynthese benötigt.

5.4. Zusammenfassung

Das DeKo-System zur Zerlegung von komplexen Wörtern arbeitet dreistufig: für die produktiven Prozesse werden wortbildungsrelevante Informationen auf allen linguistischen Ebenen systematisch erfasst. Parallel erfolgt der Aufbau eines Lexikons, in dem solche Informationen für Stämme und Affixe kodiert werden können. Die Regeln für die Zerlegung der komplexen Wörter werden in einem Finite-State-Formalismus kodiert. Auf der Zerlegung wird dann die Struktur aufgebaut.

Für die Sprachsynthese wird außerdem auf den morphologisch analysierten Wörtern Syllabifizierung, Akzentzuweisung und phonetische Transkription vorgenommen.

Literaturverzeichnis

- BAAYEN, R. H. (1992): "Quantitative aspects of morphological productivity". In: *Yearbook of Morphology 1991*, herausgegeben von Booik, G. und v. Marle, J., Dordrecht: Foris, S. 109–150.
- BAAYEN, R. H. (2001): *Word Frequency Distributions*. Dordrecht: Kluwer.
- BAAYEN, R. H. UND LIEBER, R. (1991): "Productivity and English derivation: a corpus-based study". *Linguistics* 29: S. 801–843.

- DOMENIG, M. UND TEN HACKEN, P. (1992): *Word Manager: A system for Morphological Dictionaries*. Hildesheim: Olms.
- EVERT, S. UND LÜDELING, A. (2001): "Measuring morphological productivity: Is automatic preprocessing sufficient?" In: *Corpus Linguistics 2001*. Lancaster.
- FISCHBACH, C. UND KILBURY, J. (1999): "Realisierung paradigmatischer Derivationsmorphologie in finite-state Umgebungen". In: *Multilinguale Corpora. Codierung, Strukturierung, Analyse. 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung*, herausgegeben von Gippert, J. Prag: Enigma, S. 245–252.
- FLEISCHER, W. UND BARZ, I. (1992): *Wortbildung der deutschen Gegenwartssprache*. Tübingen: Niemeyer.
- FUHRHOP, N. (1998): *Grenzfälle morphologischer Einheiten*. Tübingen: Stauffenberg.
- GUENTHNER, F. (1996): "Electronic lexica and corpora research at CIS". *International Journal of Corpus Linguistics* 1 (2).
- KÜHNHOLD, I.; PUTZER, O. UND WELLMANN, H. (1978): *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache 3: Das Adjektiv*. Düsseldorf: Schwann.
- KÜHNHOLD, I. UND WELLMANN, H. (1973): *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache 1: Das Verb*. Düsseldorf: Schwann.
- LÜDELING, A.; EVERT, S. UND HEID, U. (2000a): "On measuring morphological productivity". In: *Proceedings of the KONVENS 2000. Sprachkommunikation*. Ilmenau, S. 57–61.
- LÜDELING, A.; SCHMID, T.; HEID, U.; SÄUBERLICH, B.; FITSCHEN, A. UND MÖBIUS, B. (2000b): "Ein integriertes Lexikon". Technischer Bericht, IMS, Universität Stuttgart.
- MOHRI, M. (1997): "Finite-state transducers in language and speech processing". *Computational Linguistics* 23 (2): S. 269–311.
- MOHRI, M.; PEREIRA, F. UND RILEY, M. (2000): "FSM library – general purpose finite-state machine software tools". Online verfügbar: <http://www.research.att.com/sw/tools/fsm/>.
- MÜLLER, K.; MÖBIUS, B. UND PRESCHER, D. (2000): "Inducing probabilistic syllable classes using multivariate clustering". In: *Proceedings of the 38th Annual Meeting of the ACL*. Hong Kong, S. 225–232.
- ORTNER, L.; BOLLHAGEN-MÜLLER, E.; ORTNER, H.; WELLMANN, H.; PÜMPEL-MADER, M. UND GÄRTNER, H. (1991): *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache 4: Substantivkomposita*. Berlin, New York: de Gruyter.
- PÜMPEL-MADER, M.; GASSNER-KOCH, E. UND WELLMANN, H. (1992): *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache 5: Adjektivkomposita und Partizipialbildungen*. Berlin: de Gruyter.
- RACKOW, U.; DAGAN, I. UND SCHWALL, U. (1992): "Automatic translation of noun compounds". In: *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92)*. ICCL, Nantes, S. 1249–1253.
- ROCHE, E. UND SCHABES, Y. (Herausgeber) (1997): *Finite-State Language Processing*. Cambridge: MIT Press.
- SPROAT, R. (2000): "Lextools: a toolkit for finite-state linguistic analysis". Online verfügbar: <http://www.research.att.com/sw/tools/lextools/>.
- VAN LEEUWEN, J. (Herausgeber) (1990): *Handbook of Theoretical Computer Science*. Amsterdam, Cambridge: Elsevier, MIT Press.
- WELLMANN, H. (1975): *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache 2: Das Substantiv*. Düsseldorf: Schwann.